



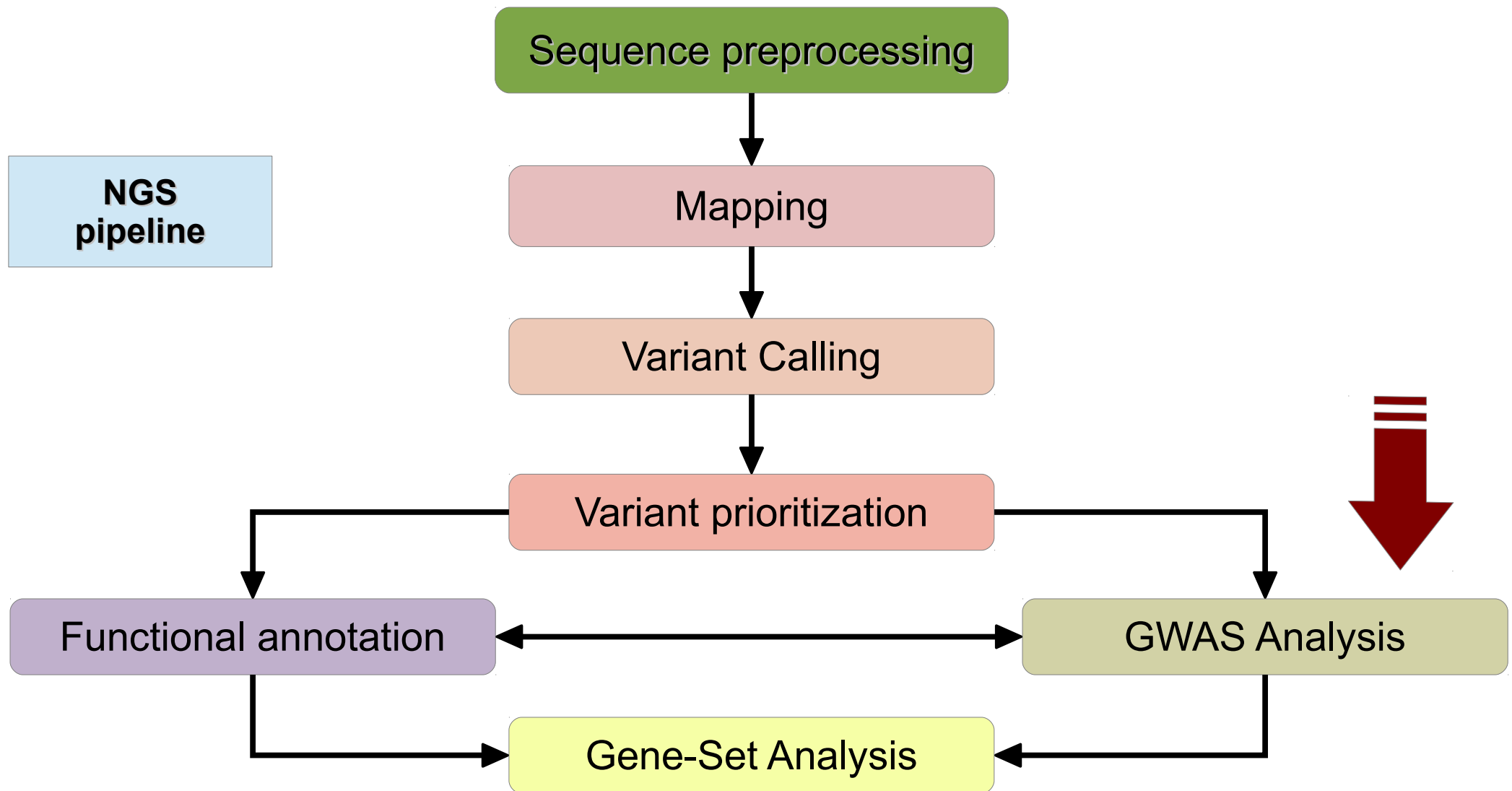
# **IX** International Course of **Massive Data Analysis** **FOR GENOMICS**



Ignacio Medina  
Cristina Y. González

Genomic-wide Association Studies

# Where are we?



# 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



**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 catalog

<http://www.ncbi.nlm.nih.gov/projects/SNP>

The screenshot shows the NCBI dbSNP website. At the top, there's a navigation bar with links to PubMed, Nucleotide, Protein, Genome, Structure, PopSet, Taxonomy, OMIM, Books, and SNP. Below this is a search bar with the text "Search for SNP on NCBI Reference Assembly". A sidebar on the left contains a "Have a question about dbSNP? Try searching the SNP FAQ Archive!" link. The main content area features an "ANNOUNCEMENT" for "09/20/2012: NCBI dbSNP Build 137 Mouse and Cow Release". It states that dbSNP Mouse\_10090 and Cow\_9913 (Bos taurus) Build 137 data are now available, including Mouse (tax\_id 10900) and Mm\_Celera (GCF\_000001635.20) and Mm\_Celera (GCF\_000002165.2). Below the announcement is a "Search by IDs on All Assemblies" section with a note that rs# and ss# must be prefixed with "rs" or "ss", respectively. There are input fields for ID and Reference cluster ID(rs#), and Search and Reset buttons.

