IX International Course of Massive Data Analysis FOR GENOMICS

ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS

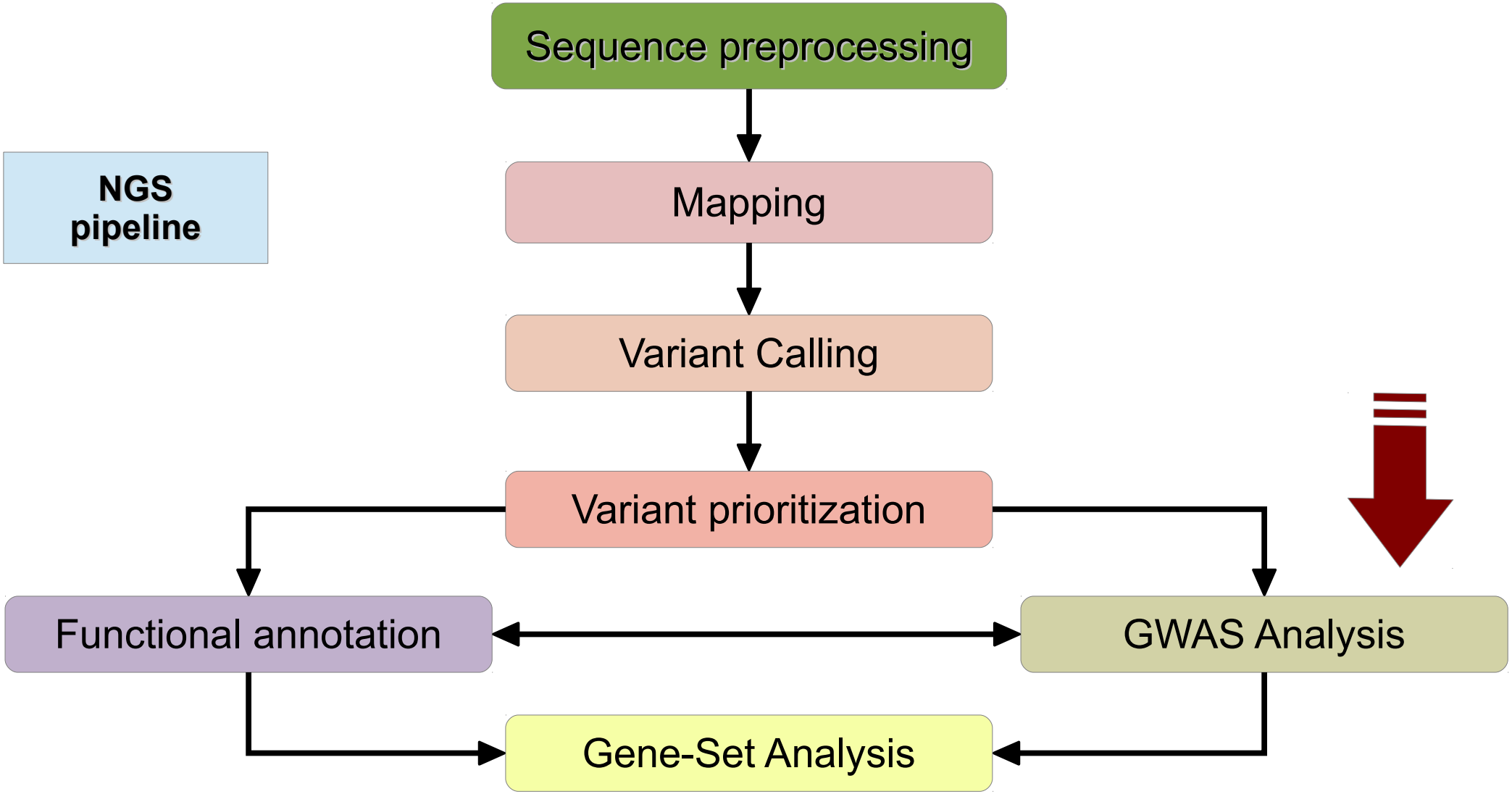
PRINCIPE FELIPE
CENTRO DE INVESTIGACION

INB

Ignacio Medina
Cristina Y. González

Genomic-wide Association Studies

Where are we?



