Genome-Wide Association Studies (GWAS): Computational Themes and Caveats

Sorin Istrail

October 14, 2014
Many issues in Genomewide Association Studies

We show that even for the simplest analysis, there is little consensus for the most appropriate statistical procedure.

We will attempt to make a list of the complicating factors.
Issues in Genomewide Association Studies

- Many issues in Genomewide Association Studies
- We show that even for the simplest analysis, there is little consensus for the most appropriate statistical procedure
- We will attempt to make a list of the complicating factors.

**Goal of Genomewide Association Studies:** *Identify patterns of polymorphisms that vary systematically between individuals with different disease states (in particular, healthy and disease) and could therefore represent the effect of risk-enhancing or protective alleles.*
In classifying patterns of the states, we will deal mostly biallelic SNPs, but in practice there are more complex patterns, including triallelic SNP instances.
In classifying patterns of the states, we will deal mostly biallelic SNPs, but in practice there are more complex patterns, including triallelic SNP instances.

Association studies should address the following issues.

- Hardy-Weinberg equilibrium testing
- Inference of phase and missing data
- SNP tagging
- Single SNP test of association
- Multi SNP test of association
- Population stratification
- Multiple Testing Correction
In classifying patterns of the states, we will deal mostly biallelic SNPs, but in practice there are more complex patterns, including triallelic SNP instances.

Association studies should address the following issues.

(A) Hardy-Weinberg equilibrium testing
In classifying patterns of the states, we will deal mostly biallelic SNPs, but in practice there are more complex patterns, including triallelic SNP instances.

Association studies should address the following issues.

(A) Hardy-Weinberg equilibrium testing
(B) Inference of phase and missing data
In classifying patterns of the states, we will deal mostly biallelic SNPs, but in practice there are more complex patterns, including triallelic SNP instances.

Association studies should address the following issues.

(A) Hardy-Weinberg equilibrium testing
(B) Inference of phase and missing data
(C) SNP tagging
In classifying patterns of the states, we will deal mostly biallelic SNPs, but in practice there are more complex patterns, including triallelic SNP instances.

Association studies should address the following issues.

(A) Hardy-Weinberg equilibrium testing
(B) Inference of phase and missing data
(C) SNP tagging
(D) Single SNP test of association
In classifying patterns of the states, we will deal mostly biallelic SNPs, but in practice there are more complex patterns, including triallelic SNP instances.

Association studies should address the following issues.

(A) Hardy-Weinberg equilibrium testing
(B) Inference of phase and missing data
(C) SNP tagging
(D) Single SNP test of association
(E) Multi SNP test of association
In classifying patterns of the states, we will deal mostly biallelic SNPs, but in practice there are more complex patterns, including triallelic SNP instances.

Association studies should address the following issues.

(A) Hardy-Weinberg equilibrium testing
(B) Inference of phase and missing data
(C) SNP tagging
(D) Single SNP test of association
(E) Multi SNP test of association
(F) Population stratification
In classifying patterns of the states, we will deal mostly biallelic SNPs, but in practice there are more complex patterns, including triallelic SNP instances.

Association studies should address the following issues.

(A) Hardy-Weinberg equilibrium testing
(B) Inference of phase and missing data
(C) SNP tagging
(D) Single SNP test of association
(E) Multi SNP test of association
(F) Population stratification
(G) Multiple Testing Correction
Different types of Population Association Studies.

- **Candidate Polymorphism Association Study**
  - These studies focus on individual polymorphism that is suspected of being indicated in disease causation. The purpose of the study is to perform further validation and testing.
Different types of Population Association Studies.

- **Candidate Polymorphism Association Study**
  - These studies focus on individual polymorphism that is suspected of being indicated in disease causation. The purpose of the study is to perform further validation and testing.

- **Candidate Gene Association Study**
Different types of Population Association Studies.

- **Candidate Polymorphism Association Study**
  - These studies focus on individual polymorphism that is suspected of being indicated in disease causation. The purpose of the study is to perform further validation and testing.

- **Candidate Gene Association Study**
  - Typically involves 5-50 SNPs within a gene, defined as to include coding region flanking genes, regulatory modules, splicing regions
Issues in Genomewide Association Studies

Different types of Population Association Studies.

- **Candidate Polymorphism Association Study**
  - These studies focus on individual polymorphism that is suspected of being indicated in disease causation. The purpose of the study is to perform further validation and testing.

- **Candidate Gene Association Study**
  - Typically involves 5-50 SNPs within a gene, defined as to include coding region flanking genes, regulatory modules, splicing regions.
  - Focuses on genes indicated in prior studies.
Issues in Genomewide Association Studies

Different types of Population Association Studies.

- **Candidate Polymorphism Association Study**
  - These studies focus on individual polymorphism that is suspected of being indicated in disease causation. The purpose of the study is to perform further validation and testing.