## Genotyping projects

The screenshot shows the International HapMap Project website. It features a header with the project name and a world map. Below the header, there's a "Project Information" section with links to "About the Project", "HapMap Publications", "HapMap Tutorial", "HapMap Mailing List", and "HapMap Project Participants". There's also a "Project Data" section with links to "HapMap3 Genome Browser release #29 (Phases 1, 2 & 3 - merged genotypes & frequencies)", "HapMap3 Genome Browser release #27 (Phase 3 - genotypes & frequencies)", "HapMap3 Genome Browser release #27 (Phase 1, 2 & 3 - merged genotypes & frequencies)", "HapMap3 Genome Browser release #2 (Phase 3 - genotypes, frequencies & LD)", and "HapMap3 Genome Browser release #24 (Phase 1 & 2 - full dataset)". A "News" section on the right lists several updates, including a "HapMap help desk announcement" from 2011-06-13, a "HapMap help desk service interruption notice" from 2011-04-20, and a "HapMap phase II recombination rate on GRCh37" from 2011-01-19.

The screenshot shows the 1000 Genomes Project website. It features a header with the project name and a world map. Below the header, there's a "Project Information" section with links to "About the Project", "HapMap Publications", "HapMap Tutorial", "HapMap Mailing List", and "HapMap Project Participants". There's also a "Project Data" section with links to "HapMap3 Genome Browser release #29 (Phases 1, 2 & 3 - merged genotypes & frequencies)", "HapMap3 Genome Browser release #27 (Phase 3 - genotypes & frequencies)", "HapMap3 Genome Browser release #27 (Phase 1, 2 & 3 - merged genotypes & frequencies)", "HapMap3 Genome Browser release #2 (Phase 3 - genotypes, frequencies & LD)", and "HapMap3 Genome Browser release #24 (Phase 1 & 2 - full dataset)". A "News" section on the right lists several updates, including a "HapMap help desk announcement" from 2011-06-13, a "HapMap help desk service interruption notice" from 2011-04-20, and a "HapMap phase II recombination rate on GRCh37" from 2011-01-19.

Ignacio Medina  
Cristina Y. González

# Genomic-Wide Association Studies

# Introduction

## Classic GWAS II, resources

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

<http://hapmap.ncbi.nlm.nih.gov>



The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "[About the International HapMap Project](#)" for more information.

Project Information	News
<a href="#">About the Project</a> <a href="#">HapMap Publications</a> <a href="#">HapMap Tutorial</a> <a href="#">HapMap Mailing List</a> <a href="#">HapMap Project Participants</a>	<ul style="list-style-type: none"><li>2011-06-13: <b>HapMap help desk announcement</b> There was a problem with the HapMap help desk system. In the past several weeks, emails sent to hapmap-help@ncbi.nlm.nih.gov not reach the help desk, and thus user requests were not addressed. Please resend your email request if you sent email to HapMap help desk in the past several weeks. Sorry for the inconvenience.</li><li>2011-04-20: <b>Hapmap help desk service interruption notice</b> There will be no help desk support from 05/03/2011 to 05/23/2011. Sorry for the inconvenience.</li><li>2011-02-02: <b>Haploview issues with rel 28 data</b> Recently, there are several questions about Haploview data format errors when users tried to analyze HapMap release 28 current Haploview version (4.2) does not recognize the new individuals in release 28 and the software will generate an error "Hapmap data format error: NA18876" when trying to open the data. Haploview is developed and maintained by an organization different from HapMap. Please contact Haploview help desk (haploview@broadinstitute.org) for questions specific to this software.</li><li>2011-01-19: <b>HapMap phase II recombination rate on GRCh37</b> The liftover of the HapMap II genetic map from human genome build b35 to GRCh37 is available. Data is <a href="#">available for download</a>.</li><li>2010-08-18: <b>HapMap Public Release #28</b></li></ul>

