Statistical Power, Population Stratification and Experimental Design

Christian Werner

(Quantitative geneticist and biostatistician) **EiB, CIMMYT**, Texcoco (Mexico)

Filippo Biscarini

HerrFalloppio

(Biostatistician, bioinformatician and quantitative geneticist) **CNR-IBBA**, Milan (Italy)

Oscar González-Recio

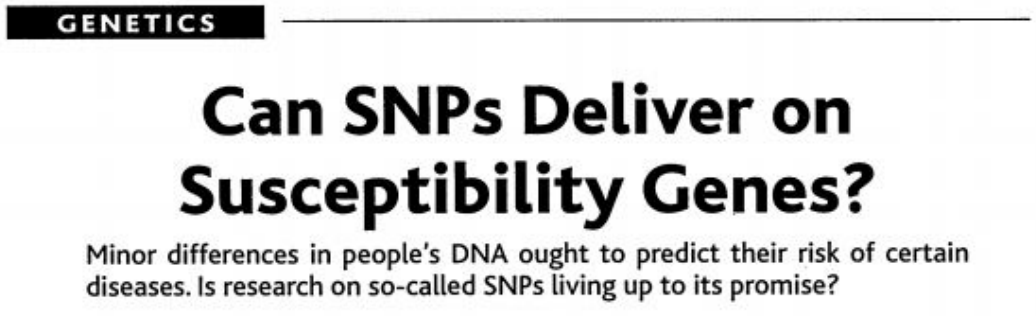
OscarGenomics

(Computational biologist and quantitative geneticist) **INIA-UPM**, Madrid (Spain)

Statistical Power (A needle in a haystack)

- Find causal mutation

Statistical Power (A needle in a haystack)



GENETICS

Can SNPs Deliver on Susceptibility Genes?

Minor differences in people's DNA ought to predict their risk of certain diseases. Is research on so-called SNPs living up to its promise?

- Altshuler et al. (2000) discussed retested 13 published associations of SNPs with type II diabetes in an independent population.

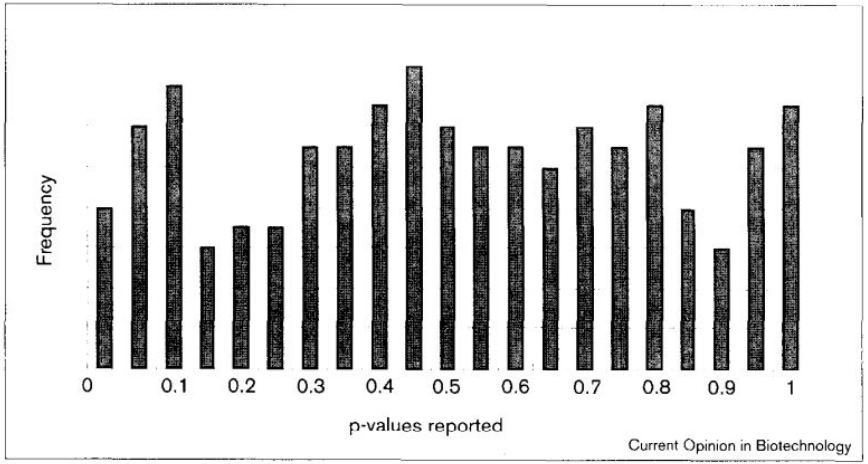
Only one was significant.

Statistical Power (A needle in a haystack)

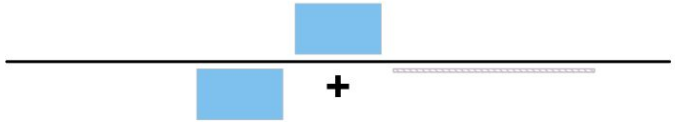
- Terwilliger and Weiss (1998)

Figure 4

The distribution of all reported p-values from association studies in either *American Journal of Medical Genetics* (*Neuropsychiatric Genetics*) or *Psychiatric Genetics* in 1997 is shown. A total of 222 reported p-values are graphed in the figure, and an additional 39 tests were listed as 'nonsignificant' at the 0.05 level with no statistical details in the manuscript. If all of the results were obtained under the null hypothesis, the expected distribution would be uniform. As can be seen in this figure, there is very good fit to the uniform expectation ($\chi^2_{(20)} = 12.98$; $p > 0.87$), indicating that the published p-values are consistent with the absence of gene effects in all the published analyses.