- **Candidate Gene Association Study**
  - Typically involves 5-50 SNPs within a gene, defined as to include coding region flanking genes, regulatory modules, splicing regions
  - Focuses on genes indicated in prior studies
  - Utilizes information from linkage studies (pedigree information, such as sequences of parents, grandparents, etc.)
Different types of Population Association Studies.

- **Candidate Polymorphism Association Study**
  - These studies focus on individual polymorphism that is suspected of being indicated in disease causation. The purpose of the study is to perform further validation and testing.

- **Candidate Gene Association Study**
  - Typically involves 5-50 SNPs within a gene, defined as to include coding region flanking genes, regulatory modules, splicing regions
  - Focuses on genes indicated in prior studies
  - Utilizes information from linkage studies (pedigree information, such as sequences of parents, grandparents, etc.)
  - Homology information: homology to a gene in a model organism (such as mouse)
Issues in Genomewide Association Studies

Different types of Population Association Studies.

- **Candidate Polymorphism Association Study**
  - These studies focus on individual polymorphism that is suspected of being indicated in disease causation. The purpose of the study is to perform further validation and testing.

- **Candidate Gene Association Study**
  - Typically involves 5-50 SNPs within a gene, defined as to include coding region flanking genes, regulatory modules, splicing regions
  - Focuses on genes indicated in prior studies
  - Utilizes information from linkage studies (pedigree information, such as sequences of parents, grandparents, etc.)
  - Homology information: homology to a gene in a model organism (such as mouse)
  - Can be applied multiple times for multiple gene testing
Issues in Genomewide Association Studies

- Fine Mapping
## Issues in Genomewide Association Studies

### Fine Mapping
- Studies conducted in a candidate region (1-10Mb) that might involve several hundred SNPs (spanning 5-50 genes)
• Fine Mapping
  • Studies conducted in a candidate region (1-10Mb) that might involve several hundred SNPs (spanning 5-50 genes)

• Genome Wide Association Studies
Issues in Genomewide Association Studies

- Fine Mapping
  - Studies conducted in a candidate region (1-10Mb) that might involve several hundred SNPs (spanning 5-50 genes)

- Genome Wide Association Studies
  - This study aims at identifying common causal variants throughout the genome
Issues in Genomewide Association Studies

- Fine Mapping
  - Studies conducted in a candidate region (1-10Mb) that might involve several hundred SNPs (spanning 5-50 genes)

- Genome Wide Association Studies
  - This study aims at identifying common causal variants throughout the genome
  - Requires at least 300,000 SNPs that are well chosen.

Sorin Istrail | Genome-Wide Association Studies (GWAS): Computational Themes and Caveats
Issues in Genomewide Association Studies

- **Fine Mapping**
  - Studies conducted in a candidate region (1-10Mb) that might involve several hundred SNPs (spanning 5-50 genes)

- **Genome Wide Association Studies**
  - This study aims at identifying common causal variants throughout the genome
  - Requires at least 300,000 SNPs that are well chosen.
  - More are needed for certain populations, such as African populations, due to greater genetic diversity
Fine Mapping

- Studies conducted in a candidate region (1-10Mb) that might involve several hundred SNPs (spanning 5-50 genes)

Genome Wide Association Studies

- This study aims at identifying common causal variants throughout the genome
- Requires at least 300,000 SNPs that are well chosen.
- More are needed for certain populations, such as African populations, due to greater genetic diversity
- Made possible by availability of SNPs in HAPMAP project and high throughput sequencing technology (Affymetrix, Illumina)
Four your final presentation, you should consider which association study characterizes your paper.
Four your final presentation, you should consider which association study characterizes your paper.

You should also consider whether other types of polymorphisms are taken into account, such as

- insertions
- deletions
- microsatellites
- copy number variation
Rationale for Association Studies

Population: difficulty in defining
biases in sampling
biases due to population substructure
Rationale for Association Studies

Population: difficulty in defining
biases in sampling
biases due to population substructure

Population association compares “unrelated” individuals, which means assuming the relationships are unknown and presumed to be distinct.
Rationale for Association Studies

Population: difficulty in defining biases in sampling biases due to population substructure

- Population association compares “unrelated” individuals, which means assuming the relationships are unknown and presumed to be distinct.

- Since phenotypes cannot be followed in population history, we instead focus on genotype information.
Rationale for Association Studies

Population: difficulty in defining biases in sampling biases due to population substructure

- Population association compares “unrelated” individuals, which means assuming the relationships are unknown and presumed to be distinct.

- Since phenotypes cannot be followed in population history, we instead focus on genotype information.

- The goal is to look for markers in the genotype.
Population: difficulty in defining biases in sampling biases due to population substructure

- Population association compares “unrelated” individuals, which means assuming the relationships are unknown and presumed to be distinct.
- Since phenotypes cannot be followed in population history, we instead focus on genotype information.
- The goal is to look for markers in the genotype.
- An increased density of markers corresponds to more information for use in association studies.
Difficulties in association studies arise because of linkage disequilibrium (LD) and recombination.
Difficulties in association studies arise because of linkage disequilibrium (LD) and recombination.

