Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.01

Lecture 20: Haplotype testing and Minimum GWAS analysis steps

Jason Mezey
jgm45@cornell.edu
April 17, 2017 (T) 8:40-9:55
Announcements

- Project will be posted today (see next slides)!

BTRY 4830/6830 & PBSB.5201.01
Quantitative Genomics and Genetics Spring 2018

Project (Version 1)

Posted April 17; Due 11:59PM May 8
Announcements

1 Introduction and instructions

The goal of the class project is for you to demonstrate what you have learned by performing a GWAS analysis on real data. To accomplish this, assume that you have been provided data by a collaborator who wants to identify positions of causal polymorphisms (loci). You will perform an in-depth analysis and write a report for your collaborator that explains your methods and results.

Instructions: While we provide some general guidelines for how to proceed below, the techniques you use to analyze the data and how you construct your report will be up to you. Do however note the following instructions (PLEASE READ THESE CAREFULLY!!):

(1) Your project must be uploaded by 11:59PM, May 8 - if it is late for any reason, standard grading policies apply.

(2) You are allowed to work together with other students in the class to analyze these data. However, note that turning in a report that describes exactly the same analyses as a fellow student is not a good strategy for getting a good grade. Also note that you must write your own report.

(3) This is an ‘open book’ assignment, such that you are allowed to use any resources online, in books, etc. You may also ask third-party (i.e. people not in the class) for suggestions on what analyses to perform but you cannot have a third-party do any of the analyses (or write any code for you!).

(4) You are also allowed to use any software or programming language that you would like as part of your analysis. However, we expect that some of the tasks will be performed in R (also note that you are welcome to use any packages, functions, etc. in R).

(5) Your final project will include a SINGLE report file and a SINGLE file including all of your R code (ideally and .rmd file!) and / or commands or scripts you used to run other software packages. That is, for your R code, the best way to maximize your grade is to have well commented code that we can run from the command line. If you use other software for some of the tasks, a reasonable approach is to include commented out descriptions in your code that provides details on how you ran the software, e.g. what parameters did you use, etc.

(6) The report file must be no more than 8 pages (single-sided), with NO MORE than 5 pages of text and NO MORE than 3 pages of figures / tables.

(7) For your report, you must describe what you did in detail (a good guide is have you provided enough detail such that someone reading your report could replicate what you have done?). You also need to describe the results you have obtained from your analysis. You may also wish to include some text to describe interpretations and conclusions that may be of interest to your collaborator, including statistical and possibly, biological interpretations. For your Figures and Tables, note that clarity and clear labels is a strategy for maximizing your grade.

(8) We will grade on two broad criteria: 1. the overall quality of the analyses / report, 2. the amount of effort put into your project. Note that ‘effort’ does not mean run many analyses without thinking carefully about why you are running them or how they fit together to provide a clear picture of results. A guide maximizing your grade on effort is to think carefully about how to produce the best possible report that you can and then put in as many hours as you wish to devote to the project accomplishing this objective (your effort level will be clear to us).
2 The experiment and data

The experiment: Among the recent large scale human genomics resources is Genetic European Variation in Health and Disease (gEUVADIS):

http://www.geuvadis.org/

with samples from 4 different European populations. Each of these individuals were part of the 1000 Genomes project and their genomes were sequenced and analyzed to identify SNP genotypes. For expression profiling, lymphoblastoid cell lines (LCL) were generated from each sample and mRNA levels were quantified through RNA sequencing.

Each of these gene expression measurements may be thought of as a phenotype and one can do a GWAS analysis on each individually, which is called an ?expression Quantitative Trait locus? or eQTL analysis, an unnecessarily fancy name for a GWAS when the phenotype is gene expression.

What you have been provided is a small subset of these data that are publicly available. Specifically, you have been provided 50,000 of the SNP genotypes for 344 samples from the CEU (Utah residents with European ancestry), FIN (Finnish), GBR (British) and, TSI (Toscani) population. For these same individuals, you have also been provided the expression levels of five genes. You have also been provided information on the population and gender of each of these individuals, and information regarding the position of each gene and SNP in the genome. A description of the broader data set from which these data were extracted can be found in::

http://www.geuvadis.org/web/geuvadis/RNAseq-project

and in other papers relating to analysis of the GEUVADIS data.

The data: These have been provided to you in five total files: ‘QG18_phenotypes.csv’, ‘QG18_genotypes.csv’, ‘QG18_covars.csv’, ‘QG18_gene_info.csv’, ‘QG18_SNP_info.csv’.

‘QG18_phenotypes.csv’ contains the phenotype data for 344 samples and 5 genes.

‘QG18_genotypes.csv’ contains the SNP data for 344 samples and 50000 genotypes.