Ignacio Medina
Cristina Y. González

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

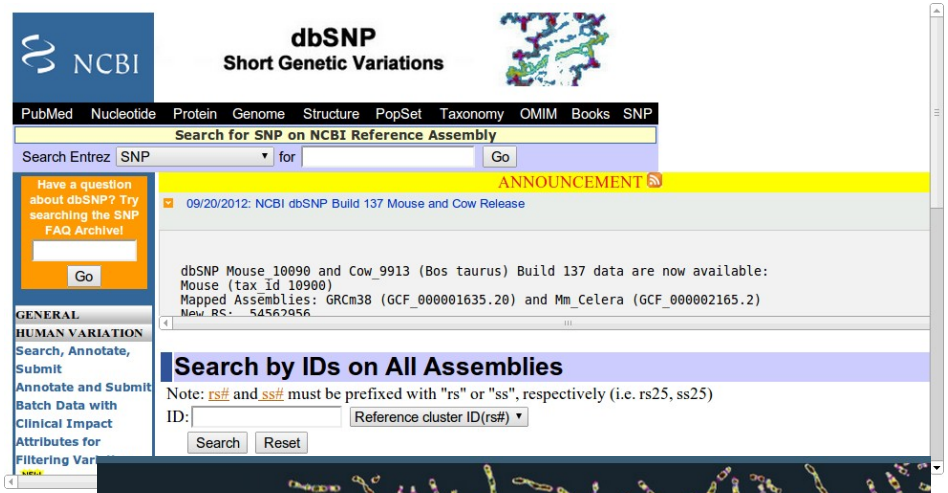
- Genotyping technology has made possible GWAS analysis, today we can genotype more than 1 million SNPs and Copy Number Variants with microarrays

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

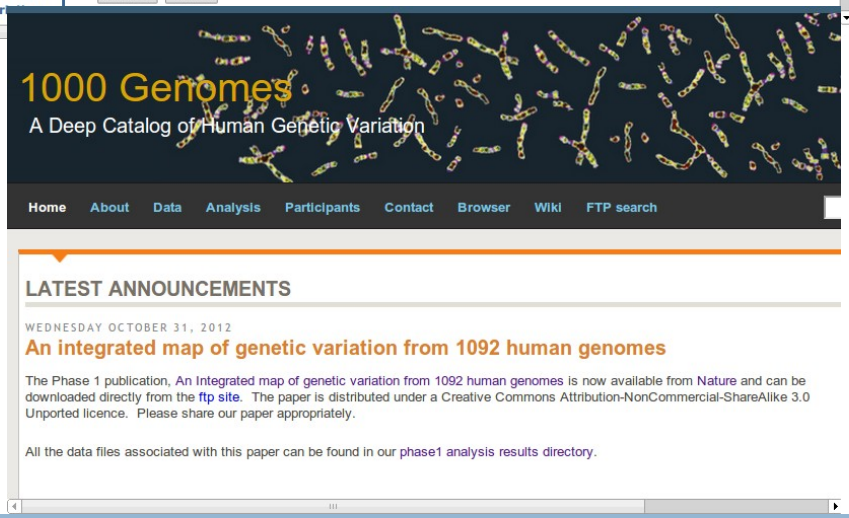
Illumina Omni5 features more than 4.3 million high-value markers. And room for 500k of your own

