



# GWAS (GENOME WIDE ASSOCIATION STUDY)

PRESENTED BY: HIRA SHAHID  
JUNIOR RESEARCH ENGINEER  
BSC

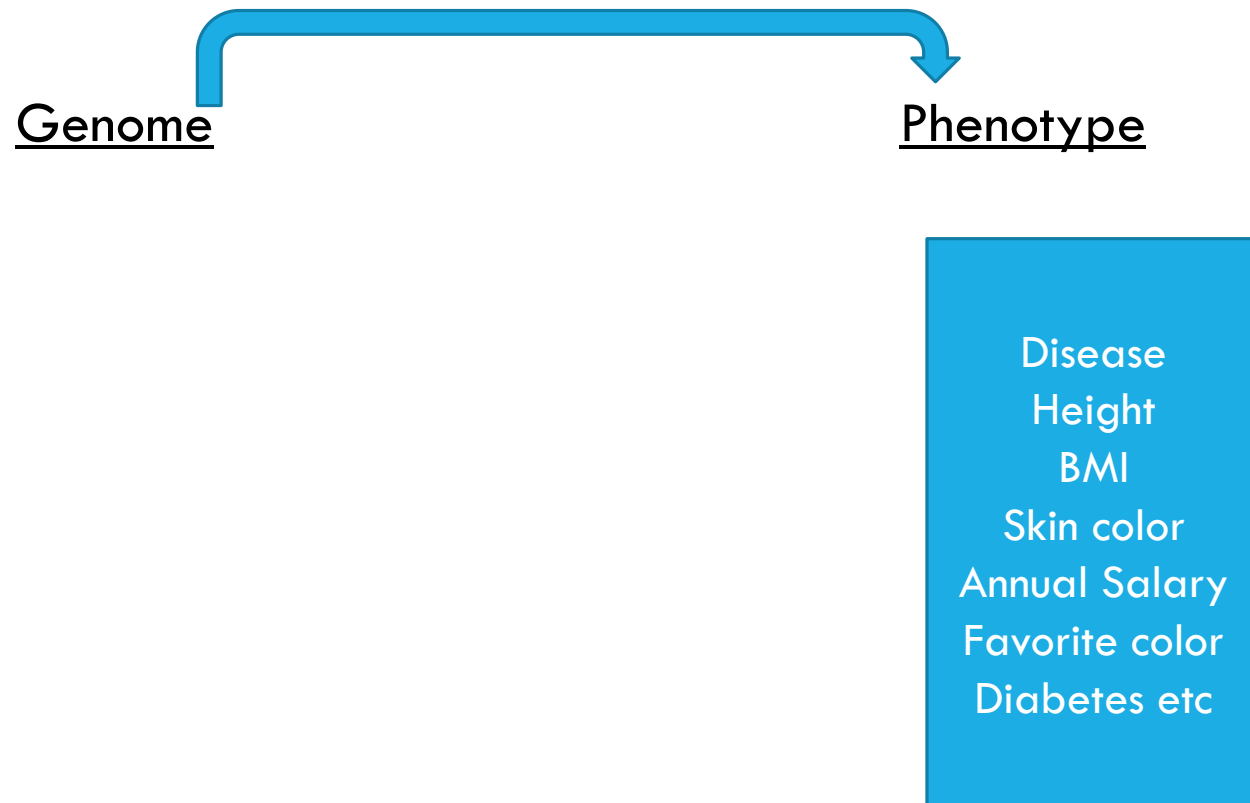
Why are we different? Why do certain people get a disease?

What are the mechanisms underlying these differences?