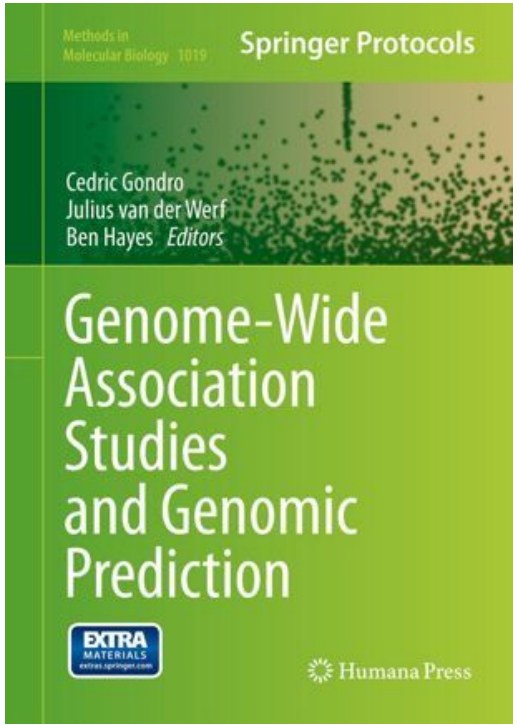


$$P(\text{Real effect}|\text{Signif.}) = \frac{P(\text{Real effect} \cap \text{Signif.})}{P(\text{Signif.})} = \frac{0.24}{0.275} = 0.873$$



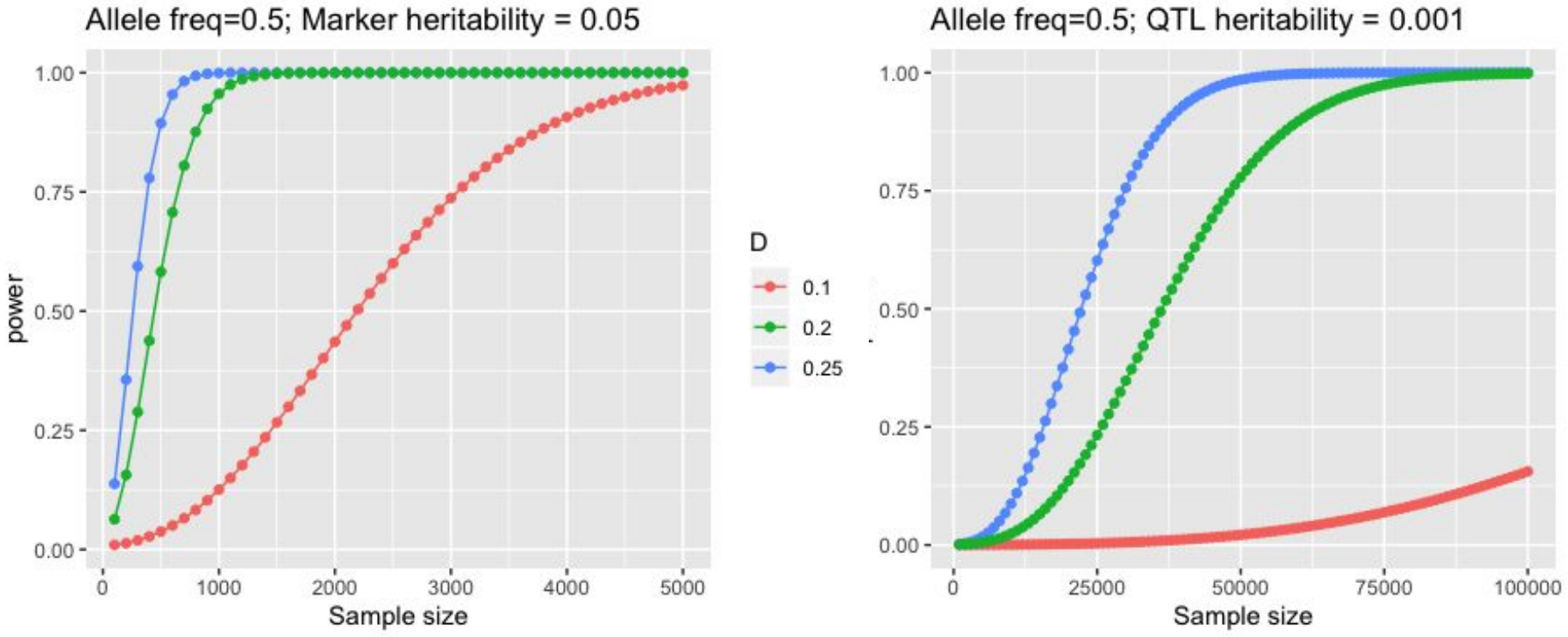
http://shiny-eio.upc.edu/bne/efectos2/

Statistical Power (A needle in a haystack)