‘QG18_gene_info.csv’ contains information about each gene that was measured. The ?chromosome? column indicates the chromosome where the gene is located, ?start? marks the position in the chromosome where the region of the gene begins and ?end? marks the position where the region ends. ?symbol? contains the common gene name of the measured transcript and ?probe? contains the ids of the transcripts that match with the column names of the phenotype data.

‘QG18_SNP_info.csv’ contains the additional information on the genotypes and has four columns. The 1st column contains the chromosome number of each SNP, the 2nd column contains the physical position of the SNP on the chromosome, the 3rd column contains the abbreviation used to the ?rsID? = the name of each SNP in order.
3 Your assignment and hints for getting started

Your GWAS assignment is to find the position of as many causal polymorphisms as possible for the five expressed genes using the data (note that each ‘hit’ will potentially indicate an eQTL). You may / should use any and as many analysis approaches as you think that are useful to accomplish this goal. In your report, you will need to describe in detail what you did, why you did it, and describe results in a manner that your ?non-statistical? collaborator will be able to understand, e.g. explain your terms, provide interpretations, etc.

A few hints:

- Apply the applicable steps of a ‘minimum GWAS’ analysis.
- In your report, justify why you applied each individual step and statistical approach.
- In your report, provide a summary of your results and what they mean.
- You may want to consider going to various resources online (e.g. genecards, UCSC genome browser, dbSNP, many others) to incorporate biological information into your interpretation and hypotheses concerning what you may have found.
- Ask Afrah, Zijun, and Jason for thoughts and ideas!

Good luck!
Review: Haplotype testing I

- We have just extended our GWAS framework to handle additional phenotypes.
- We can also extend our GWAS framework to handle genotypes defined using a different approach.
- In this case, let’s consider using haplotype alleles in our testing framework.
- Note that a haplotype collapses genetic marker information but in some cases, testing using haplotypes is more effective than testing one genetic marker at a time.
Review: Haplotype testing II

- **Haplotype** - a series of ordered, linked alleles that are inherited together

- For the moment, let’s consider a haplotype to define a “function” that takes a set of alleles at several loci A, B, C, D, etc. and outputs a haplotype allele:

  \[ h = f(A_i, B_j, ...) \]

- For example, if these loci are each a SNP with the following alleles (A,G), (A,T),(G,C),(G,C) we could define the following haplotype alleles:

  \[ h_1 = (A, A, C, C') \quad h_2 = (G, T, G, G) \]
Review: Haplotype testing III

- Note that how we define haplotype alleles is somewhat arbitrary but in general, we define a haplotype for a set of genetic markers (loci) that are physically linked that are frequently occur in a population.
- How many markers is somewhat arbitrary, e.g. we often define sets that match observed patterns of LD.
- How many haplotype alleles we define is also somewhat arbitrary, where we define haplotype alleles that have appreciable frequency in the population.
  - For example, four the four loci with alleles (A,G), (A,T), (G,C), (G,C) how many haplotype alleles could we define?
  - However, it could be that only the following two combinations have relatively “high” allele frequencies (say >0.05 = arbitrary!)

\[
h_1 = (A, A, C, C) \quad h_2 = (G, T, G, G)
\]
- In such a case, we can collapse the many alleles into just a few!
Review: Haplotype testing IV

- As an example of haplotype allele collapsing, say for our case of four loci (A,G), (A,T),(G,C),(G,C), we have lots of LD (!!) such that there are only 4 alleles in the population (i.e. all other combinations have frequency of zero!):


- Let’s also say that the frequencies of the third and fourth of these in the population are < 0.01

- In this case, we can define just two haplotype alleles that collapse the other alleles as follows (where * means “any” genetic marker allele):

  \[ h_1 = (A, A, *, C) \quad h_2 = (G, T, *, G) \]

  \[ h_1 = h_1^* \cup h_3^* \quad h_2 = h_2^* \cup h_4^* \]

- NOTE: we are therefore loosing information using this approach!!
Review: GWAS with haplotypes I

- Once we have defined haplotype alleles, we can proceed with a GWAS using our framework (just substitute haplotype alleles and genotypes for genetic marker alleles and genotypes!)

- For example, in a case where we only have two haplotype alleles, we can code our independent variables for our regression model as follows:

  \[ X_a(h_1 h_1) = -1, \quad X_a(h_1 h_2) = 0, \quad X_a(h_2 h_2) = 1 \]

  \[ X_d(h_1 h_1) = -1, \quad X_d(h_1 h_2) = 1, \quad X_d(h_2 h_2) = -1 \]

- All other aspects remain the same (although what is the effect on our interpretation of where the causal polymorphism is located?)
Review: GWAS with haplotypes II

• Given that we are losing information by using a haplotype testing approach in a GWAS, why might we want to use this approach?