**How genetic variation controls the phenotype?**

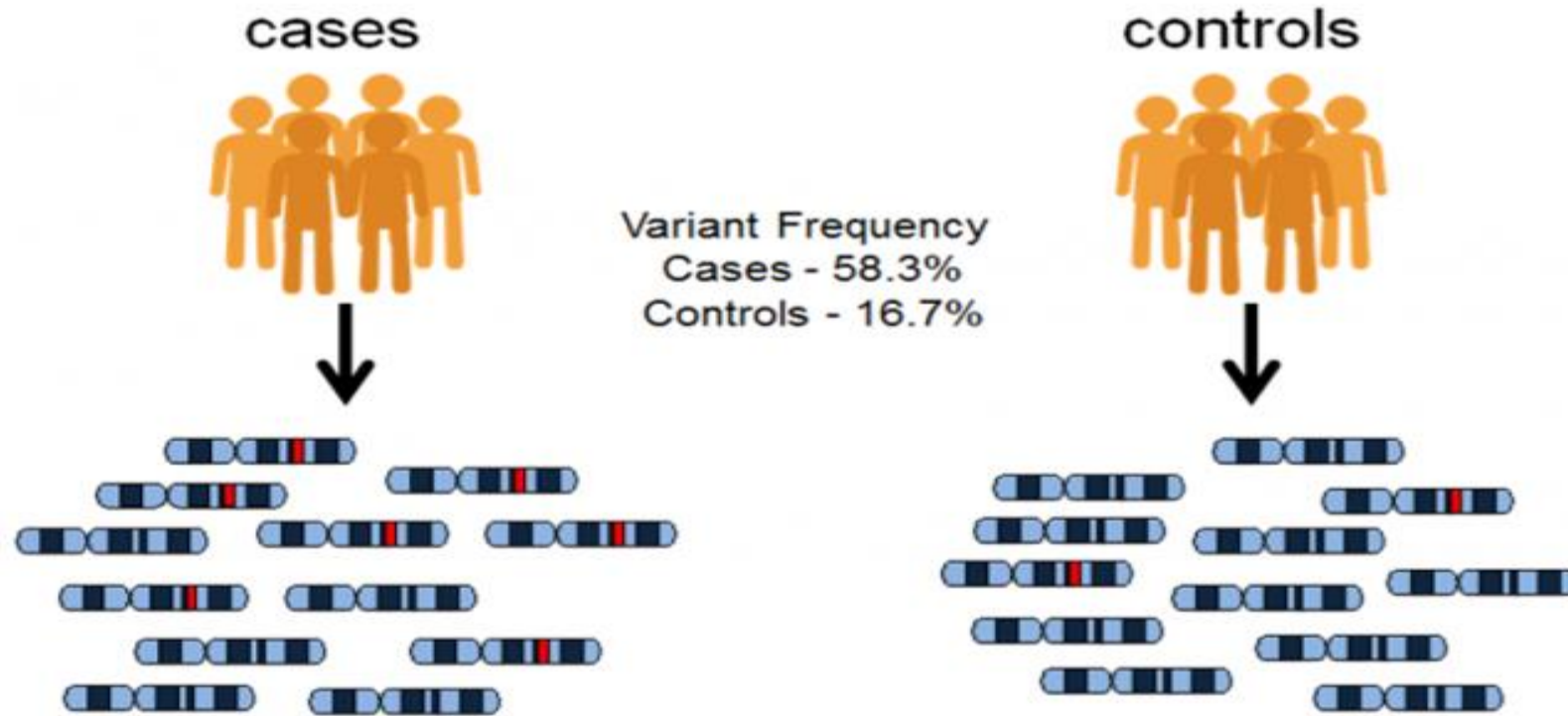


GENOME WIDE ASSOCIATION STUDIES (GWAS) ARE HYPOTHESIS-FREE METHODS FOR IDENTIFYING ASSOCIATIONS BETWEEN GENETIC REGIONS (LOCI) AND PHENOTYPES (INCLUDING DISEASES).



