Recent Developments in Genome-Wide Association Studies

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

June 3, 2009

Karyotypes

Single Nucleotide Polymorphisms

Haplotypes

### Haplotypes


<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Description</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CAGATCGCTGAATGAAATCGCATCTGT</td>
<td>35%</td>
</tr>
<tr>
<td>2</td>
<td>CAGATCGCTGAATGGATCCCATCAGT</td>
<td>30%</td>
</tr>
<tr>
<td>3</td>
<td>CGGATTGCTGCATGGATCCCATCAGT</td>
<td>15%</td>
</tr>
<tr>
<td>4</td>
<td>CGGATTGCTGCATGAATCGCATCTGT</td>
<td>10%</td>
</tr>
<tr>
<td>Others</td>
<td>Several other haplotypes</td>
<td>[10%]</td>
</tr>
</tbody>
</table>

### Haplotype Blocks

#### Kim and Dionne (2007)
A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping


Table 1
Estimated coverage of commercially available fixed marker genotyping platforms

<table>
<thead>
<tr>
<th>Platform</th>
<th>HapMap population sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>YRI</td>
</tr>
<tr>
<td>Affymetrix GeneChip 500K</td>
<td>46</td>
</tr>
<tr>
<td>Affymetrix SNP Array 6.0</td>
<td>66</td>
</tr>
<tr>
<td>Illumina HumanHap300</td>
<td>33</td>
</tr>
<tr>
<td>Illumina HumanHap550</td>
<td>55</td>
</tr>
<tr>
<td>Illumina HumanHap650Y</td>
<td>66</td>
</tr>
<tr>
<td>Perlegen 600K</td>
<td>47</td>
</tr>
</tbody>
</table>

Data represent percent of SNPs tagged at $r^2 \geq 0.8$. Values assume all SNPs on the platform are informative and pass quality control. YRI, Yoruba in Ibadan, Nigeria; CEU, subsample of Utah residents of Northern European ancestry selected from Centre d’Etude du Polymorphisme Humain samples; CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo. From the International HapMap Consortium, 2007 (3).


---

Review Papers

How to Interpret a Genome-wide Association Study

Thomas A. Pearson, MD, MPH, PhD
Terry A. Manolio, MD, PhD

Genome-wide association (GWA) studies use high-throughput genotyping technologies to analyze hundreds of thousands of single-nucleotide polymorphisms (SNPs) and relate them to clinical conditions and measurable traits. Since 2005, nearly 100 loci for as many as 40 common diseases and traits have been identified and replicated in GWA studies. Many of these genes had not previously been suspected of having a role in the disease under study, and some in genomic regions containing no known genes. GWA studies are a major advance in discovering genetic variants influencing disease but also have important limitations, including their potential for false-positive and false-negative results and for biases related to selection of study participants and genotyping errors. Although these studies are clearly many steps removed from actual clinical use, and specific applications of GWA findings in prevention and treatment are actively being pursued, at present these studies mainly represent a valuable discovery tool for examining genetic function and derailing pathophysiologic mechanisms. This article describes the design, interpretation, application, and limitations of GWA studies for clinicians and scientists for whom this evolving science may have great relevance.

Review Papers

A tutorial on statistical methods for population association studies

David J. Balding

Abstract | Although genetic association studies have been with us for many years, even for the simplest analyses there is little consensus on the most appropriate statistical procedures. Here I give an overview of statistical approaches to population association studies, including preliminary analyses (Hardy–Weinberg equilibrium testing, inference of phase and missing data, and SNP tagging), and single-SNP and multipoint tests for association. My goal is to outline the key methods with a brief discussion of problems (population structure and multiple testing), avenues for solutions and some ongoing developments.


Terminology

- **Alleles** Alternate forms of a gene or chromosomal locus that differ in DNA sequence
- **Candidate gene** A gene believed to influence expression of complex phenotypes due to known biological and/or physiological properties of its products, or to its location near a region of association or linkage
- **Copy number variants** Stretches of genomic sequence of roughly 1 kb to 3 Mb in size that are deleted or are duplicated in varying numbers
- **False discovery rate** Proportion of significant associations that are actually false positives
- **Functional studies** Investigations of the role or mechanism of a genetic variant in causation of a disease or trait
- **Gene-environment interactions** Modification of gene-disease associations in the presence of environmental factors

Terminology

- **Genome-wide association study** Any study of genetic variation across the entire human genome designed to identify genetic association with observable traits or the presence or absence of a disease, usually referring to studies with genetic marker density of 100,000 or more to represent a large proportion of variation in the human genome.

- **Genotyping call rate** Proportion of samples or SNPs for which a specific allele SNP can be reliably identified by a genotyping method.

- **Haplotype** A group of specific alleles at neighboring genes or markers that tend to be inherited together.

- **HapMap** Genome-wide database of patterns of common human genetic sequence variation among multiple ancestral population samples.