• As one example consider the following case of haplotypes in a population:

\[
\begin{align*}
A_1 & \quad B_1 & (C_1)^* & \quad D_2 & \quad E_1 \\
A_1 & \quad B_2 & (C_1)^* & \quad D_1 & \quad E_1 \\
A_2 & \quad B_1 & (C_1)^* & \quad D_1 & \quad E_1 \\
A_1 & \quad B_1 & (C_1)^* & \quad D_1 & \quad E_2 \\
A_2 & \quad B_2 & (C_2)^* & \quad D_1 & \quad E_2 \\
A_2 & \quad B_1 & (C_2)^* & \quad D_2 & \quad E_2 \\
A_1 & \quad B_2 & (C_2)^* & \quad D_2 & \quad E_2 \\
A_2 & \quad B_2 & (C_2)^* & \quad D_2 & \quad E_1 \\
\end{align*}
\]
Advantages of haplotype testing

• In some cases (system and sample dependent!), the haplotype is a better “tag” of the causal polymorphism than any of the surrounding markers

• In such a case, the $\text{Corr}(X_h, X) > \text{Corr} (X', X)$ and therefore has a higher probability of correctly rejecting the null hypothesis

• Another “advantage” is by putting together markers, we are performing less total tests in our GWAS (in what sense is this an advantage!?)
Disadvantages of haplotype testing

- Collapsing to haplotypes may produce a better tag but it also may not (!!), i.e. sometimes (in fact often!) individual genetic markers are better tags of the causal polymorphism
- Another disadvantage is resolution, since we absolutely cannot resolve the position of the causal polymorphism to a position smaller than the range of the haplotype alleles, i.e. large haplotypes can have smaller resolution
- If we had measured the causal polymorphism in our data, should we use haplotype testing (i.e. in the future, the importance of haplotype testing may decrease)
Should I apply haplotype testing in my GWAS?

• Yes! but apply both an individual marker testing approach (always!) as well as a haplotype test (optional)

• The reason is that we never know the true answer in our GWAS (as with any statistical analysis!) so it doesn’t hurt us to explore our dataset with as many techniques as we want to apply

• In fact, this will be a continuing theme of the class, i.e. keep analyzing GWAS with as many methods as you find useful

• However, since we never know the right answer for certain, if we get conflicting results, which one do we interpret as “correct”!? 
Where do haplotypes come from?

- A deep discussion of the origin of haplotypes (remember: a fuzzy definition!) is another subject that is in the realm of population genetics and therefore we cannot discuss this in detail in this class (again: I encourage you to take a class on population genetics!)

- However, we can get an intuition about where haplotypes come from by remembering that the origin of new haplotype alleles are mutations and that new haplotype alleles can be produced by recombination

- In fact, these two processes also underlie the amount of LD in the population and therefore what blocks of alleles are inherited as a haplotype (and we therefore use them to define haplotypes using system specific criteria)
Defining haplotypes

- We could spend multiple lectures on how people define haplotypes for given systems and the algorithms used for this purpose (so we will just briefly mention the main concepts here)

- To define haplotypes, we need to “phase” measured genotype markers, decide on the number of genotype markers to put together into a haplotype block, and decide how many haplotype alleles to consider

- Remember: there are no universal rules for doing this (system dependent!)
Phasing haplotypes

- To get a sense of the phasing problem, consider a case where we have two markers that are right next to each other on a chromosome and we know we want to put them together in a haplotype block.

- Say one marker is (A,T) and the other marker is (G,C) and we are considering a diploid individual who is a heterozygote for both of these markers, which of the marker alleles are physically linked in this individual?

- Figuring this out for individuals in a sample is the phasing problem and there are many algorithms for accomplishing this goal (note that in the future, technology may make this a non-issue...).
Deciding on how many genotypes to include in a haplotype block

- Again, while there is no set rule, how we decide on genotypes to include in a haplotype block depends on LD.

- The general rule: if we have a set of markers in high LD with each other but low LD with other markers, we use this as a guide for defining the haplotype block.
Deciding on how many haplotype alleles to consider

- Again, there are no set rules for how many haplotype alleles to define, but in general, we define a set where the frequency in a population is above some MAF threshold (which depends on the system)

- With a MAF cutoff of say 0.05, this generally limits us to 2-5 haplotype alleles (e.g. in humans!)

- There are however cases where we might want to consider rarer haplotypes (what are some of these?)
Haplotype GWAS wrap-up

• Haplotypes are a physical and sampling consequence of how genetic systems work (just like LD!)

• Definitions of haplotype blocks and haplotype alleles depend on the system and context (fuzzy definition)

• Regardless of how we define them, once we have haplotype alleles, we can use them as we would genetic markers in our GWAS analysis framework

• While optional, it is never a bad idea to perform a haplotype analysis of your GWAS in addition to your single marker analysis (ALWAYS do a single marker analysis)
Minimal GWAS 1

• You have now reached a stage when you are ready to perform a real GWAS data on your own (please note that there is more to learn and analyzing GWAS well requires that you jump in and analyze!!)