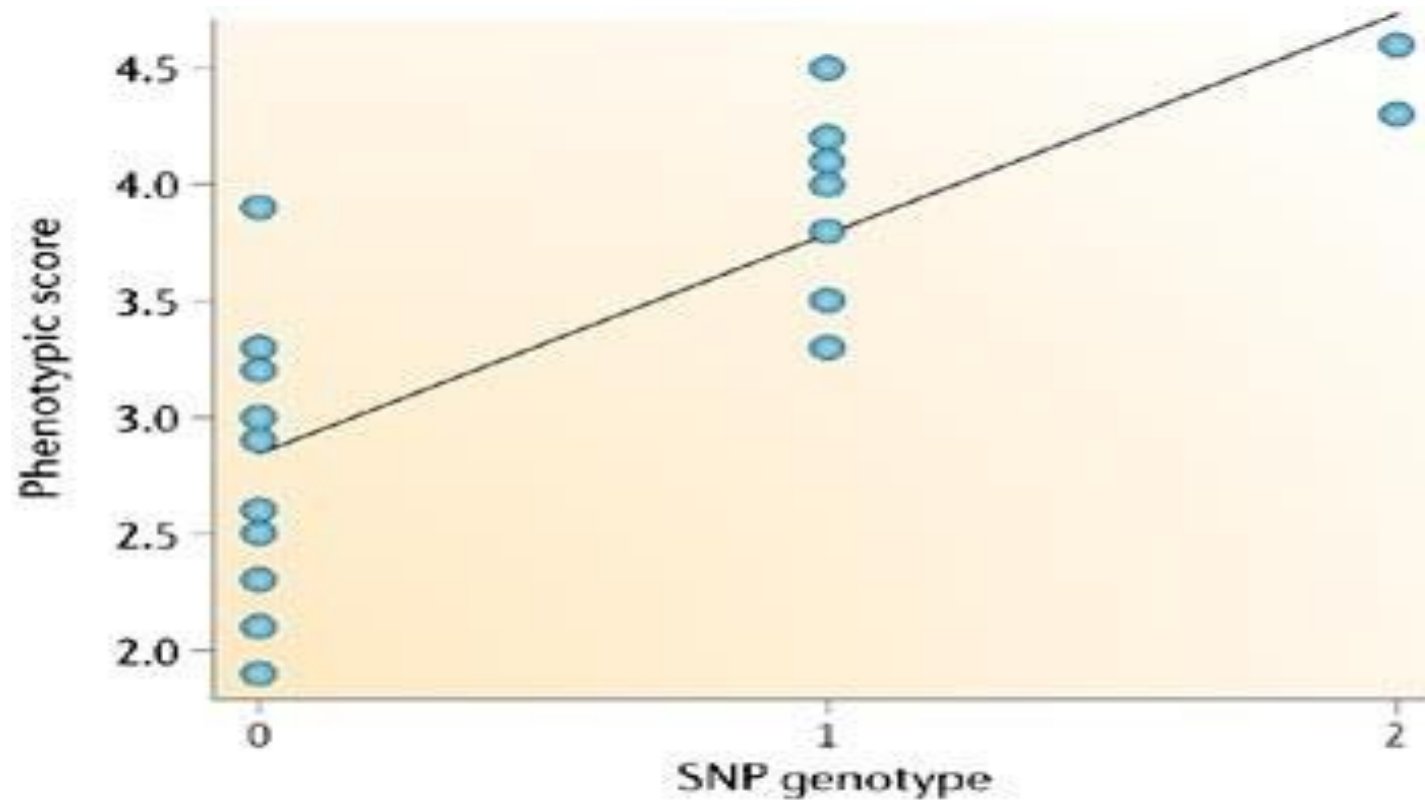
# GENOME WIDE ASSOCIATION STUDY

## CASE CONTROL



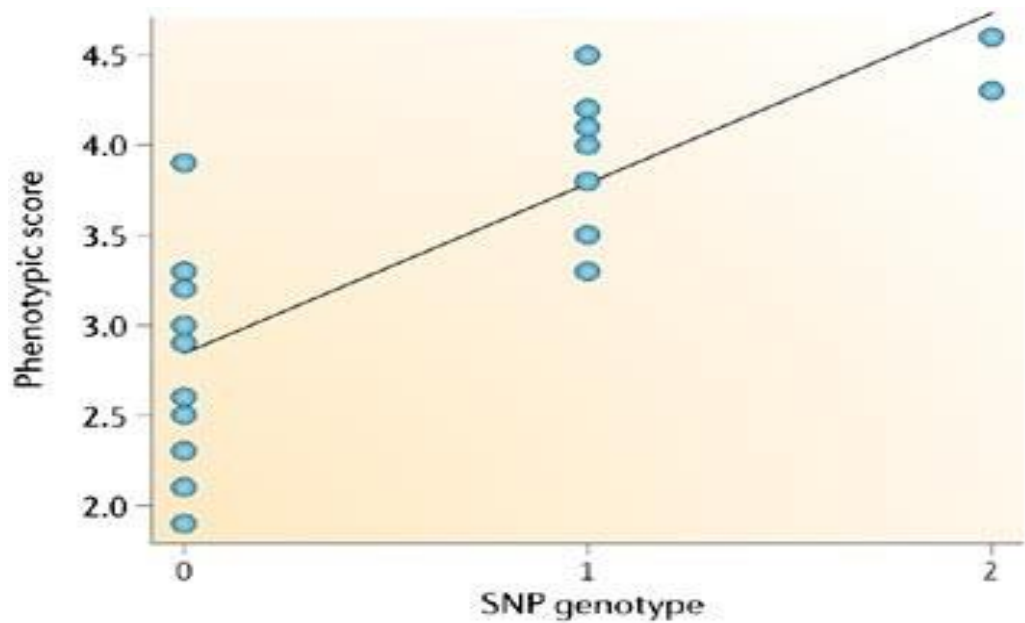


# GENOME WIDE ASSOCIATION STUDY CONTINUOUS TRAIT



Copyright © 2006 Nature Publishing Group  
Nature Reviews | **Genetics**

# GWAS MODEL



Copyright © 2006 Nature Publishing Group  
Nature Reviews | **Genetics**

$$y_j = \beta_i g_{ij} + \varepsilon_j$$

individual:  $j$

SNP:  $i$

phenotype:  $y$

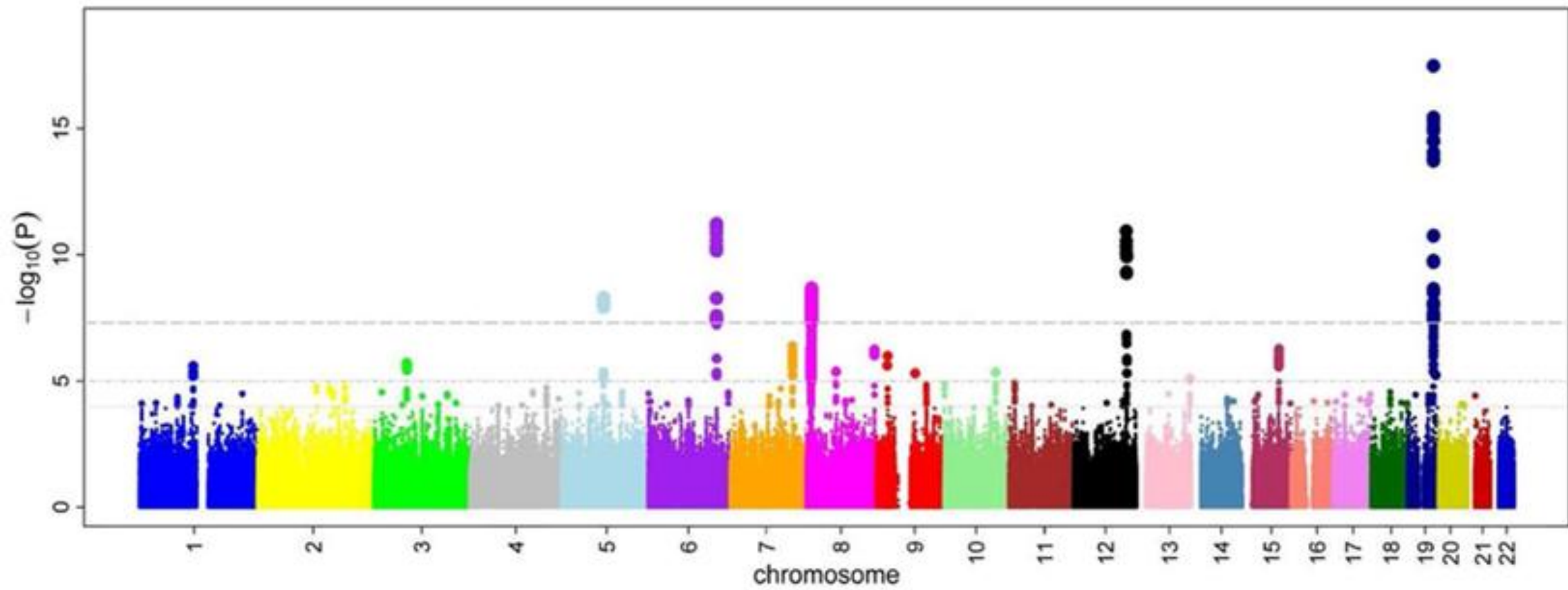
genotype:  $g$

Effect of SNP  $i$ :  $\beta$

all other effects:  $\varepsilon$

# RESULT PLOT

Null: Genotype has no estimated effect  
Alternate: Genotype has non zero estimated effect.



# GWAS METHOD

1. QC
2. Population Stratification
3. Imputation
3. Association
4. Post GWAS Analysis





# GWAS METHODS




Received: 7 February 2017 | Revised: 11 December 2017 | Accepted: 20 December 2017

DOI: 10.1002/mpr.1608

**ORIGINAL ARTICLE**

WILEY

# A tutorial on conducting genome-wide association studies: Quality control and statistical analysis

Andries T. Marees<sup>1,2,3,4,5</sup>  | Hilde de Kluiver<sup>6</sup> | Sven Stringer<sup>7</sup> | Florence Vorspan<sup>1,2,3,4,8,9</sup> |  
Emmanuel Curis<sup>3,10,11</sup> | Cynthia Marie-Claire<sup>2,3,4</sup> | Eske M. Derks<sup>1,5</sup>