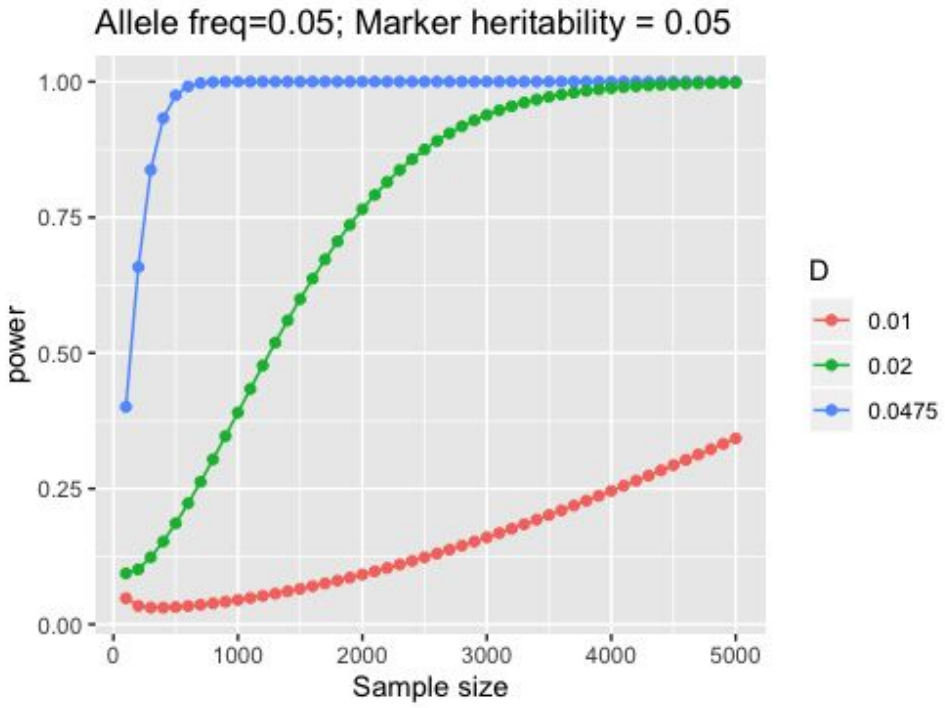
Chapter 3

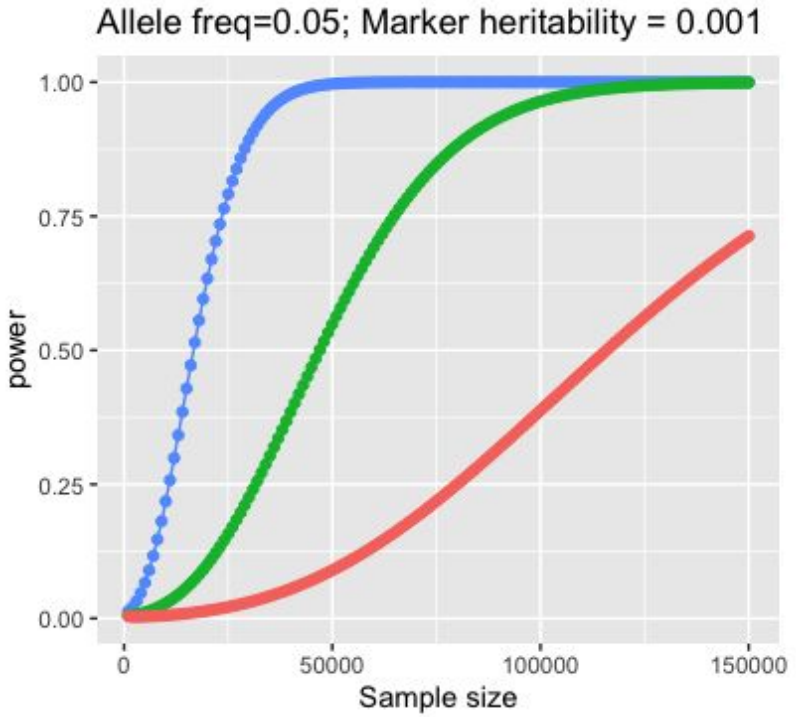
power of GWAS experiments (common variants)