**Project Data**  
HapMap Genome Browser release #28 ( Phases 1, 2 & 3 - merged genotypes & frequencies )  
HapMap3 Genome Browser release #3 ( Phase 3 - genotypes & frequencies )  
HapMap Genome Browser release #27 ( Phase 1, 2 & 3 - merged genotypes & frequencies )  
HapMap3 Genome Browser release #2 ( Phase 3 - genotypes, frequencies & LD )  
HapMap Genome Browser release #24 ( Phase 1 & 2 - full dataset )  
GWAS Karyogram

**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>

**plink...**  
Whole genome association analysis toolset

Latest PLINK release is v1.07 (10-Oct-2009)

[Introduction](#) | [Basics](#) | [Download](#) | [Reference](#) | [Formats](#) | [Data management](#) | [Summary stats](#) | [Filters](#) | [Stratification](#) | [IBS/IBD](#) | [Association](#) | [Family-based](#) | [Permutation](#) | [LD calculations](#) | [Haplotypes](#) | [Conditional tests](#) | [Proxy association](#) | [Imputation](#) | [Dosage data](#) | [Meta-analysis](#) | [Result annotation](#) | [Clumping](#) | [Gene Report](#) | [Epistasis](#) | [Rare CNVs](#) | [Common CNVs](#) | [R-plugins](#) | [SNP annotation](#) | [Simulation](#) | [Profiles](#) | [ID helper](#) | [Resources](#) | [Flow chart](#) | [Misc](#) | [FAQ](#) | [gPLINK](#)

**1. Introduction**  
**2. Basic information**

- Citing PLINK
- Reporting problems
- What's new?
- PDF documentation

**3. Download and general notes**

- Stable download
- Development code
- General notes
- MS-DOS notes
- Unix/Linux notes
- Compilation
- Using the command line
- Viewing output files
- Version history

**4. Command reference table**

- List of options
- List of output files
- Under development

**5. Basic usage/data formats**

- Running PLINK
- PED files
- MAP files
- Transposed files
- Long-format files
- Binary PED files
- Alternate phenotypes
- Covariate files
- Cluster files
- Set files

**6. Data management**

- Recode
- Reorder
- Write SNP list
- Update SNP map
- Update allele information
- Force reference allele

**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.  
The focus of **PLINK** is purely on *analysis* of genotype/phenotype data, so there is no support for steps prior to this (e.g. study design and planning, generating genotype or CNV calls from raw data). Through integration with gPLINK and Haploview, there is some support for the subsequent visualization, annotation and storage of results.  
**PLINK** (one syllable) is being developed by Shaun Purcell at the Center for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH), and the Broad Institute of Harvard & MIT, with the support of others.  
**New in 1.07:** meta-analysis, result annotation and analysis of dosage data.

**Quick links**  
[PLINK tutorial](#)  
[gPLINK](#)  
[Join e-mail list](#)  
[Resources](#)  
[FAQs](#) | [PDF](#)  
[Citing PLINK](#)  
[Bugs, questions?](#)

**Data management**

- Read data in a variety of formats
- Recode and reorder files
- Merge two or more files
- Extract subsets (SNPs or individuals)
- Flip strand of SNPs
- Compress data in a binary file format

**Summary statistics for quality control**

- Allele, genotypes frequencies, HWE tests
- Missing genotype rates
- Inbreeding, IBS and IBD statistics for individuals and pairs of individuals
- non-Mendelian transmission in family data
- Sex checks based on X chromosome SNPs
- Tests of non-random genotyping failure

**Population stratification detection**

Ignacio Medina  
Cristina Y. González

# Genomic-Wide Association Studies



# Introduction

## Classic GWAS results

### THE LANCET Neurology

Search for  in  All Fields

[Home](#) | [Journals](#) | [Specialties](#) | [Clinical](#) | [Global Health](#) | [Multimedia](#) | [Conferences](#) | [Information for](#)

The Lancet Neurology, [Volume 7, Issue 11](#), Pages 1067 - 1072, November 2008  
doi:10.1016/S1474-4422(08)70241-2

This article can be found in the following collections: [Genetics & Genomics](#); [Neurology \(Genetics & neurology\)](#)

## Genome-wide association studies in neurological disorders

[Javier Simón-Sánchez](#) MS a b, [Andrew Singleton](#) PhD c d

### Summary

#### Background

During the past decade, the genetic causes of monogenic forms of disease have been successfully defined; this work has helped the progression of basic scientific investigation into many disorders, and has helped to characterise several molecular biological processes. An important goal of genetic research is to extend this work and define genetic risk factor loci for complex disorders. The aim is for these data not only to offer further basic understanding of the disease process, but also to provide the opportunity

NCBI Resources ☒ How To ☒

**PubMed.gov**

[Display Settings:](#) ☒ Abstract

[Send to:](#) ☒

[Nature](#), 2007 Jun 7;447(7145):661-78.

## Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.

[Wellcome Trust Case Control Consortium](#).

[Collaborators \(258\)](#)

#### Abstract

There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We describe a joint GWA study (using the Affymetrix GeneChip 500K Mapping Array Set) undertaken in the British population, which has examined approximately 2,000 individuals for each of 7 major diseases and a shared set of approximately 3,000 controls. Case-control comparisons identified 24 independent association signals at  $P < 5 \times 10^{-7}$ : 1 in bipolar disorder, 1 in coronary artery disease, 9 in Crohn's disease, 3 in rheumatoid arthritis, 7 in type 1 diabetes and 3 in type 2 diabetes. On the basis of prior findings and replication studies thus far completed, almost all of these signals reflect genuine susceptibility effects. We observed association at many previously identified loci, and found compelling evidence that some loci confer risk for more than one of the diseases studied. Across all diseases, we identified a large number of further signals (including 58 loci with single-point  $P$  values between  $10^{-5}$  and  $5 \times 10^{-7}$ ) likely to yield additional susceptibility loci. The importance of appropriately large samples was confirmed by the modest effect sizes observed at most loci identified. This study thus represents a thorough validation of the GWA approach. It has also demonstrated that careful use of a shared control group represents a safe and effective approach to GWA analyses of multiple disease phenotypes; has generated a genome-wide genotype database for future studies of common diseases in the British population; and shown that, provided that European ancestry is excluded, the extent of population stratification in the British population database offers new avenues for exploring the pathophysiology of these important disorders. We anticipate that the results and software, which will be widely available to other investigators, will provide a powerful resource for genetic research.

**nature**

**FREE** Author Manuscript in PubMed Central

**Save items**

**Related citations in PubMed**

Genomics: guilt by association. [Nature. 2007]

Final Report on Carcinogens Backg [Rep Carcinog Backgr Doc. 2010]

Genome-wide association study of CNVs in 16,000 cases of [Nature. 2010]

[Review](#) New IBD genetics: common pathways with other disease: [Gut. 2011]

[Review](#) Genome-wide association scans identify [Inflamm Bowel Dis. 2007]

Matches in page for gwas nature wtcc

**Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Wellcome Trust Case Control...**

[More matches >>](#)

### THE LANCET

Search for  in  All Fields

[Home](#) | [Journals](#) | [Specialties](#) | [Clinical](#) | [Global Health](#) | [Multimedia](#) | [Conferences](#) | [Information for](#)

The Lancet, [Volume 380, Issue 9844](#), Pages 815 - 823, 1 September 2012  
doi:10.1016/S0140-6736(12)60681-3

This article can be found in the following collection: [Genetics & Genomics](#)  
Published Online: 03 July 2012

[< Previous Article](#) | [Next Article >](#)

## Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study

arcOGEN Consortium and arcOGEN Collaborators<sup>‡</sup>

### Summary

#### Background

Osteoarthritis is the most common form of arthritis worldwide and is a major cause of pain and disability in elderly people. The health economic burden of osteoarthritis is increasing commensurate with obesity prevalence and longevity. Osteoarthritis has a high prevalence in populations with efficient sample sizes and

### THE LANCET

Search for  in  All Fields

[Home](#) | [Journals](#) | [Specialties](#) | [Clinical](#) | [Global Health](#) | [Multimedia](#) | [Conferences](#) | [Information for](#)

The Lancet, [Volume 377, Issue 9766](#), Pages 641 - 649, 19 February 2011  
doi:10.1016/S0140-6736(10)62345-8

This article can be found in the following collections: [Neurology \(Dementias\)](#)  
Published Online: 02 February 2011

[< Previous Article](#) | [Next Article >](#)

## Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies

International Parkinson Disease Genomics Consortium<sup>‡</sup>

### Summary

#### Background

Genome-wide association studies (GWAS) for Parkinson's disease have linked two loci (*MAPT* and *SNCA*) to risk of Parkinson's disease. We aimed to identify novel risk loci for Parkinson's disease.

#### Methods

We did a meta-analysis of datasets from five Parkinson's disease GWAS from the USA and Europe to identify loci associated with Parkinson's disease (discovery phase). We then did replication analyses of significantly associated loci in an independent sample series. Estimates of population-attributable risk were calculated from estimates from the discovery and replication phases combined, and risk-profile estimates for loci identified in the discovery phase were calculated.