# PLINK

\*.ped

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

\*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

\*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
0	1	0			
0	1	1			

\*.bed

Contains binary version of the SNP info of the \*.ped file.  
(not in a format readable for humans)

\*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C
2	2	0			
3	3	0			

Covariate file

C1	C2	C3
1.00812835	0.00606235	-0.000871105
0.0600943	0.0318994	-0.0827743
0.0431903	0.00133068	-0.000276131

Legend

FID	Family ID	rs{x}	Alleles per subject per SNP
IID	Individual ID	Chr	Chromosome
PID	Paternal ID	SNP	SNP name
MID	Maternal ID	GD	Genetic distance (morgans)
Sex	Sex of subject	BPP	Base-pair position (bp units)

FID	IID
1	1
2	2
3	3

s (e.g., Multidimensional  
10S) components)

Phenotype

C{x}

Covariate  
Scaling (M)

lisa.surfsara.nl - PuTTY

```
amarees@login1:~/genetic_data$ plink --bfile MY_DATA --assoc --out gwas_results
```

Path to the directory containing your files\*

Indicate the usage of PLINK\*\*

Specify the input file name

Specify the options

Specify the output filename

**TABLE 1** Overview of seven QC steps that should be conducted prior to genetic association analysis

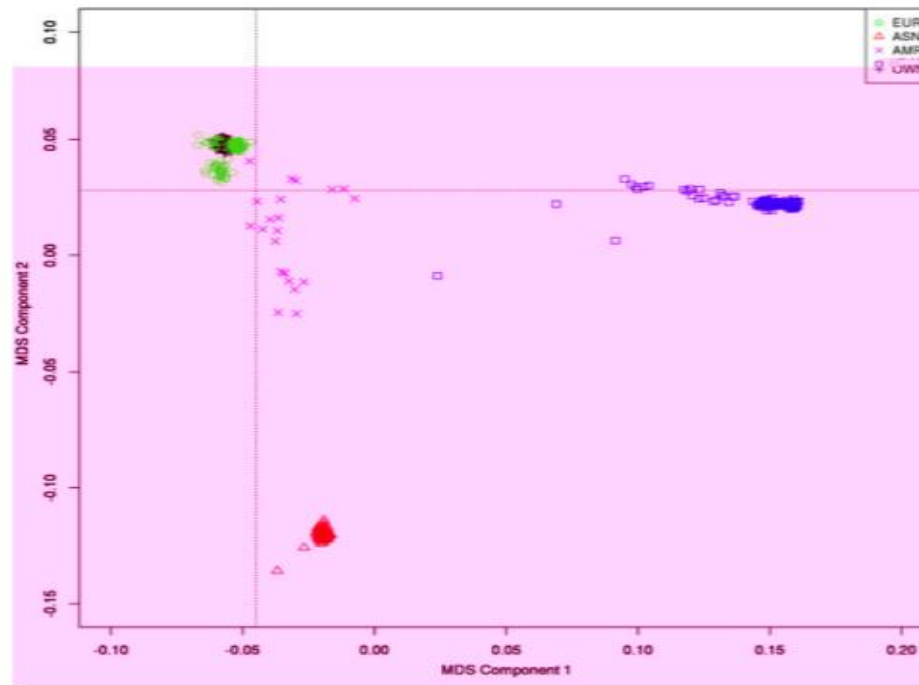
Step	Command	Function	Thresholds and explanation
1: Missingness of SNPs and individuals	--geno  --mind	Excludes SNPs that are missing in a large proportion of the subjects. In this step, SNPs with low genotype calls are removed.  Excludes individuals who have high rates of genotype missingness. In this step, individual with low genotype calls are removed.	We recommend to first filter SNPs and individuals based on a relaxed threshold (0.2; >20%), as this will filter out SNPs and individuals with very high levels of missingness. Then a filter with a more stringent threshold can be applied (0.02).  Note, SNP filtering should be performed before individual filtering.
2: Sex discrepancy	--check-sex	Checks for discrepancies between sex of the individuals recorded in the dataset and their sex based on X chromosome heterozygosity/homozygosity rates.	Can indicate sample mix-ups. If many subjects have this discrepancy, the data should be checked carefully. Males should have an X chromosome homozygosity estimate >0.8 and females should have a value <0.2.
3: Minor allele frequency (MAF)	--maf	Includes only SNPs above the set MAF threshold.	SNPs with a low MAF are rare, therefore power is lacking for detecting SNP-phenotype associations. These SNPs are also more prone to genotyping errors. The MAF threshold should depend on your sample size, larger samples can use lower MAF thresholds. Respectively, for large (N = 100.000) vs. moderate samples (N = 10000), 0.01 and 0.05 are commonly



4: Hardy–Weinberg equilibrium (HWE)	--hwe	Excludes markers which deviate from Hardy–Weinberg equilibrium.	Common indicator of genotyping error, may also indicate evolutionary selection. For <u>binary traits</u> we suggest to exclude: HWE $p$ value $<1e-10$ in cases and $<1e-6$ in controls. Less strict case threshold avoids discarding disease-associated SNPs under selection (see online tutorial at <a href="https://github.com/MareesAT/GWA_tutorial/">https://github.com/MareesAT/GWA_tutorial/</a> ). For <u>quantitative traits</u> , we recommend HWE $p$ value $<1e-6$ .
5: Heterozygosity	For an example script see <a href="https://github.com/MareesAT/GWA_tutorial/">https://github.com/MareesAT/GWA_tutorial/</a>	Excludes individuals with high or low heterozygosity rates	Deviations can indicate sample contamination, inbreeding. We suggest removing individuals who deviate $\pm 3$ SD from the samples' heterozygosity rate mean.
6: Relatedness	--genome  --min	Calculates identity by descent (IBD) of all sample pairs. Sets threshold and creates a list of individuals with relatedness above the chosen threshold. Meaning that subjects who are related at, for example, $\pi\text{-hat} > 0.2$ (i.e., second degree relatives) can be detected.	Use independent SNPs ( <b>pruning</b> ) for this analysis and limit it to autosomal chromosomes only. Cryptic relatedness can interfere with the association analysis. If you have a family-based sample (e.g., parent-offspring), you do not need to remove related pairs but the statistical analysis should take family relatedness into account. However, for a population based sample we suggest to use a $\pi\text{-hat}$ threshold of 0.2, which is in line with the literature (Anderson et al., 2010; Guo et al., 2014).
7: Population stratification	--genome  --cluster --mds-plot k	Calculates identity by descent (IBD) of all sample pairs. Produces a $k$ -dimensional representation of any substructure in the data, based on IBS.	Use independent SNPs ( <b>pruning</b> ) for this analysis and limit it to autosomal chromosomes only. $K$ is the number of dimensions, which needs to be defined (typically 10). This is an important step of the QC that consists of multiple proceedings but for reasons of completeness we briefly refer to this step in the table. This step will be described in more detail in section “controlling for population stratification.”