R. D. Ball (2013) Genomewide Association studies and Genome-Wide prediction.

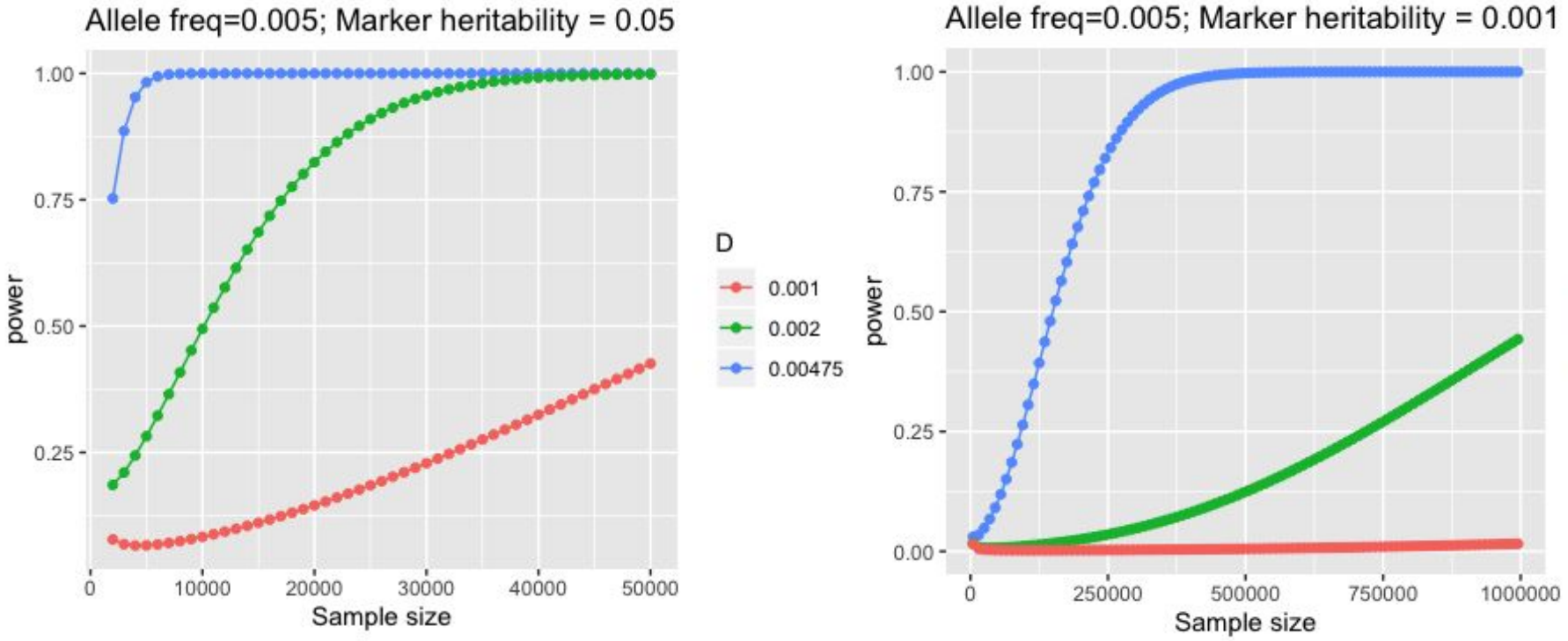
power of GWAS experiments (uncommon variants)





R. D. Ball (2013) Genomewide Association studies and Genome-Wide prediction.

power of GWAS experiments (rare variants)