Because of the number of SNPs, all we can do is to hope to find indirect association.
Difficulties in association studies arise because of linkage disequilibrium (LD) and recombination.

Because of the number of SNPs, all we can do is to hope to find indirect association.

The hope is that by typing a dense set of markers, we will observe markers in direct association with unobserved causal locus, and in indirect association with disease phenotypes.
There are several major problems in the different stages of association studies. These include
There are several major problems in the different stages of association studies. These include:

- collection of data
There are several major problems in the different stages of association studies. These include:

- Collection of data
- Computational methods for phasing genotypes

Hardy-Weinberg equilibrium: significant deviation from HW needs to be addressed/scrutinized (carried out using Pearson $\chi^2$, or Fisher exact test if counts are low). Missing data (can be as high as 10-20%): EM method for inferring missing data requires assumptions on probability distribution.)

By varying the different techniques applied, obtain different association studies.
There are several major problems in the different stages of association studies. These include:

- Collection of data
- Computational methods for phasing genotypes
- Hardy-Weinberg equilibrium: significant deviation from HW needs to be addressed/scrutinized (carried out using Pearson $\chi^2$, or Fisher exact test if counts are low)
Issues in Genomewide Association Studies

- There are several major problems in the different stages of association studies. These include:
  - collection of data
  - computational methods for phasing genotypes
  - Hardy-Weinberg equilibrium: significant deviation from HW needs to be addressed/scrutinized (carried out using Pearson $\chi^2$, or Fisher exact test if counts are low)
  - Inference of phase
There are several major problems in the different stages of association studies. These include:

- Collection of data
- Computational methods for phasing genotypes
- Hardy-Weinberg equilibrium: significant deviation from HW needs to be addressed/scrutinized (carried out using Pearson $\chi^2$, or Fisher exact test if counts are low)
- Inference of phase
- Missing data (can be as high as 10-20%): EM method for inferring missing data requires assumptions on probability distribution
There are several major problems in the different stages of association studies. These include:

- Collection of data
- Computational methods for phasing genotypes
- Hardy-Weinberg equilibrium: significant deviation from HW needs to be addressed/scrutinized (carried out using Pearson $\chi^2$, or Fisher exact test if counts are low)
- Inference of phase
- Missing data (can be as high as 10-20%): EM method for inferring missing data requires assumptions on probability distribution
- By varying the different techniques applied, obtain different association studies
Most studies involve choices that are ad hoc. Therefore, it is important to address the question of robustness for the studies conducted.
Most studies involve choices that are ad hoc. Therefore, it is important to address the question of robustness for the studies conducted.

(1) Genome is so large that patterns that are suggestive of a causal polymorphism could well arise by chance.
Most studies involve choices that are ad hoc. Therefore, it is important to address the question of robustness for the studies conducted.

(1) Genome is so large that patterns that are suggestive of a causal polymorphism could well arise by chance

(2) Hardy-Weinberg, SNP selection, phasing are as yet unresolved methods
Most studies involve choices that are ad hoc. Therefore, it is important to address the question of robustness for the studies conducted.

1. Genome is so large that patterns that are suggestive of a causal polymorphism could well arise by chance
2. Hardy-Weinberg, SNP selection, phasing are as yet unresolved methods
3. Missing Data (The Missingness Problem)
Most studies involve choices that are ad hoc. Therefore, it is important to address the question of robustness for the studies conducted.

1. Genome is so large that patterns that are suggestive of a causal polymorphism could well arise by chance
2. Hardy-Weinberg, SNP selection, phasing are as yet unresolved methods
3. Missing Data (The Missingness Problem)
4. Case samples are often collected differently than controls, giving different rates of missingness. This is often a problem even if the study is carried out blind to the case-control status
(5) LD is a nonquantitative phenomenon, and there is no natural scale for it
(5) LD is a nonquantitative phenomenon, and there is no natural scale for it

(6) In practice, tagging is only effective in capturing common variants (tagging in one population only poorly performs in another
(5) LD is a nonquantitative phenomenon, and there is no natural scale for it.

(6) In practice, tagging is only effective in capturing common variants (tagging in one population only poorly performs in another).

(7) Population structure can generate spurious phenotype associations:
- Notion of subpopulation is a theoretical construct that only imperfectly reflects reality.
- The question of the correct number of subpopulations can never fully be resolved.
(5) LD is a nonquantitative phenomenon, and there is no natural scale for it

(6) In practice, tagging is only effective in capturing common variants (tagging in one population only poorly performs in another)

(7) Population structure can generate spurious phenotype associations
   - notion of subpopulation is a theoretical construct that only imperfectly reflects reality
   - the question of the correct number of subpopulations can never fully be resolved

(8) To phase or not to phase