# CONTROLLING FOR POPULATION STRATIFICATION



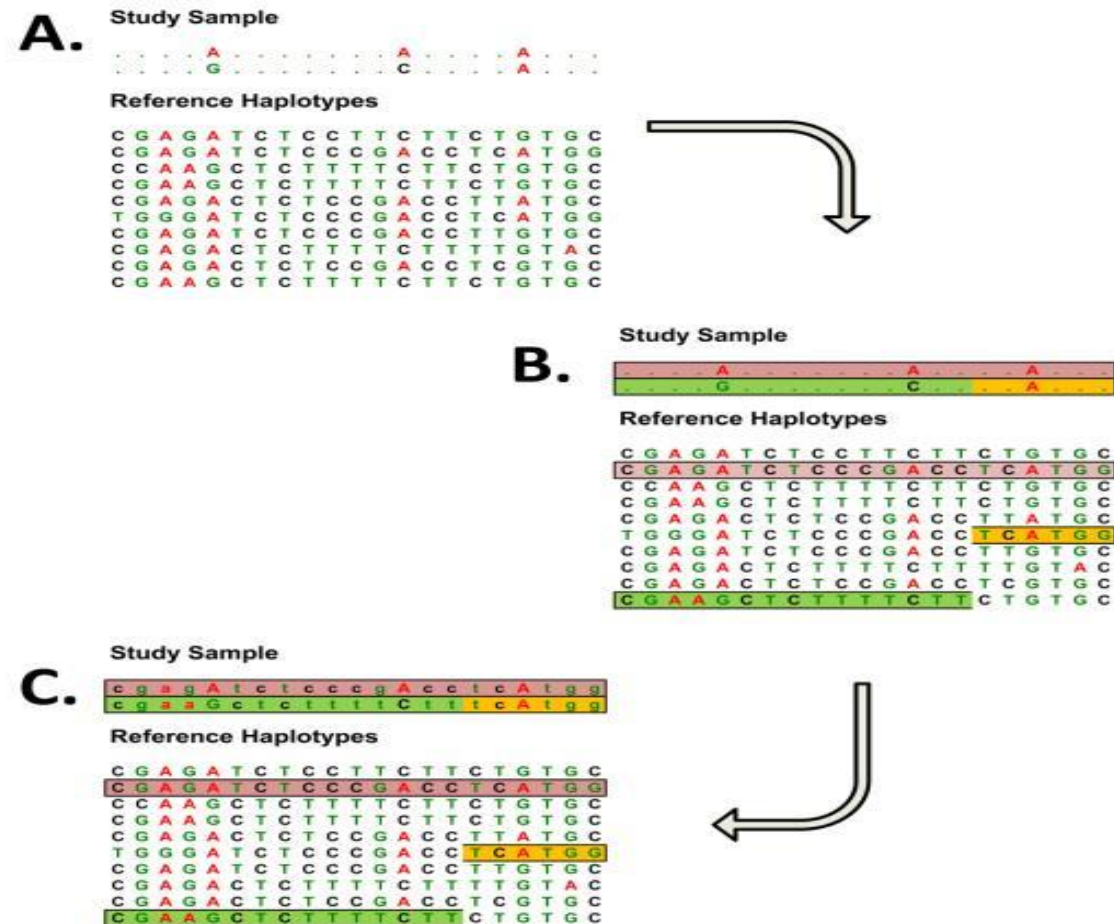
**FIGURE 3** Multidimensional scaling (MDS) plot of 1KG against the CEU of the HapMap data (which could be seen as your "own" data in this example, as it is being used in the online tutorial at [https://github.com/MareesAT/GWA\\_tutorial/](https://github.com/MareesAT/GWA_tutorial/)). The black crosses (+ = "OWN") in the upper left part represent the first two MDS components of the individuals in the HapMap sample (the colored symbols represent the 1KG data (○ = European; □ = African;

# CHOPSTICK GENE

Once upon a time, an ethnogeneticist decided to figure out why some people eat with chopsticks and others do not. His experiment was simple. He rounded up several hundred students from a local university, asked them how often they used chopsticks, then collected buccal DNA samples and mapped them for a series of anonymous and candidate genes. The results were astounding. One of the markers, located right in the middle of a region previously linked to several behavioral traits, showed a huge correlation to chopstick use, enough to account for nearly half of the observed variance. When the experiment was repeated with students from a different university, precisely the same marker lit up. Eureka! The delighted scientist popped a bottle of champagne and quickly submitted an article to Molecular Psychiatry heralding the discovery of the 'successful-use-of-selected-handinstruments gene' (SUSHI). It took another 2 years to discover that SUSHI is a histocompatibility antigen gene that has nothing to do with chopstick use but just happens to have different allele frequencies in Asians and Caucasians, who of course differ in chopstick use for purely cultural rather than biological reasons. Even though the association data were highly significant and readily replicated, they were biologically meaningless

# GENOTYPE IMPUTATION

Stretches of shared haplotype between sample and reference panel are identified (Panel B) and missing genotypes for each study sample can be filled in by copying alleles observed in matching reference haplotypes (Panel C).



# STATISTICAL TESTS OF ASSOCIATION FOR BINARY AND QUANTITATIVE TRAITS

The `--assoc` option in PLINK performs a  $\chi^2$  test of association that does not allow the inclusion of covariates.