Genotyping catalog

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



Genotyping projects



Ignacio Medina
Cristina Y. González

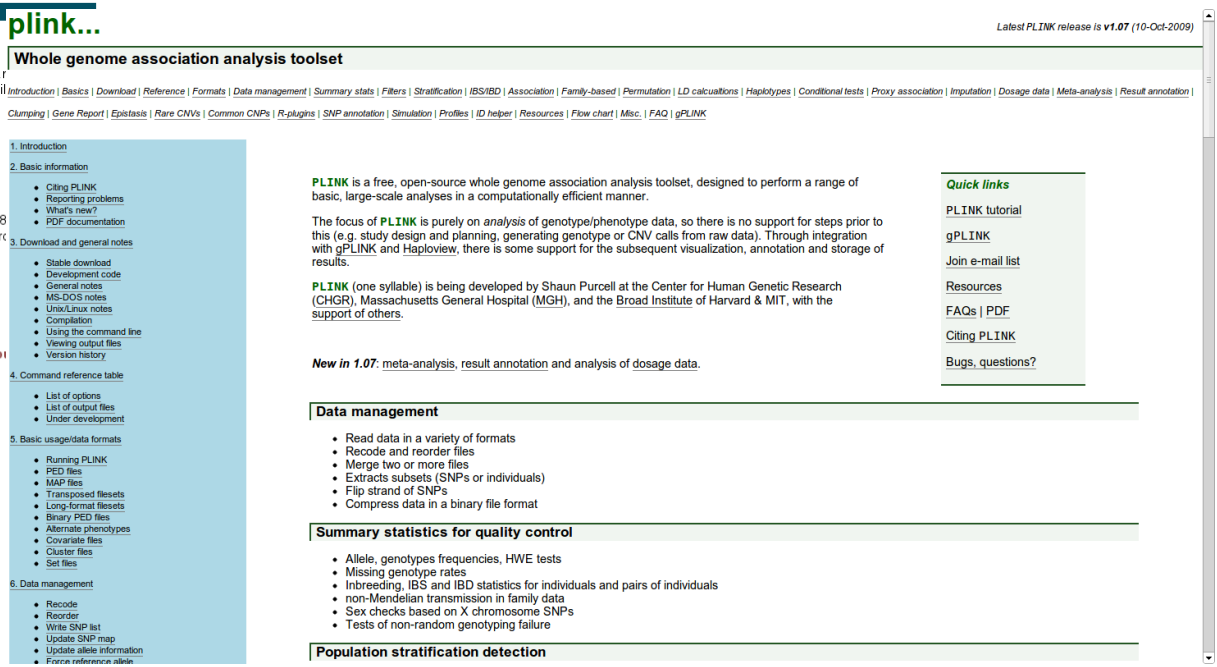
Introduction
Classic GWAS II, resources

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

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

Introduction
Classic GWAS results

THE LANCET Neurology

Search for in All Fields GO Advanc

Home | Journals | Specialties | Clinical | Global Health | Multimedia | Conferences | Information fo

The Lancet Neurology, Volume 7, Issue 11, Pages 1067 - 1072, November 2008
doi:10.1016/S1474-4422(08)70241-2 Cite or Link Using DOI
This article can be found in the following collections: Genetics & Genomics; Neurology (Genetics & neurology)

< Previous Article

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 Sign in to NCBI

PubMed.gov US National Library of Medicine National Institutes of Health

PubMed Search Advanced Help

Display Settings: Abstract Send to:

nature FREE Author Manuscript in PubMed Central

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 x 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 x 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, provi
European ancestry are excluded, the extent of population stratification in the British population
findings offer new avenues for exploring the pathophysiology of these important disorders. We a
results and software, which will be widely available to other investigators, will provide a powerful
genetics research.

Save items

Add to Favorites

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 GO Advanc

Home | Journals | Specialties | Clinical | Global Health | Multimedia | Conferences | Information fo

The Lancet, Volume 380, Issue 9844, Pages 815 - 823, 1 September 2012
doi:10.1016/S0140-6736(12)60681-3 Cite or Link Using DOI
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[‡]

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

THE LANCET

Search for in All Fields GO Advanc

Home | Journals | Specialties | Clinical | Global Health | Multimedia | Conferences | Information fo

The Lancet, Volume 377, Issue 9766, Pages 641 - 649, 19 February 2011
doi:10.1016/S0140-6736(10)62345-8 Cite or Link Using DOI
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[‡]

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 ($p<5\times10^{-8}$). Six were previously identified loci (*MAPT*, *SNCA*, *HLA-DRB5*, *BST1*, *GAK* and *LRRK2*) and five were newly identified loci (*ACMSD*, *STK39*, *MCCC1/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.

