Computational Workflows for Genome-Wide Association Study: I

Sorin Istrail

Department of Computer Science
Brown University, Providence
sorin@cs.brown.edu

November 3, 2015

Outline

Association Study

A genetic variant is genotyped in a population for which phenotypic information is available (such as disease occurence, or a range of different trait values). If a correlation is observed between genotype and phenotype, there is said to be an association between the variant and the disease or trait.

Quantitative Trait

A biological trait that shows continuous variation (such as height) rather than falling into distinct categories (such as the binary types: diabetic or healthy (i.e. with disease and without the disease)). The genetic basis of these traits generally involve the effects of multiple genes and gene-environment interactions. Examples of quantitative traits that contribute to disease are body mass index, blood pressure, blood lipid levels.

A first description of GWAS

A dense set of SNPs across the genome is genotyped to survey the most *common genetic variation* for a role in disease or to identify heritable quantitative traits that risk factors for disease.

Pro and Con for GWAS: Basic Questions

1. What constitute a statistical well-powered study?
2. What are the limitations of what such a study can discover?
3. How do we interpret their results?
4. What is the rationale for GWAS?
5. How do we design a GWAS to address rigorously its critical issues:
   - Power
   - Efficiency
   - Comprehensiveness
   - Interpretation
   - Analsys
6. What are different approaches/workflows for GWAS?
7. What are the technical and analytical issues that hinder their

Mapping the Genes the Underlie Common Disease and Quantitative Traits

There are two classes or methods

1. *Candidate Gene studies* - use of associations or re-sequencing approaches
2. *Genome-wide studies* - use Linkage Mapping and GWAS

Note that Admixture Mapping is a fundamentally different approach. (We will discuss it later.)

Candidate Gene

A gene for which there is evidence for its possible role in the trait or disease that is under study.

Linkage Mapping

Correlation between nearby variants such that the alleles at neighboring markers (observed on the same chromosome) are associated within a population more often than if they were unlinked.

Admixture Mapping

Predicting the recent ancestry of chromosomal segments across the genome to identify regions for which recent ancestry in a particular population correlates with disease or trait values. Such regions are more likely to contain casual variants that are more common in the ancestral population.

Disease Penetrance

The proportion of individuals with a specific genotype who "manifest the genotype at the phenotype level. "For example if all individuals with a specific disease genotype show the disease phenotype, then the genotype is said to be "completely penetrant."

Heritability

The proportion of variation in a given characteristic or state that can be attributed to (additive) genetic factors.

The Monogenic "Mendelian" Disease Case a success story of Genome-wide Linkage Mapping

- Genome-wide linkage analysis is the method traditionally used to identify disease genes, and has been traditionally used to identify disease genes; very successful for mapping genes that underlie monogenic "Mendelian" diseases.
- For linkage analysis to succeed markers that flank the disease gene must segregate with the disease in families.
- Variants that cause monogenic disease are often rare (probably because negative selection reduces the frequencies of variants that cause disease characterized by early-onset morbidity and mortality) so each segregating disease allele will be found in the same 10-20 cM chromosomal background within each family

Complex common disease - GW linkage mapping not successful

- Genome-wide linkage analysis applied to common disease was less successful.
- Some few cases were successful: inflammatory bowel disease, schizophrenia, and type1 diabetes.
- For most common diseases linkage analysis only a limited success and genes found only explain a small fraction of the overall heritability of the disease.