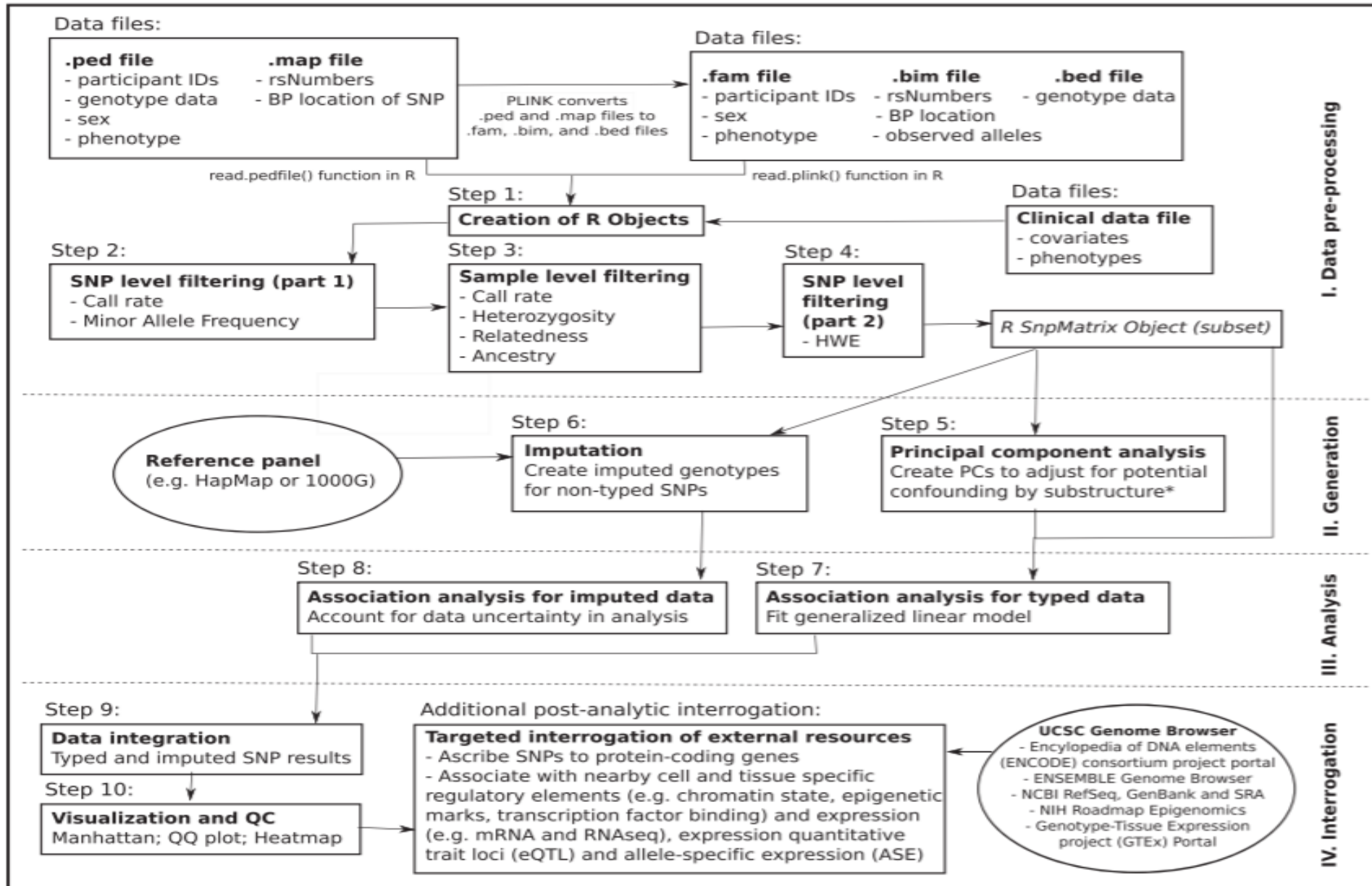
With the `--logistic` option, a logistic regression analysis will be performed which allows the inclusion of covariates. The `--logistic` option is more flexible than the `--assoc` option, yet it comes at the price of increased computational time.

Binary outcome measure

Quantitative outcome measure

Correction for multiple testing

# GENOME-WIDE ASSOCIATION (GWA) ANALYSIS WORKFLOW





Received 28 February 2015,

Accepted 6 July 2015

Published online 6 September 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6605

# A guide to genome-wide association analysis and post-analytic interrogation

**Eric Reed,<sup>a</sup> Sara Nunez,<sup>a</sup> David Kulp,<sup>b</sup> Jing Qian,<sup>c</sup>  
Muredach P. Reilly<sup>d</sup> and Andrea S. Foulkes<sup>a\*†</sup>**

```
# ---- packages ----  
# Run this once interactively to download and install Bioconductor packages and other packages.  
  
source("http://bioconductor.org/biocLite.R")  
biocLite("snpStats")  
biocLite("SNPRelate")  
biocLite("rtracklayer")  
biocLite("biomaRt")  
install.packages(c('plyr', 'GenABEL', 'LDheatmap', 'doParallel', 'ggplot2', 'coin', 'igraph', 'devtools'))  
  
library(devtools)  
install_url("http://cran.r-project.org/src/contrib/Archive/postgwas/postgwas_1.11.tar.gz")
```

# Package ‘GenABEL’

September 3, 2009

**Type** Package

**Title** genome-wide SNP association analysis

**Version** 1.4-4

**Date** 2009-09-02

**Author** Yurii Aulchenko, Maksim Struchalin

**Maintainer** Yurii Aulchenko <i.aoultchenko@erasmusmc.nl>

**Depends** R (>= 2.4.0), methods, genetics, haplo.stats, MASS

**Suggests** qvalue

**Description** a package for genome-wide association analysis between quantitative or binary traits and single-nucleotide polymorphisms (SNPs).