#### Findings

The discovery phase consisted of 5333 case and 12 019 control samples, with genotyped and imputed data at 7 689 524 SNPs. The replication phase consisted of 7053 case and 9007 control samples. We identified 11 loci that surpassed the threshold for genome-wide significance ( $p < 5 \times 10^{-8}$ ). Six were previously identified loci (*MAPT*, *SNCA*, *HLA-DRB5*, *BST1*, *GAK* and *LRRK2*) and five were newly identified loci (*ACMSD*, *STK39*, *MC6CC1/LAMP3*, *SYT11*, and *CCDC62/HIP1R*). The combined population-attributable risk was 60.3% (95% CI 43.7–69.3). In the risk-profile analysis, the odds ratio in the highest quintile of disease risk was 2.51 (95% CI 2.23–2.83) compared with 1.00 in the lowest quintile of disease risk.

Ignacio Medina  
Cristina Y. González

# Genomic-Wide Association Studies

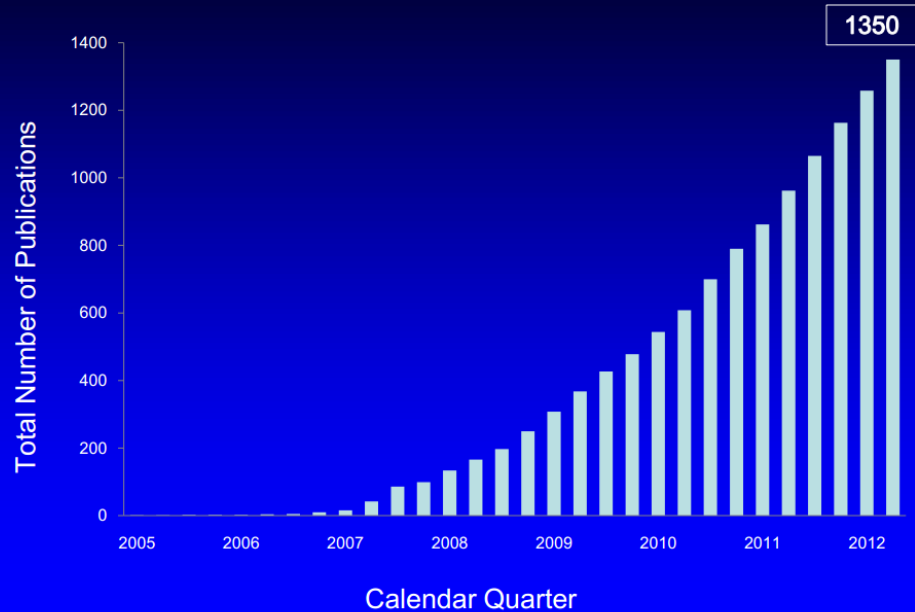
# Introduction

## GWAS catalog

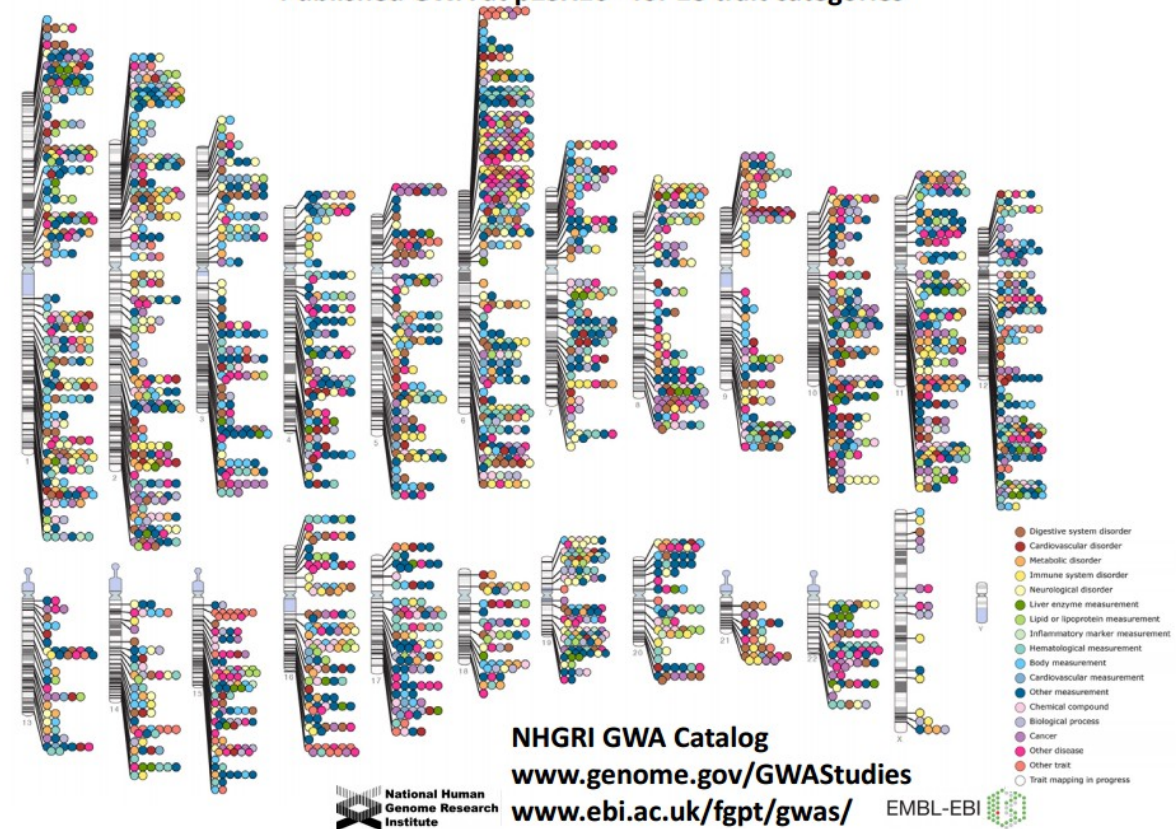
A Catalog of Published Genome-Wide Association Studies