Factors responsible for the lack of success of linkage analysis

1. low heritability of most complex traits
2. the inability of the standard set of microsatellite markers -which are spaced 10cM apart - to extract complete information about inheritance
3. the imprecise definition of phenotype
4. inadequately powered study design
5. less powerful for identifying common genetic variants that have modest effect on disease
6. poor power for detecting common alleles that have lowpenetrance

Candidate Gene Studies - Resequencing

- Candidate Genes Studies - a practical alternative to linkage analysis
- Candidate genes are hypothesis-based studies; genes are selected based on location in a region of linkage or on based on other evidence that they affect disease risk
- Most effective candidate gene analysis is done by
  - resequencing the entire genes in patients and controls
  - searching for a variant or set of variants that is enriched or depleted in disease cases
- These are laborious studies limited to coding regions of few candidate genes
- Properly interpreting results are difficult especially when considering rare non-coding variants

Candidate Gene Sets Association Studies

- Association studies using common allelic variants are cheaper and simpler than the complete resequencing of candidate genes, and have been proposed as a powerful mean of identifying the common variants that underlie complex traits.
- Simply put, such an association study compare the frequencies of alleles or genotypes of a particular variant between disease cases and controls. To avoid the difficult problem of population stratification family-based controls are used as an alternative approach.
- Candidate gene association studies have identified many of the genes that are known to contribute to susceptibility to common disease.
- Such studies are based on the using indirect Linkage

Genome-Wide Association Methods

- We define a genome-wide association approach as an association study that surveys most of the genome for causal genetic variants.
- No assumptions are made about the genomic locations of the causal variants, this approach would exploit the strength of the association studies without having to guess the identity of the causal genes.
- Represents an unbiased but fairly comprehensive approach that can be attempted even in the absence of convincing evidence regarding the function and location of causal genes.
- GWAS require:
  - Knowledge about common genetic variation, especially SNPs
  - ability to genotype a sufficiently comprehensive of set of

Markers for Genome-Wide Association Studies

- Markers could be SNPs, microsatellites, or any other loci on the genome
- Useful Markers tested for association must be either causal or highly correlated, that is in Linkage Disequilibrium (LD) with the causal allele.
- The view of the genome is a that a lot of it, maybe 80%, is made out of highly correlated LD "blocks" or also called haplotype block; a block is a consecutive segment of the genome where all the SNPs residing in the block are strongly correlated with each other.
- The most fundamental part is that in such and LD block most chromosomes of the population carry only one of only a few common haplotypes.

SNP "Captains": SNPs as proxies or tagging SNPs

- studies, including the ENCODE Project, which focused on comprehensive analyses of small regions of about 500 kb, showed that most of the 11 million common SNPs in the genome have clusters of neighbors that are all nearly perfectly correlated with each other = the genotype of one SNP perfectly predicts those correlated neighboring SNPs (note: two different SNPs have potentially different alleles sets, but as we know relabeling both with $0, 1$ makes this test well defined)
- One SNP then can be a proxy for several in a LD or haplotype block
- Here comes crucial the need for accurate and detailed knowledge of the empirical patterns of LD across the genome

HapMap

- Studies showed that a few hundred thousand well chosen SNPs should be adequate to provide information about most of the common variation in the genome.
- African populations would require a much larger number of tag SNPs because these populations because these populations contain more variation and less LD.
- In a nutshell, the evaluation of how many SNPs are required for these populations depends on:
  - the method used to select SNPs
  - the degree of LD in blocks and outside blocks
  - the degree of long-range LD between blocks
  - the efficiency with which SNPs in regions of low LD can be tagged
  - there are several algorithms proposed for tagging SNPs

Rare Alleles in GWAS

The problem of rare allele is crucial for the GWAS analysis.

- The frequency and penetrance of causal alleles affect the statistical power to detect these alleles.Power increases with increased frequency and increased penetrance.
- It is considered that instead of these two parameters, one might consider a new measure, a single parameter: **the amount of variation in phenotype that can be "explained" by the genetic variant in question**
- The consequence of such a new measure is that **rare highly penetrant alleles** and **common low penetrant alleles** are on an equal footing.
- The conclusion is then that the power to detect such an allele therefore depends on what is ultimately the most relevant

Population

In population genetics, the term population does not refer to the entire species, but to a group of organisms of the same species living within a sufficiently restricted geographic area, such that any member can mate with any other member of opposite sex. There are difficulties with this definition. One relates to the fact that geography creates some typically non-random pattern in the spatial distribution of organisms; the members are not uniformly distributed but they are in clusters or colonies, hard to define formally.