# GUIDANCE

An easy-to-use platform for comprehensive GWAS and PheWAS

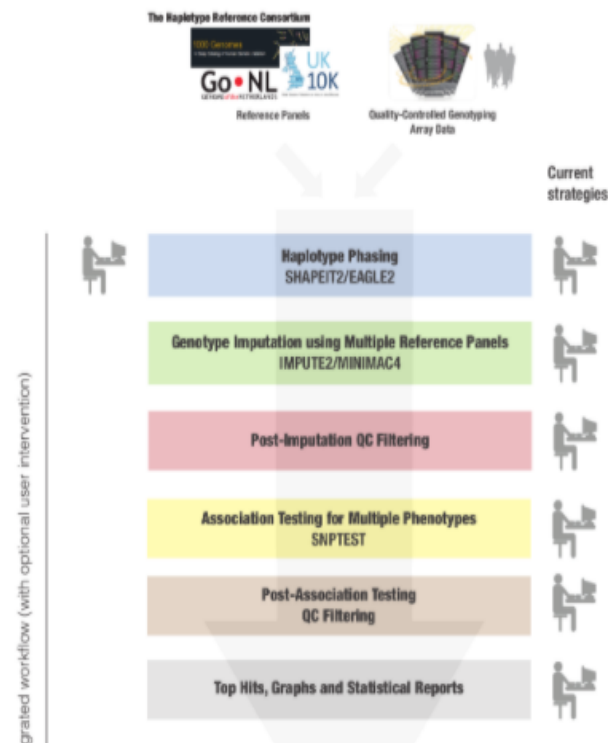
[Home](#)[How To Install](#)[How To Run](#)[Downloads](#)[FAQ](#)[Mailing list](#)

## Home

### Overview

Genome-wide Association Studies (GWAS) are a promising approach for uncovering genetic variants related to complex diseases. Although GWAS have allowed the identification of thousands of trait-disease associations, the current methodology is far from being able to explain a small fraction of the estimated heritability of each trait, even when analyzing hundreds of thousands of samples. To allow an efficient analysis of the current and upcoming GWAS datasets we propose a novel strategy, called GUIDANCE, to make the most of the information available in GWAS dataset and taking advantage of novel imputation reference panels.

GUIDANCE is an integrated framework that is able to perform haplotype phasing, genotype imputation, association testing assuming different models of inheritance and phenome-wide association analysis (PheWAS) analysis of large GWAS datasets. Moreover, this application allows performing all these steps in a single execution, as well as in a modular way with



bioRxiv is receiving many new papers on coronavirus SARS-CoV-2. A reminder: these are preliminary reports that have not been peer-reviewed and should not guide clinical practice/health-related behavior, or be reported in news media as established information.

New Results

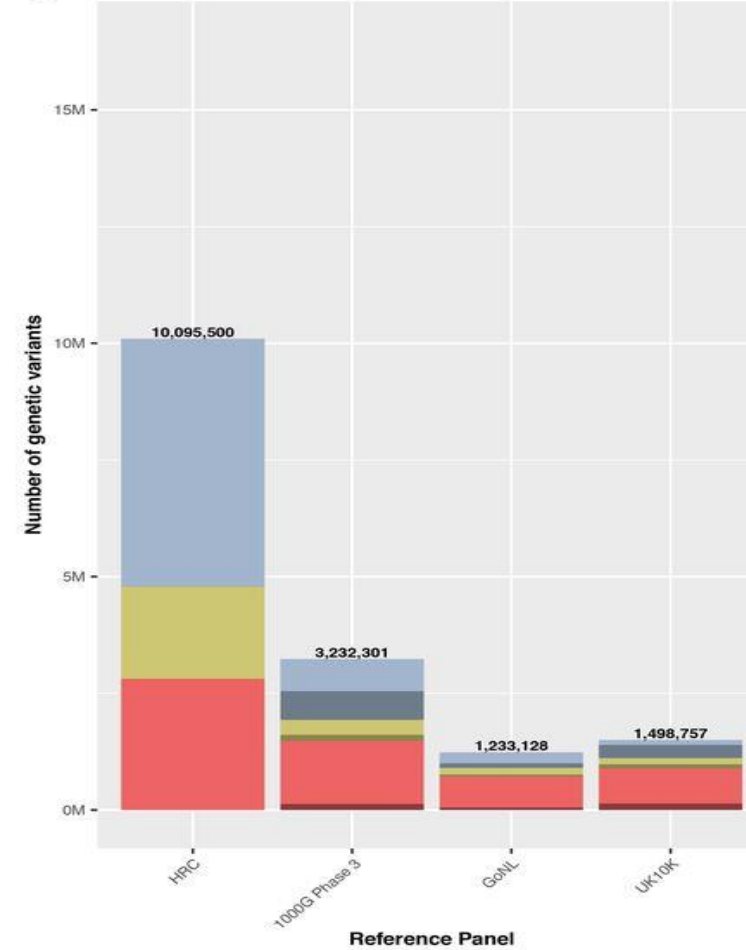
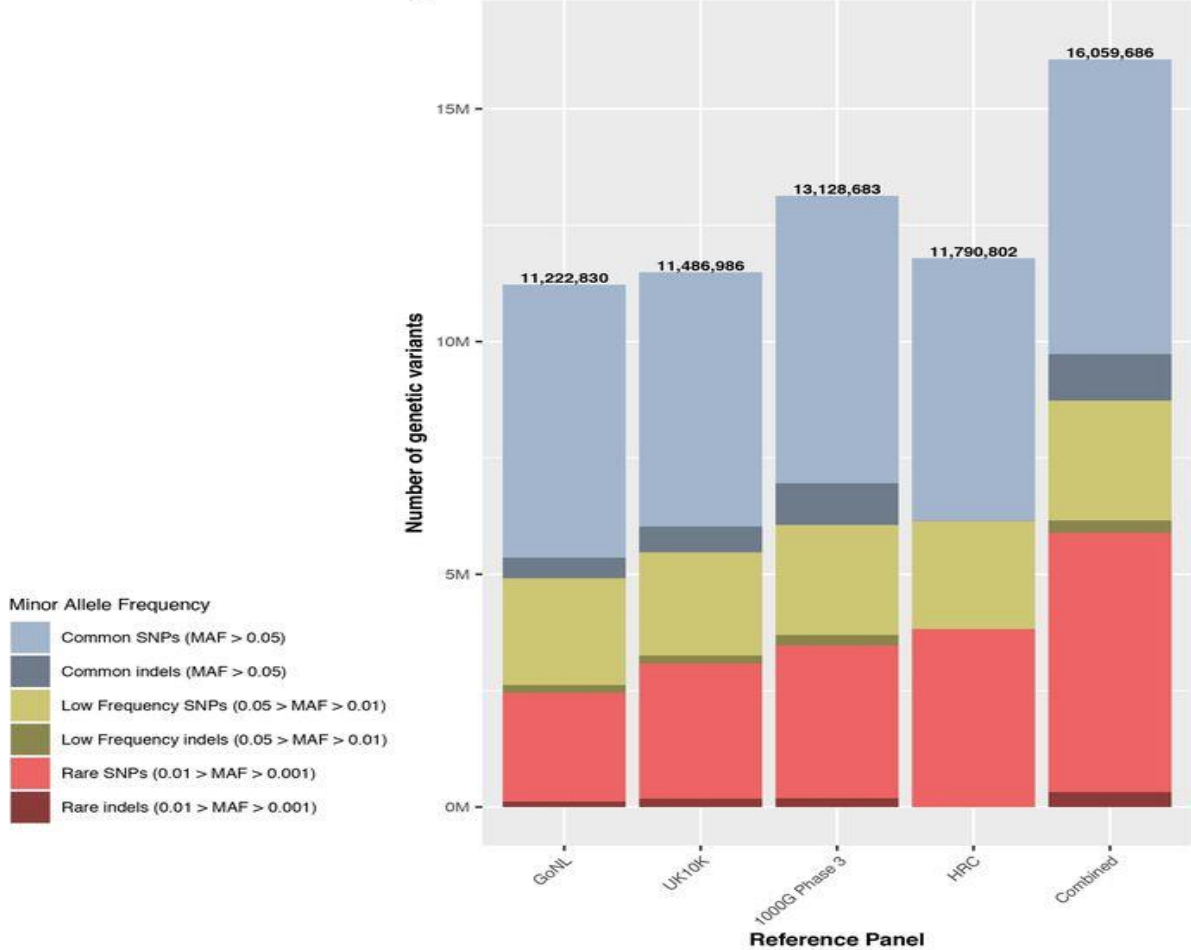
[Comment on this paper](#)