<http://www.genome.gov/gwastudies>

### Published GWA Reports, 2005 – 6/2012



Published Genome-Wide Associations through 07/2012  
Published GWA at  $p \leq 5 \times 10^{-8}$  for 18 trait categories



Ignacio Medina  
Cristina Y. González

# Genomic-Wide Association Studies

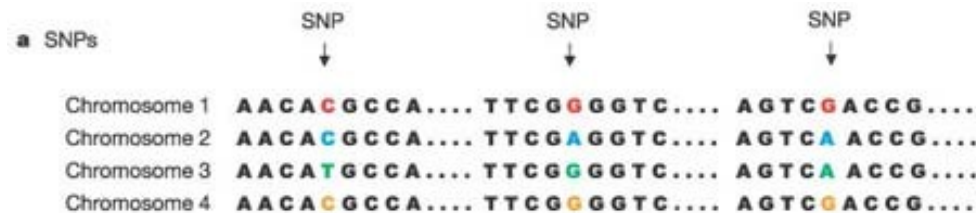


# Introduction

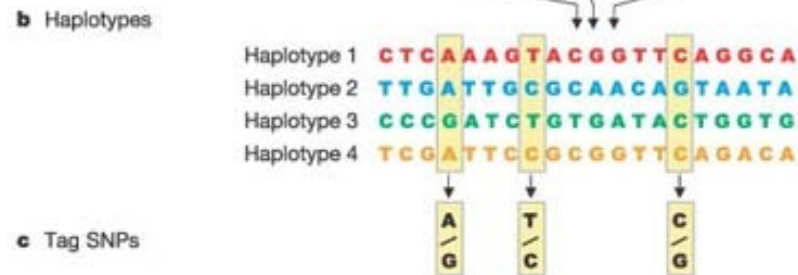
## What is a Haplotype?

- 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

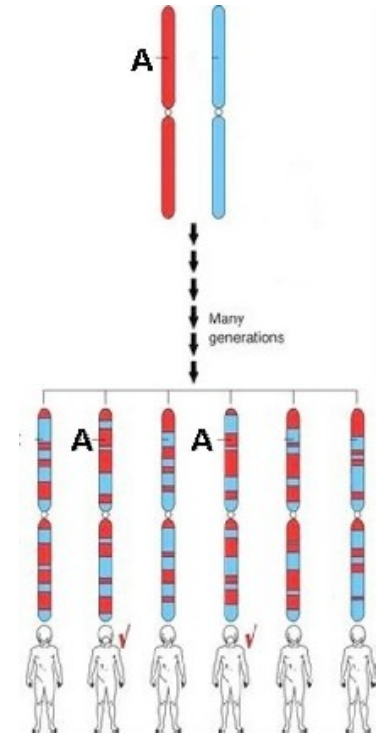


b) Haplotypes made up of a **combination** of different alleles at 20 nearby SNPs



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

“A” mutation is linked to the **red** haplotype



# 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.

# Introduction

## Linkage Disequilibrium (LD)

Two adjacent SNPs (A and B) or genetic markers are genotyped in a population.