---

Terminology

- **Hardy Weinberg equilibrium** Population distribution of 2 alleles (with frequencies p and q) such that the distribution is stable from generation to generation and genotypes occur at frequencies of $p^2$, $2pq$, and $q^2$ for the major allele homozygote, heterozygote, and minor allele homozygote, respectively.

- **Linkage disequilibrium** Association between 2 alleles located near each other on a chromosome, such that they are inherited together more frequently than expected by chance.

- **Mendelian disease** Condition caused almost entirely by a single major gene, such as cystic fibrosis or Huntington’s disease, in which disease is manifested in only 1 (recessive) or 2 (dominant) of the 3 possible genotype groups.

- **Minor allele** The allele of a biallelic polymorphism that is less frequent in the study population.

Terminology

- **Minor allele frequency** Proportion of the less common of 2 alleles in a population (with 2 alleles carried by each person at each autosomal locus) ranging from less than 1% to less than 50%

- **Non-Mendelian disease** (also "common" or "complex" disease) Condition influenced by multiple genes and environmental factors and not showing Mendelian inheritance patterns

- **Nonsynonymous SNP** A polymorphism that results in a change in the amino acid sequence of a protein (and therefore may affect the function of the protein)

- **Platform** Arrays or chips on which high-throughput genotyping is performed


---

Terminology

- **Polymorphic** A gene or site with multiple allelic forms. The term polymorphism usually implies a minor allele frequency of at least 1%

- **Population stratification** (also "population structure") A form of confounding in genetic association studies caused by genetic differences between cases and controls unrelated to disease but due to sampling them from populations of different ancestries

- **Single-nucleotide polymorphism** Most common form of genetic variation in the genome, in which a single-base substitution has created 2 forms of a DNA sequence that differ by a single nucleotide

- **Tag SNP** A readily measured SNP that is in strong linkage disequilibrium with multiple other SNPs so that it can serve as a proxy for these SNPs on large-scale genotyping platforms

Case-control Design


Family-Based Designs

**Linkage and Association**

**Linkage analysis**
A method for localizing genes that is based on the co-inheritance of genetic markers and phenotypes in families over several generations.

**Association studies**
A gene-discovery strategy that compares allele frequencies in cases and controls to assess the contribution of genetic variants to phenotypes in specific populations.

Linkage versus Association

Rare, high penetrance mutations – use linkage

Common, low penetrance variants – use association

Frequency in population

Family-Based Designs

Family versus case-control

The main differences (IMHO):

- **Case-control designs often have more power (per genotype).**
- **Case-control studies are often cheaper, and simpler to conduct.**
- **Family based designs are immune to population stratification.**
- **Family based designs can test for linkage and association.**
Quality Control

- Samples
  - Genotype uncertainty / missingness / DNA quality
  - DNA quality cases versus controls
  - Contamination
  - Gender mismatch
  - Klinefelter, Turner syndrome
  - Relatedness of individuals
  - Genetic background / population stratification

- SNPs
  - Troublemakers
  - Minor allele frequency
  - Hardy-Weinberg equilibrium

Population Stratification

\[ \text{Balding DJ (2006). Nat Rev Genet. 7(10): 781-91.} \]
Principal components analysis corrects for stratification in genome-wide association studies

Allkes L. Price1,2, Nick J Patterson3, Robert M Plenge2,3, Michael E Weinblatt4, Nancy A Shadick5 & David Reich1,2

Population stratification—allele frequency differences between cases and controls due to systematic ancestry differences—causes spurious associations in disease studies. We describe a method that enables explicit detection and correction of population stratification on a genome-wide scale. Our method uses principal components analysis to explicitly model ancestry differences between cases and controls. The resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. Our simple, efficient approach can easily be applied to disease studies with hundreds of thousands of markers.

**Principal Components**


---

**Ingo Ruczinski**

**Developments in GWAs**

---

**Principal Components**


---

**Ingo Ruczinski**

**Developments in GWAs**
PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham

Whole-genome association studies (GWAS) bring new computational, as well as analytic, challenges to researchers. Many existing genetic-analysis tools are not designed to handle such large data sets in a convenient manner and do not necessarily exploit the new opportunities that whole-genome data bring. To address these issues, we developed PLINK, an open-source C/C++ GWAS tool set. With PLINK, large data sets comprising hundreds of thousands of markers genotyped for thousands of individuals can be rapidly manipulated and analyzed in their entirety. As well as providing tools to make the basic analytic steps computationally efficient, PLINK also supports some novel approaches to whole-genome data that take advantage of whole-genome coverage. We introduce PLINK and describe the five main domains of function: data management, summary statistics, population stratification, association analysis, and identity-by-descent estimation. In particular, we focus on the estimation and use of identity-by-state and identity-by-descent information in the context of population-based whole-genome studies. This information can be used to detect and correct for population stratification and to identify extended chromosomal segments that are shared identical by descent between very distinctly related individuals. Analysis of the patterns of segmental sharing has the potential to map disease loci that contain multiple rare variants in a population-based linkage analysis.


