Genomic-wide Association Studies

Ignacio Medina
Cristina Y. González

Genomic-wide Association Studies
Where are we?

Sequence preprocessing

Mapping

Variant Calling

Variant prioritization

Functional annotation

GWAS Analysis

Gene-Set Analysis

Ignacio Medina
Cristina Y. González

Genomic-Wide Association Studies
Index

• Introduction
• HPG Variant Suite
• Hands on
Introduction
Basics of Genetic Association Analysis

- Goal: to establish a statistical association between two variables: a disease trait and a genetic marker.

- Disease trait can be a dichotomous or quantitative measured variable.

- Genetic marker can be:
  - a known or suspected disease-causing mutation, or
  - a marker without any known effect on DNA (SNPs, …), in this case the association is created by Linkage Disequilibrium (LD) between the marker and disease allele.

- Two different study designs can be used:
  - Unrelated subjects, population study
  - Family studies

- Which is/are the genomic variant/s associated with my phenotype? Where is the disease locus located in the genome?
**Introduction**

**Classic GWAS I, technologies**

- Genotyping technology has made possible GWAS analysis, today we can genotype more than 1 million SNPs and Copy Number Variants with microarrays.

**Genotyping catalog**

**Affymetrix Genome-Wide Human SNP Array 6.0** features 1.8 million genetic markers, including more than 906,600 SNPs and more than 946,000 probes for the detection of copy number variation.

**Illumina Omni5** features more than 4.3 million high-value markers. And room for 500k of your own.

**Genotyping projects**

- **International HapMap Project**
- **1000 Genomes** A Deep Catalog of Human Genetic Variation

---

Ignacio Medina
Cristina Y. González

Genomic-Wide Association Studies

The International **HapMap** Project is a partnership of scientists to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals.


**PLINK** is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

http://pngu.mgh.harvard.edu/~purcell/plink

---

### Ignacio Medina

### Cristina Y. González

---

### Genomic-Wide Association Studies
Introduction

Classic GWAS results
Introduction
GWAS catalog

A Catalog of Published Genome-Wide Association Studies

http://www.genome.gov/gwastudies

Published GWA Reports, 2005 – 6/2012

Published Genome-Wide Associations through 07/2012
Published GWA at p≤5x10⁻⁸ for 18 trait categories

Ignacio Medina
Cristina Y. González

Genomic-Wide Association Studies
Introduction
What is a Haplotype?

- **A haplotype** is a sequence of alleles stretching along an extended segment of DNA – a sort of super allele!

a) Short stretch of DNA for 4 different people – 3 SNPs are present

b) Haplotypes made up of a combination of different alleles at 20 nearby SNPs

c) Genotyping just 3 “tag” SNPs can distinguish all 4 haplotypes

“A” mutation is linked to the red haplotype

Ignacio Medina
Cristina Y. González

Genomic-Wide Association Studies
Introduction
Linkage Disequilibrium (LD)

- **LD** is non-independence (nonrandomness) of alleles at different sites
- Example:
  - Suppose that allele A at locus 1 and allele B at locus 2 are at frequencies \(p_A\) and \(p_B\), respectively, in the population.
  - If the two loci are independent, then we would expect to see the \(AB\) haplotype at frequency \(p_A p_B\).
  - If the population frequency of the \(AB\) haplotype is either higher or lower than this - implying that particular alleles tend to be observed together - then the two loci are said to be in LD.
Two adjacent SNPs (A and B) or genetic markers are genotyped in a population.

**Under Linkage "Equilibrium" (LE)**

- $f_{AB} = f_{AfB}$
- $f_{aB} = f_{afB}$
- $f_{Ab} = f_{AfB}$
- $f_{Ab} = f_{AfB}$

**Under Linkage Disequilibrium (LD)**

- $f_{AB} = f_{AfB} + D$
- $f_{aB} = f_{afB} + D$
- $f_{Ab} = f_{AfB} + D$
- $f_{Ab} = f_{AfB} + D$

where $D$ is the LD coefficient:

- $D = f_{AB} \times f_{ab} - f_{aB} \times f_{Ab}$ or
- $D = f_{AB} - f_{A} \times f_{B}$

**Assessing LD:**
- $D' = D/D_{max}$
- $r^2$
Introduction
GWAS result and NGS

- Now we now sequenced all the variants, not only genotype some markers (SNPs)
- We can execute the statistical test to see if a variant is associated with a phenotype
  - Chi-square and Regression for population studies
  - Transmission Disequilibrium Test (TDT) for families based studies
- A variant can still be a marker if causal mutated variant is not properly captured or sequenced
- In multi factorial diseases is harder to find causal variants
Introduction

Drawbacks

Variants are considered independently. However, complex phenotypes are expected to be induced by different genes in the same functional module. A different strategy must be taken: Methodologies based on **Gene-Set Analysis or networks** permit the study of functional modules (group of genes that cooperate to carry out a biological function).

The cases of the **multifactorial disease** will have different mutations (or combinations). Many cases have to be used to obtain significant associations to many markers. The only common element is the pathway (unknown at this moment) affected.
HPG Variant is a suite consisting of 3 applications
You have already used the Effect annotation tool
2 more applications available:
  - VCF tools: For VCF files preprocessing
  - GWAS: For genomic-wide association studies
HPG Variant Suite
HPG Variant VCF tools

• **HPG Variant VCF** handles files containing information about genomic variants

• As fast and efficient as possible → scientists can focus on experiments, not dataset cleanup!

• Based on a publicly available library (**vcf-lib**), so you can use it for your own applications :-)

• But... what does it do exactly?
With HPG Variant VCF, you can:

- Retrieve statistics about a file (for instance, to find out the allele frequencies or the quality of a file)
- Merge several files into one (for example, if they belong to the same experiment)
- Split a file into multiple ones (to analyze only a small part of it)
- Filter a file (to remove the records that don't meet certain requirements)
Non-suitable datasets could distort your results

Some statistical tests can be biased by missing variants or samples, it is important to “clean” the dataset
Conducts association studies from 2 points of view:
- Population
- Family

Population-based studies only consider individuals' phenotypes

Family-based studies only check families (relationships should meet certain conditions)
 Population based-studies calculate the value of a statistical distribution (chi-square, Fisher's exact test)

 Family-based studies calculate a p-value based on other criteria (transmission disequilibrium test a.k.a TDT)
Hands on
Downloads and exercises set up

• Follow the HPG Variant **Getting started tutorial**
  – Download the datasets from that website

• HPG Variant is available as:
  – Package for your favorite distribution
  – Compressed executable files (Debian 6 / Ubuntu 10.04 or greater)
  – Source files
Thanks for your attention

Any questions?