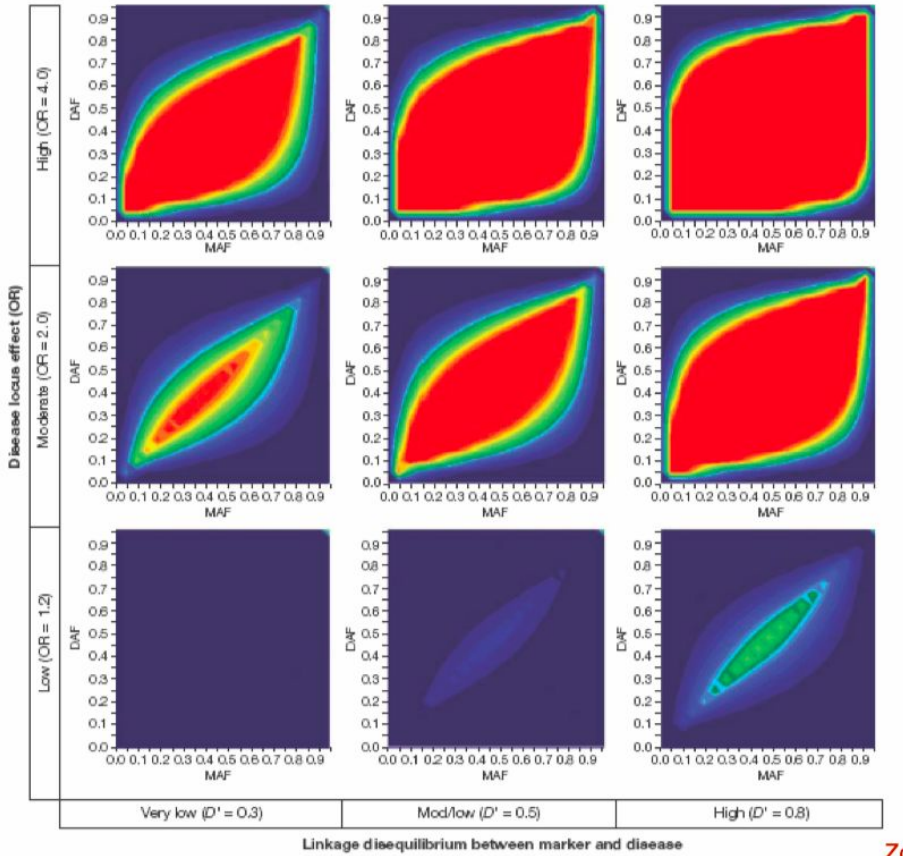
R. D. Ball (2013) Genomewide Association studies and Genome-Wide prediction.

power of GWAS experiments

- Sample size
- Magnitude of effect
- MAF marker
- MAF qtl
- Range of LD
- Likelihood of the model
- Experimental design

power of GWAS experiments

- Sample size
- Magnitude of effect
- MAF marker
- MAF qtl
- Range of LD
- Likelihood of the model
- Experimental design



Zondervan and Cardon (2004)

power of GWAS experiments

- MAIN DIFFICULTIES
- Low **effect size**
- Low LD

- When the allele-frequencies are mismatched, power is dramatically reduced
- We never know true disease allele frequency, so having a range of allele frequencies across markers is helpful

power of GWAS experiments

introduction_to_gwas/5.power_and_significance/StatisticalPower_exercise.R

power of bacterial GWAS experiments

- Main difficulties
  - All of the common GWAS
  - small sample size
  - high LD (reduces the number of independent tests)
  - Causal mutation that appear only at a handful of vertices

- Unique characteristics of bacterial populations
  - Clonal reproduction
  - Strong population stratification
  - Varying degrees of recombinations

- Even larger proportion of Type I errors

power of bacterial GWAS experiments

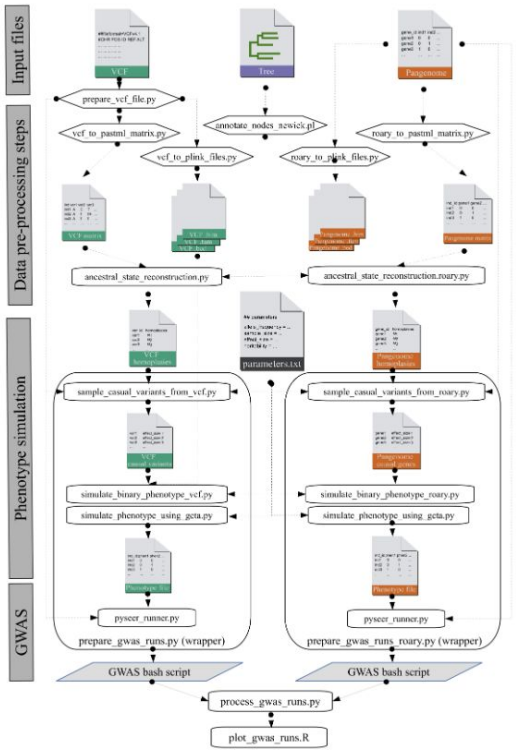
https://github.com/francesccoll/powerbacgwas