## The impact of non-additive genetic associations on age-related complex diseases

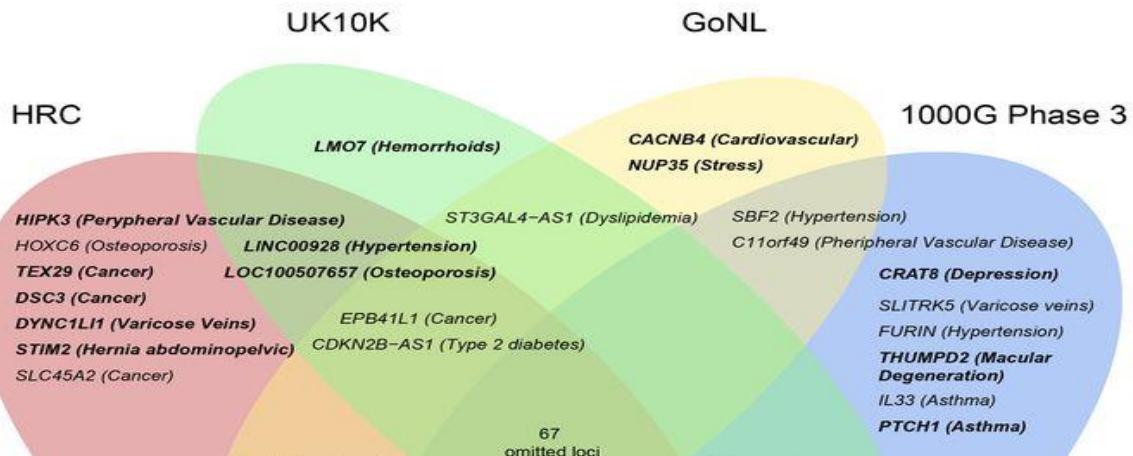
[Marta Guindo-Martínez](#), [Ramon Amela](#), [Silvia Bonàs-Guarch](#), [Montserrat Puiggròs](#), [Cecilia Salvoro](#),  
[Irene Miguel-Escalada](#), [Caitlin E Carey](#), [Joanne B. Cole](#), [Sina Rüeger](#), [Elizabeth Atkinson](#),  
[Aaron Leong](#), [Friman Sanchez](#), [Cristian Ramon-Cortes](#), [Jorge Ejarque](#), [Duncan S Palmer](#),  
[Mitja Kurki](#), FinnGen Consortium, [Krishna Aragam](#), [Jose C Florez](#), [Rosa M. Badia](#),  
[Josep M. Mercader](#), [David Torrents](#)

### Abstract

Genome-wide association studies (GWAS) are not fully comprehensive as current strategies typically test only the additive model, exclude the X chromosome, and use only one reference panel for genotype imputation. We implemented an extensive GWAS strategy, GUIDANCE, which improves genotype imputation by using multiple reference panels, includes the analysis of the X chromosome and non-additive models to test for association. We applied this methodology to 62,281 subjects across 22 age-related diseases and identified 94 genome-wide associated loci, including 26 previously unreported. We observed that 27.6% of the 94 loci would be missed if we only used standard imputation strategies and only tested the additive model. Among the new findings, we identified three novel low-frequency recessive variants with odd ratios larger than 4, which would need at least a three-fold larger sample size to be detected under the additive model. This study highlights the benefits of applying innovative strategies to better uncover the genetic architecture of complex diseases.



**c**







**THANK YOU FOR LISTENING.**