• Our final concept to allow you to do this are *minimal GWAS steps*, i.e. a list of analyses you should always do when analyzing GWAS data (you now know how to do most of these, a few you will have to do additional work to figure out)

• While these minimal steps are fuzzy (=they do not apply in every situation!) they provide a good guide to how you should think about analyzing your GWAS data (in fact, no matter how experienced you become, you will always consider these steps!)
Minimal GWAS II

- The minimal steps are as follows:
  - Make sure you understand the data and are clear on the components of the data
  - Check the phenotype data
  - Check and filter the genotype data
  - Perform a GWAS analysis and diagnostics
  - Present your final analysis and consider other evidence

- Note I: the software PLINK (google it!) is a very useful tool for some (but not all) of these steps (but you can do everything in R!)

- Note II: GWAS analysis is not “do this and you are done” - it requires that you consider the output of each step (does it make sense? what does it mean in this case?) and that you use this information to iteratively change your analysis / try different approaches to get to your goal (what is this goal!?)
Minimal GWAS III: check data

• Look at the files (!!) using a text editor (if they are too large to do this - you will need another approach)

• Make sure you can identify: phenotypes, genotypes, covariates, and that you know what all other information indicates, i.e. indicators of the structure of the data, missing data, information that is not useful, etc. (also make sure you do not have any strange formatting, etc. in your file that will mess up your analysis!)

• Make sure you understand how phenotypes are coded and what they represent (how are they collected? are they the same phenotype?) and the structure of the genotype data (are they SNPs? are there three states for each?) - ideally talk to your collaborator about this (!!)
Minimal GWAS IV: phenotype data

• Plot your phenotype data (histogram!)
• Check for odd phenotypes or outliers (remove if applicable)
• Make sure it conforms to a distribution that you expect and can model (!!) - this will determine which analysis techniques you can use
  • e.g. if the data is continuous, is it approximately normal (or can be transformed to normal?)
  • e.g. if it has two states, make sure you have coded the two states appropriately and know what they represent (are there enough in each category to do an analysis?)
  • e.g. what if your phenotype does not conform to either?
Minimal GWAS V: genotype data

• Make sure you know how many states you have for your genotypes and that they are coded appropriately

• Filter your genotypes (fuzzy rules!):
  • Remove individuals with >10% missing data across all genotypes (also remove individuals without phenotypes!)
  • Remove genotypes with >5% missing data across the entire individual
  • Remove genotypes with MAF < 5%
  • Remove individuals that fail a test of Hardy-Weinberg equilibrium (where appropriate!)
  • Remove individuals that fail transmission, sex chromosome test, etc.

• Perform a Principal Component Analysis (PCA) to check for clustering of n individuals (population structure!) or outliers, i.e. use the covariance matrix among individuals after scaling genotypes (by mean and sd) and look at the loadings of each individual on the PCs (you may have to “thin” the data!)
Minimal GWAS VI: GWAS analysis

• Perform an association analysis considering the association of each marker one at a time (always do this not matter how complicated your experimental design!)

• Apply as many individual analyses as you find informative (i.e. perform individual GWAS each with a different statistical analysis technique), e.g. trying different sets of covariates, different types of tests (see next lecture!), etc.

• CHECK QQ PLOTS FOR EACH INDIVIDUAL GWAS ANALYSIS and use this information to indicate if your analysis can be interpreted as indicating the positions of causal polymorphisms (if not, try more analyses, different filtering, etc. = experience is key!)

• For significant markers (multiple test correction!) do a “local” Manhattan plot and visualize the LD among the markers ($r^2$ or $D'$ if possible but just a correlation of you $Xa$ can work) to determine if anything might be amiss

• Compare significant “hits” among different analyses (what might be causing the differences if there are any?)
Minimal GWAS VII: present results

• List ALL of the steps (methods!) you have taken to analyze the data such that someone could replicate what you did from your description (!!!), i.e. what data did you remove? what intermediate analyses did you do? how did you analyze the data? if you used software what settings did you use?

• Plot a Manhattan and QQ plot (at least!)

• Present your hits (many ways to do this)

• Consider other information available from other sources (databases, literature) to try to determine more about the possible causal locus, i.e. are there good candidate loci, control regions, known genome structure, gene expression or other types of data, pathway information, etc.
Conceptual Overview

- Genetic System
  - Does A1 → A2 affect Y?

- Sample or experimental pop
  - Measured individuals (genotype, phenotype)

- Reject / DNR
  - Model params
  - F-test

- Regression model
  - Pr(Y|X)
Conceptual Overview

System

Question

Experiment

Sample

Inference

Prob. Models

Statistics

Assumptions
That’s it for today

• See you on Thurs.!