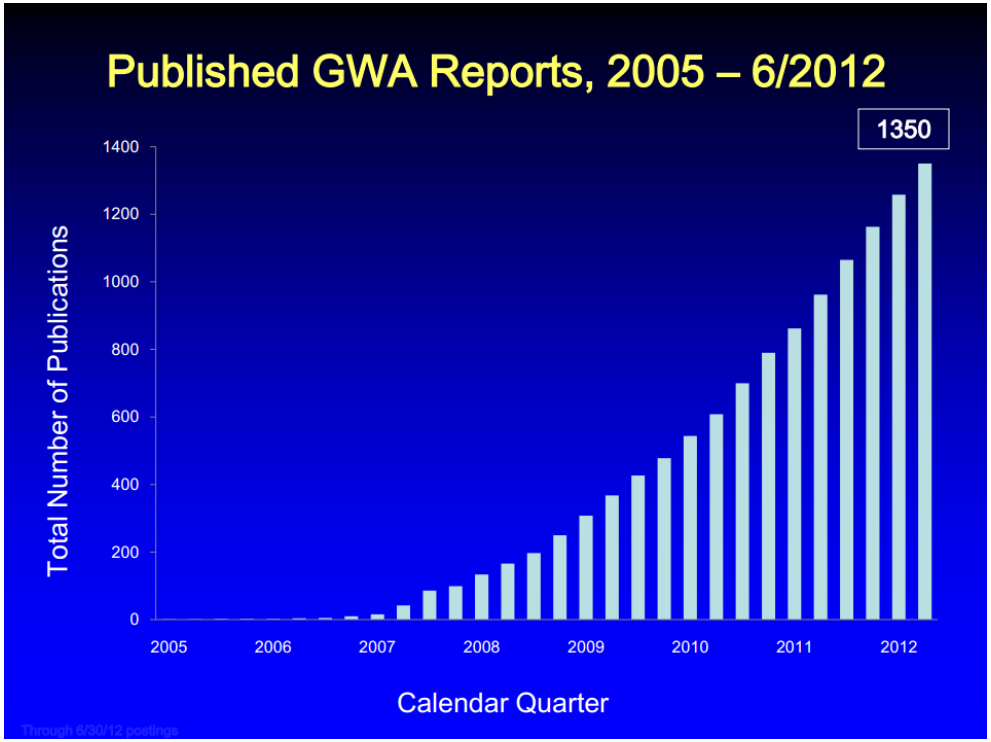
Ignacio Medina
Cristina Y. González

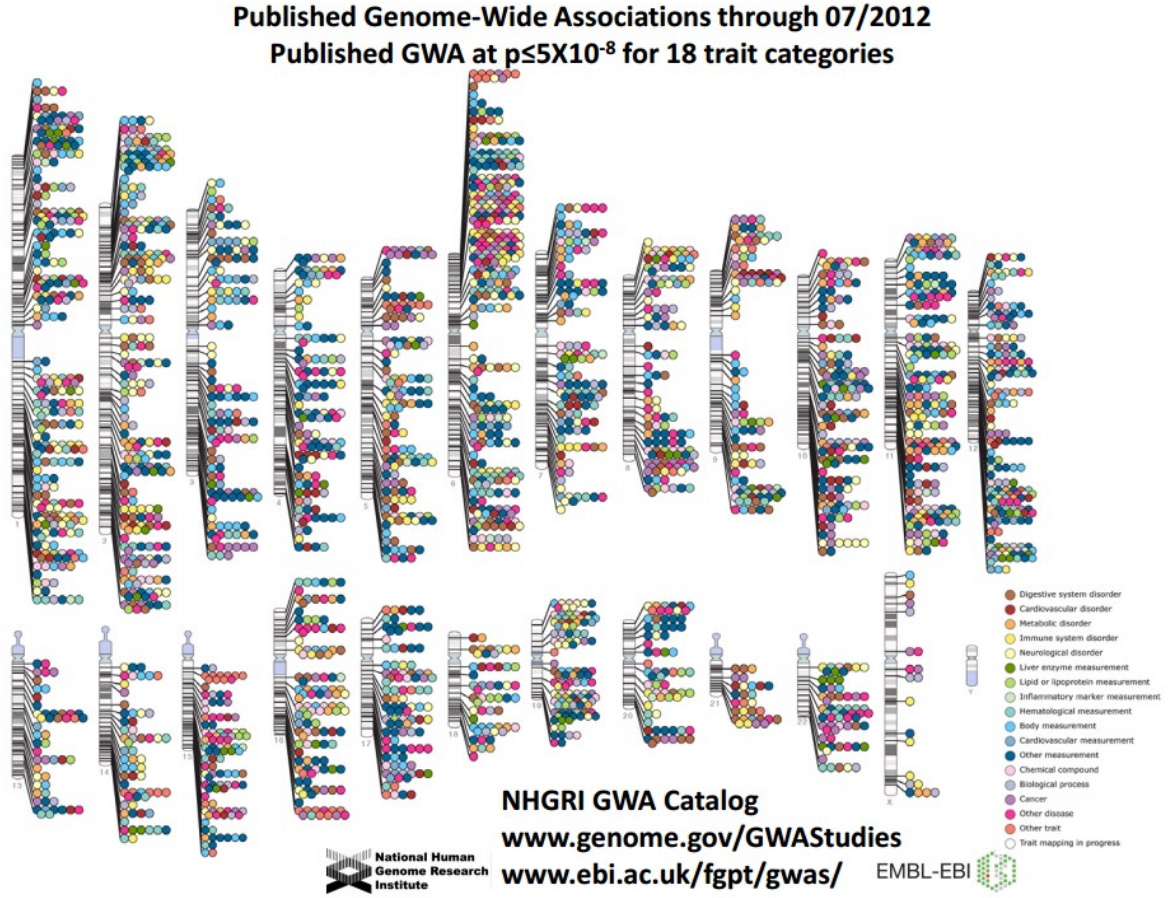
Genomic-Wide Association Studies

Introduction
GWAS catalog

A Catalog of Published Genome-Wide Association Studies

http://www.genome.gov/gwastudies





Ignacio Medina
Cristina Y. González

Genomic-Wide Association Studies

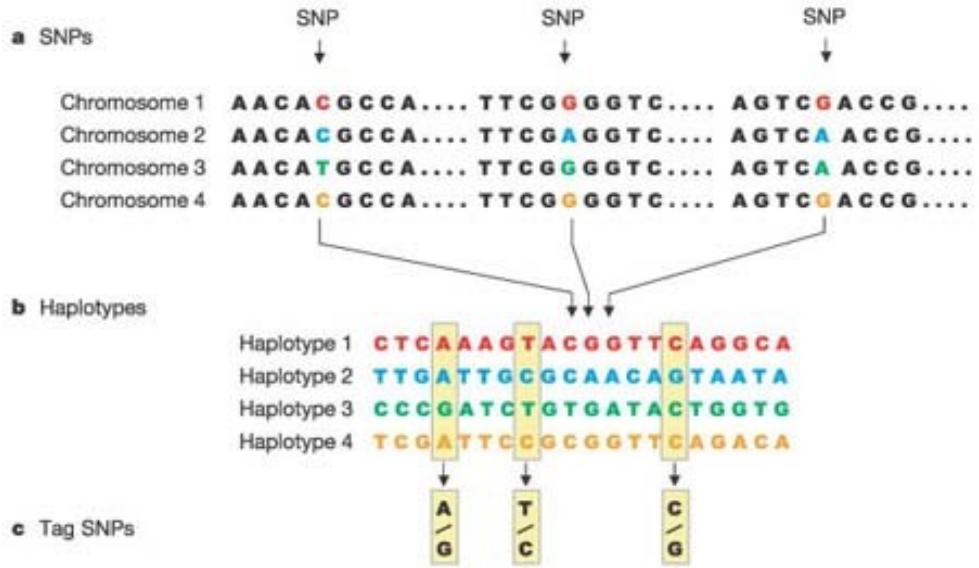
Introduction
What is a Haplotype?

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

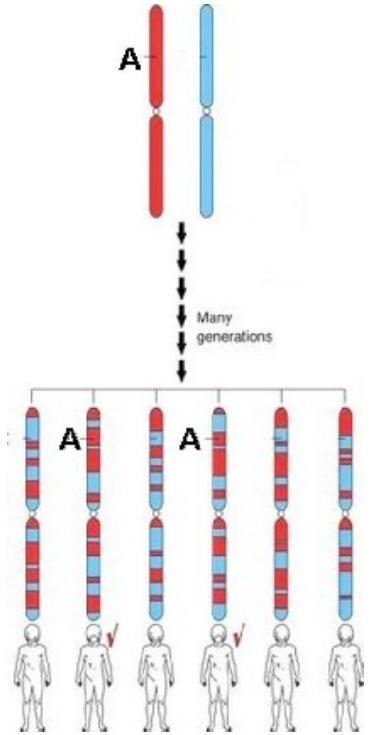
a) Short stretch of DNA for 4 different people – 3 SNPs are present

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

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 *pA* and *pB*, respectively, in the population.
  - If the two loci are independent, then we would expect to see the *AB* haplotype at frequency *pApB*.
  - 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.