There are 4 possible haplotypes

		SNP 1		
		A	a	
SNP 2	B	$f_{AB}$	$f_{aB}$	$f_B$
	b	$f_{Ab}$	$f_{ab}$	$f_b$
		$f_A$	$f_a$	



### Under Linkage “Equilibrium” (LE)

$$f_{AB} = f_A f_B$$

$$f_{aB} = f_a f_B$$

$$f_{Ab} = f_A f_b$$

$$f_{ab} = f_a f_b$$

### Under Linkage Disequilibrium (LD)

$$f_{AB} = f_A f_B + D$$

$$f_{aB} = f_a f_B + D$$

$$f_{Ab} = f_A f_b + D$$

$$f_{ab} = f_a f_b + D$$

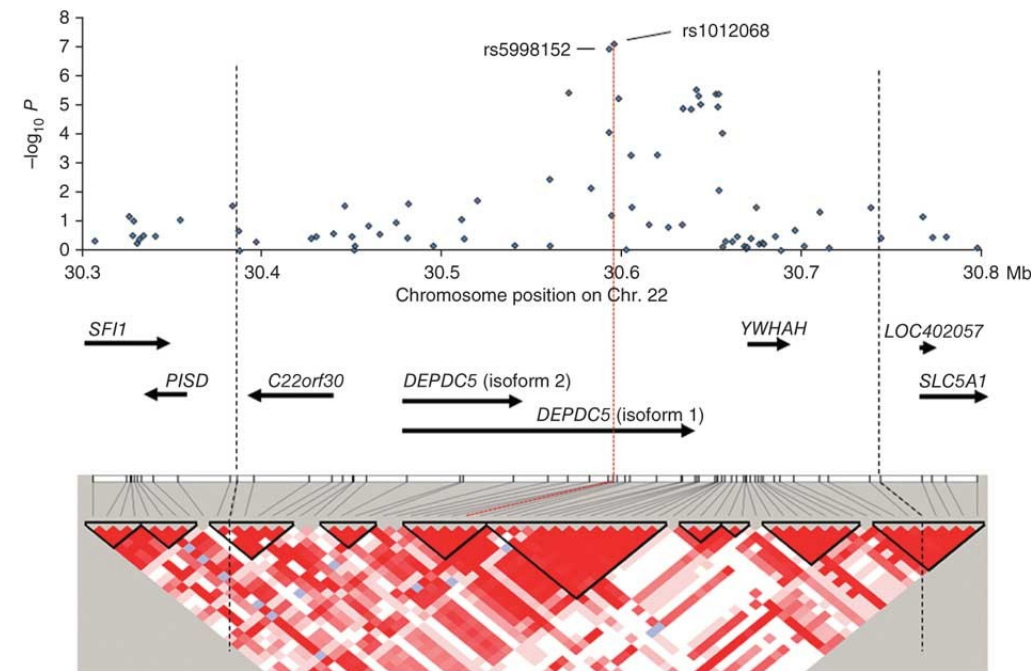
where  $D$  is the LD coefficient:

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

$$D = f_{AB} - f_A \times f_B$$

### Assessing LD:

- $D' = D/D_{\max}$
- $r^2$



# Introduction

## GWAS with NGS

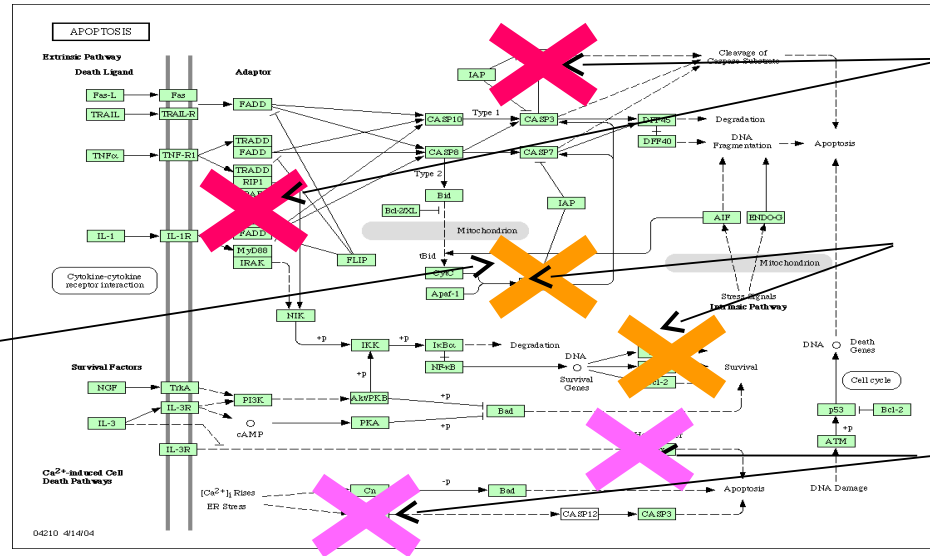
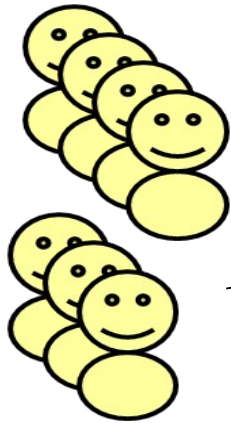
---