PowerBacGWAS: Power calculations for Bacterial GWAS

Usage

Overview

Input files, steps and scripts used to implement PowerBacGWAS pipeline:



power of bacterial GWAS experiments

Recommendations for GWAS

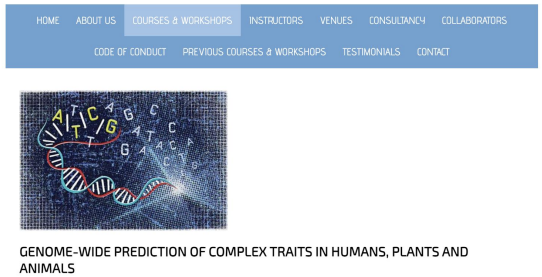
- Use gene level variation:
  - haploid SNPs
  - indels
  - mobile genetic elements (MGE)
  - insertion sequences
  - plasmids and conjugative elements
  - Copy Number Variations (CNVs)
  - Sequence Inversions
- Use k-mers as variables
- Replication analysis in the lab is “easy”. ➡ **Validate your results in an independent trial!**

power of bacterial GWAS experiments

Recommendations for GWAS

- Move to polygenic methods and work in PRS framework (each genomic variant is assigned an effect which are then summed across the genome)

Course 49

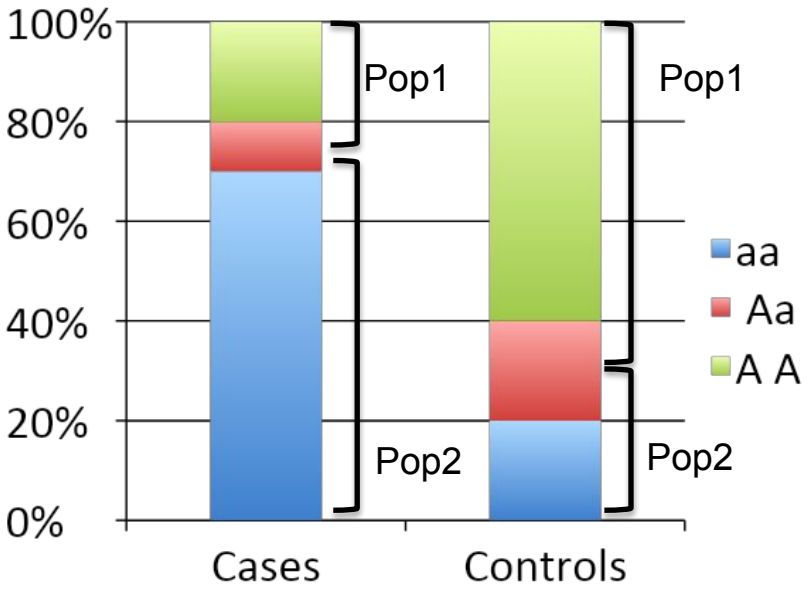


power of bacterial GWAS experiments

bacterial-GWAS are in its infancy, and many challenges need to be addressed

POPULATION STRATIFICATION

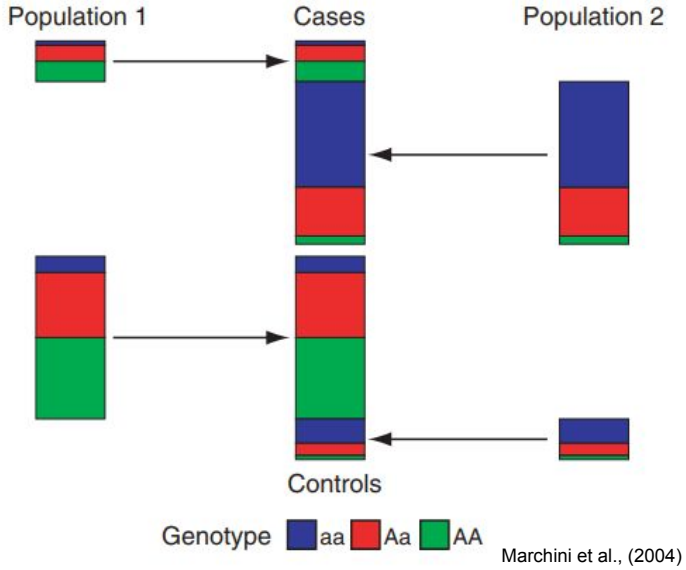
Population stratification



- Distribution of genotypes differs between cases and controls
- Might conclude that allele A (or genotype AA) related to disease