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

There are 4 possible haplotypes

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)

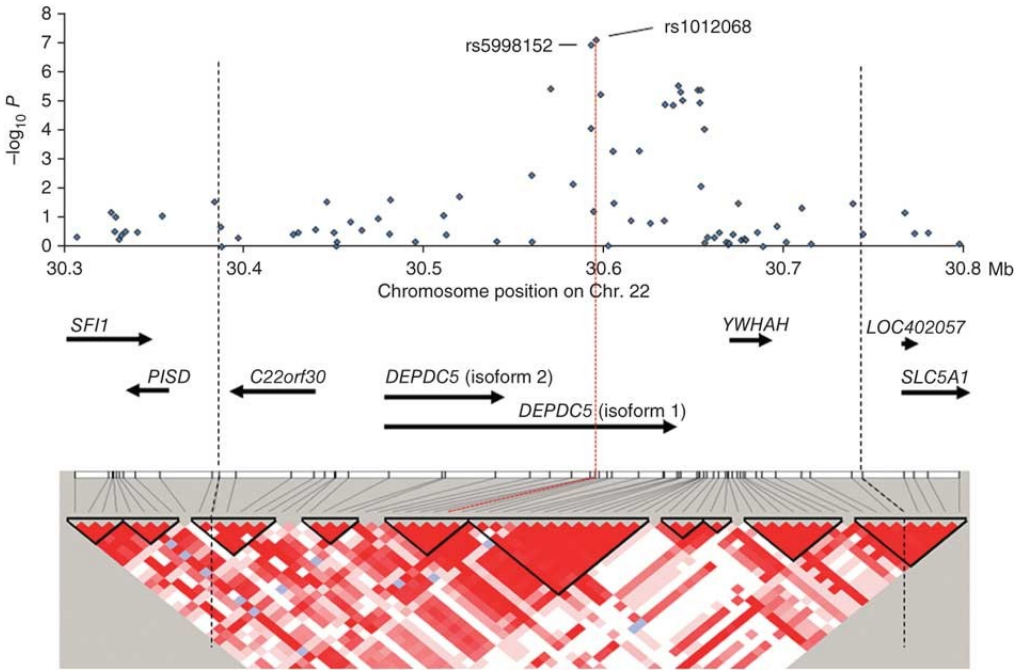
$f_{AB} = f_A f_B + D$
$f_{aB} = f_a f_B + D$
$f_{Ab} = f_A f_b + D$
$f_{Ab} = f_A f_b + D$

where D is the LD coefficient:

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

Assessing LD:

- D' = D/Dmax
- r2



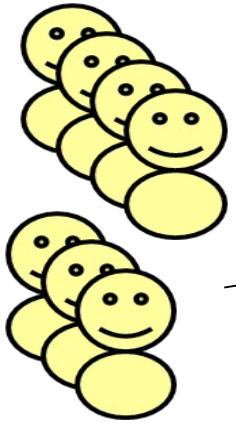
Ignacio Medina
Cristina Y. González

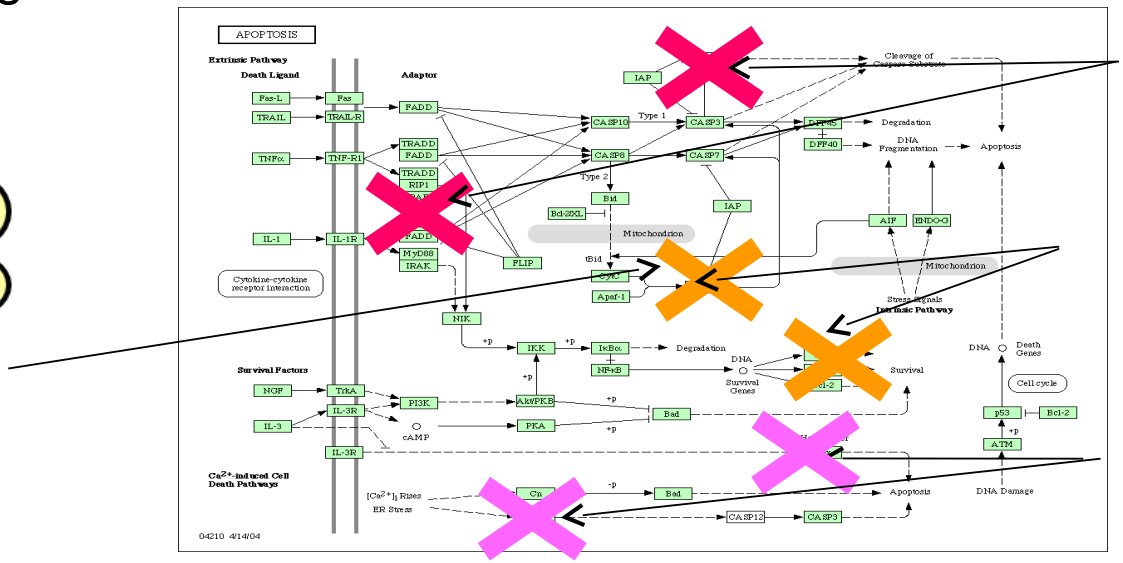
Introduction
GWAS with NGS

- 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

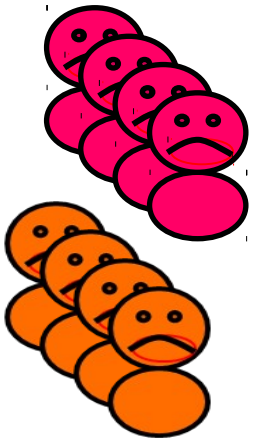
Introduction
Drawbacks

Controls



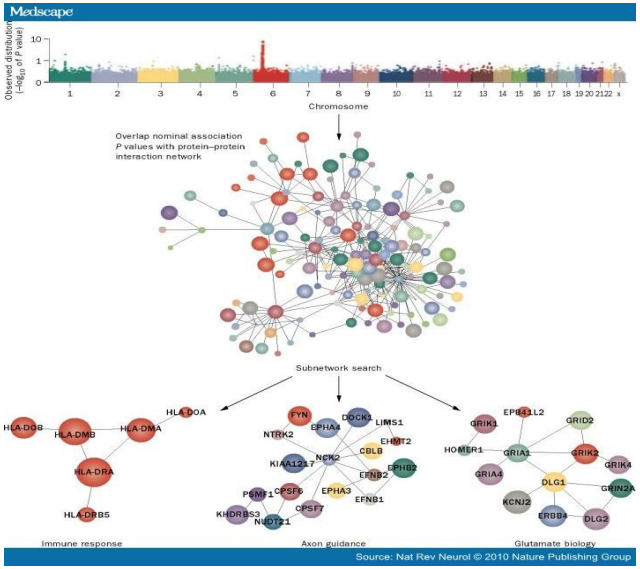


Cases



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 (unknow at this moment) affected.



Ignacio Medina
Cristina Y. González

Genomic-Wide Association Studies

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 you 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)

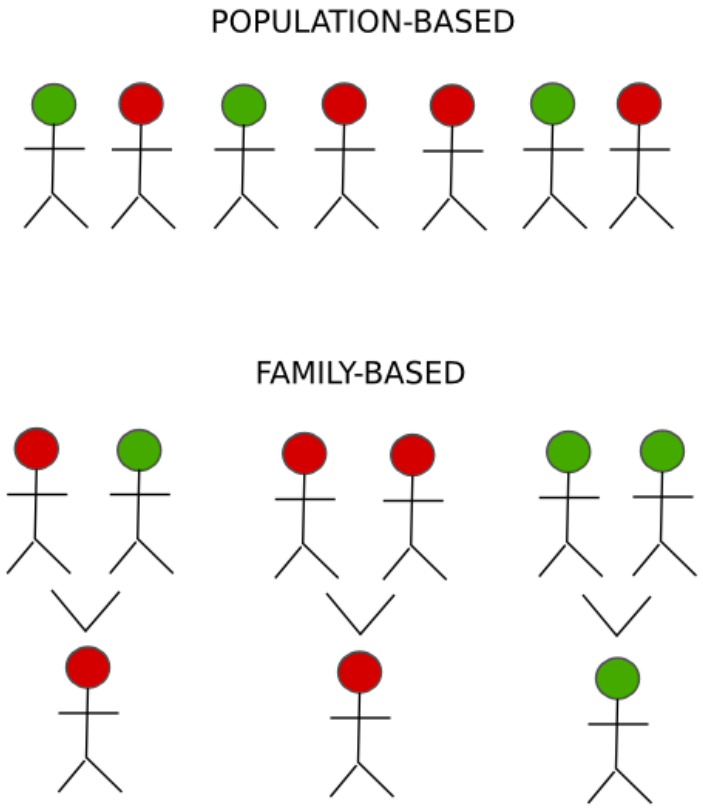
HPG Variant Suite
Why are these tools important?

- 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)



Ignacio Medina
Cristina Y. González

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?