- 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

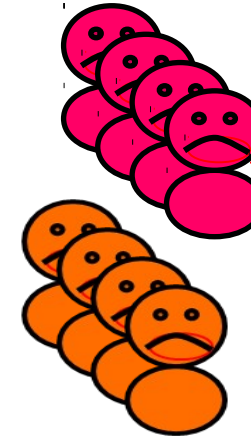
# Introduction

## Drawbacks

### Controls

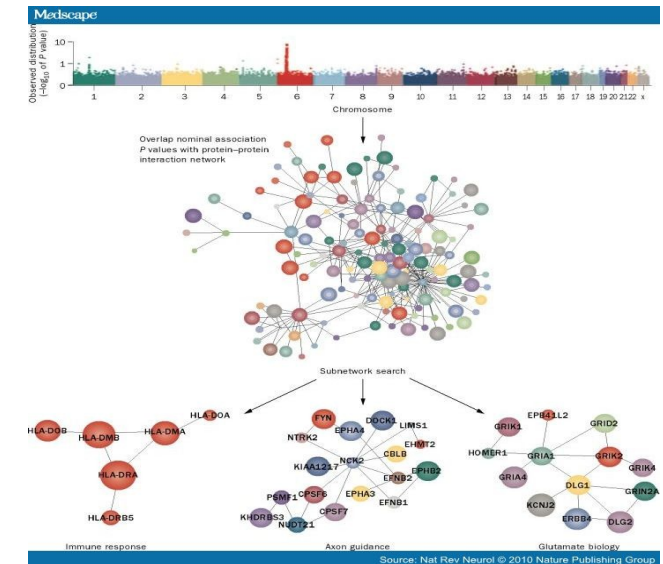


### Cases



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 Suite

Lots of features

---

- 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?

# HPG Variant Suite

## HPG Variant VCF tools

---

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)

# HPG Variant Suite

Why are these tools important?

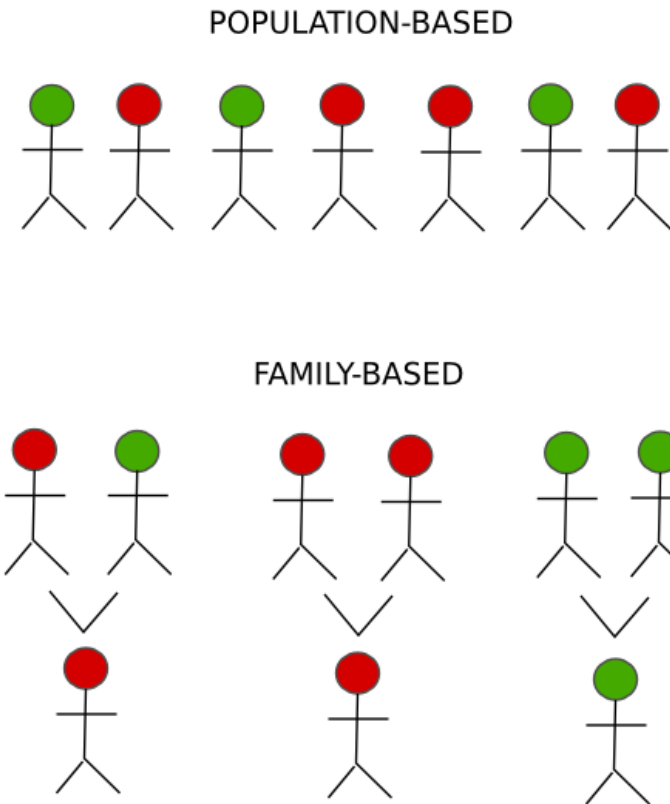
---

- 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

# HPG Variant Suite

## HPG Variant GWAS

- 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)





# HPG Variant Suite

## HPG Variant GWAS

---

- 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?**