ARTICLES


Reality Often Sucks
Results

Science in medicine

A HapMap harvest of insights into the genetics of common disease

Teri A. Manolio, Lisa D. Brookes, and Francis S. Collins
National Human Genome Research Institute, Bethesda, Maryland, USA.

The International HapMap Project was designed to create a genome-wide database of patterns of human genetic variation, with the expectation that these patterns would be useful for genetic association studies of common diseases. This expectation has been amply fulfilled with just the initial output of genome-wide association studies, identifying nearly 100 loci for nearly 40 common diseases and traits. These associations provided new insights into pathophysiology, suggesting previously unsuspected etiologic pathways for common diseases that will be of use in identifying new therapeutic targets and developing targeted interventions based on genetically defined risk. In addition, HapMap-based discoveries have shed new light on the impact of evolutionary pressures on the human genome, suggesting multiple loci important for adapting to disease-causing pathogens and new environments. In this review we examine the origin, development, and current status of the HapMap; its prospects for continued evolution; and its current and potential future impact on biomedical science.


Ingo Ruczinski Developments in GWAs

Results


Ingo Ruczinski Developments in GWAs
Catalogued findings for the following phenotypes:

- Eye diseases
- Diabetes
- Cancer
- Gastrointestinal disorders
- Cardiovascular conditions and lipid metabolism
- Neuropsychiatric conditions
- Autoimmune and infectious diseases
- Various traits


http://www.genome.gov/GWAstudies/
Results

Phenotypes under investigation in collaborative GWA studies

<table>
<thead>
<tr>
<th>GAIN⁰</th>
<th>GE²</th>
<th>STAMPEDE²</th>
<th>CGEMS³</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention deficit hyperactivity disorder</td>
<td>Type 2 diabetes</td>
<td>Early-onset myocardial infarction</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td>Major depressive disorder</td>
<td>Maternal metabolism and birth weight</td>
<td>Asthma</td>
<td>Breast cancer</td>
</tr>
<tr>
<td>Bipolar I disorder</td>
<td>Preterm birth</td>
<td>Platelet phenotypes</td>
<td>Pancreatic cancer</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>Oral clefts</td>
<td>CHD and other heart, lung and blood disorders</td>
<td>Lung cancer</td>
</tr>
<tr>
<td>Type 1 diabetic nephropathy</td>
<td>Dental caries</td>
<td>Childhood respiratory outcomes</td>
<td>Bladder cancer</td>
</tr>
<tr>
<td>Psoriasis</td>
<td>Coronary disease</td>
<td>Hematopoietic cell transplant outcome</td>
<td>Renal cancer</td>
</tr>
<tr>
<td></td>
<td>Lung cancer</td>
<td>Arteriosclerosis in hypertensives</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Addiction</td>
<td>Asthma and lung function</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cardiovascular risk factors</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Atherosclerosis pathway genes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CV events</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Early coronary artery disease</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phenotypic variability in sickle cell anemia</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Centenarians</td>
<td></td>
</tr>
</tbody>
</table>


Imputation

A comprehensive evaluation of SNP genotype imputation

Michael Nothnagel · David Ellinghaus · Stefan Schreiber · Michael Krinzik · Andre Franke

Abstract: Genome-wide association studies have contributed significantly to the genetic dissection of complex diseases. In order to increase the power of existing studies, even further, methods have been proposed to predict individual genotype on an typed list from other markers by employing high-density data as a reference. Although various imputation algorithms have been used in practice already, a comprehensive evaluation and comparison of these approaches, using genome-wide SNP data from one and the same population is still lacking. We therefore investigated four publicly available programs for genotype imputation (BEAGLE, IMPUTE, MACH, and PLINK) using data from 449 German individuals genotyped in our laboratory for their genome-wide SNP sets. In this study, BEAGLE and IMPUTE performed consistently better. We nevertheless recommend either MACH or BEAGLE for practical use because these two programs are more user-friendly and generally require less memory than IMPUTE.

Imputation


What’s missing?

1. Phenotype delineation
2. Genotypes (unobserved, mis-typed)
3. Copy number variants
4. Genetic models (complex disease, biological interactions)
5. Epigenetics and such
6. Environmental data

Ingo Ruczinski  Developments in GWAs

Trisomy

http://www.1000genomes.org
Karyotypes

General Cytogenetics Information  http://members.aol.com/chrominfo/

FISH

Courtesy of the Pevsner Laboratory
Amplification

Uniparental Isodisomy
Cancer samples

Mosaicism
From Forrest Spencer (unpublished).


From Benilton Cavalho
Genotype call comparison

BRLMM-P  CRLMM  BirdSeed

From Rafa Irizarry

http://biostat.jhsph.edu/~iruczins/