Population stratification

Separated populations



Marchini et al., (2004)

- If cases and controls not well-matched ethnically:
  - Disease is more common in one population than another AND
  - allele frequencies differ between populations at markers (chromosomal positions) unrelated to outcome
- Any allele more common in population with increased risk of disease may appear to be associated with disease

Population stratification

- Causes.

Unequal distribution of alleles may result from:

  - Subpopulations: sample made up of more than one distinct population, and cases and controls made up of different proportion of one or more of the subpopulations

  - Admixture: sample made up of individuals each with differing fractions of ancestry from distinct populations (e.g. African American individuals with differing levels of Caucasian ancestry) and the average level of admixture differs between cases and controls

Population stratification

23andMe

Population stratification

- What can you do about this?
  - Carefully match cases and controls in the data collection stage
    - Easy at gross level, difficult when structure is subtle
    - Recommend recording grandparental birthplace
  - Use a family-based study design
  - Detect and/or control for stratification in analysis stage

Population stratification

- Methods to detect and control for stratification
  - Test for stratification
  - Genomic control (GC)
  - Structured analysis (SA)
  - Principal components
  - Genetic covariance structure (Kinship matrix)

Population stratification

- Methods to detect and control for stratification
  - Test for stratification
    - Rationale: if stratification exists, should see association with disease status across many loci, not just the candidate locus of interest

1. Genotype a set of *L* unlinked (independent) marker loci in all samples of study (e.g. 15 to 20 polymorphic markers).
2. Assume markers are chosen randomly so that the probability that any one marker is tightly linked to a disease locus is very low.
3. Compare differences between genotype or allele frequencies between cases and controls using the χ^2 test for each marker.
4. To test for stratification, sum individual marker test statistics to get final χ^2 statistic with *L* degrees of freedom.

Population stratification

- Methods to detect and control for stratification
  - Test for stratification
    - H_0: the allele frequencies at each of the marker loci are the same in the case and control groups.
    - H_A: the allele frequency distribution across the *L* loci differs between the case and control groups.

    - Strategy attractive for its simplicity and ability to formally test, but it doesn't offer a solution.

(Pritchard, JK and Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65: 220-228)

Population stratification

- Methods to detect and control for stratification
  - Genomic Control
    - Rationale: if population stratification exists in the study population then it should be present at many markers across the genome
    - Strategy:

1. Genotype a set of *L* unlinked (independent) marker loci in all samples of study (e.g. 20 to 50 SNPs)
2. Assume markers are chosen randomly so that the probability that any one marker is tightly linked to a disease locus is very low
3. Compare differences between genotype or allele frequencies between cases and controls using the χ^2 test for each marker.
4. Adjust test statistic for association based on the inflation (λ) seen across a collection of unlinked (independent) markers

Population stratification

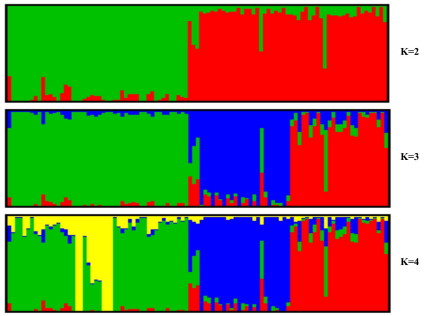
- Methods to detect and control for stratification
  - Genomic Control
    - Assume constant inflation factor (λ) across genome (it is thought to be true regardless [almost] of allele frequency, but cannot be generalized)

    - If λ = 1 then no population stratification
    - If λ < 1 then set to 1 (bounded λ)

Adjust χ 2 statistic for each markers of interest: $\chi^2 = \frac{\chi^2}{\hat{\lambda}}$

(Devlin B and Roeder K (1999) Genomic control for association studies. Biometrics 55:369-387)

Population stratification



S1 Figure. Admixture plots of K=2-4 showing population structure of different Angora sub-populations

- Methods to detect and control for stratification
  - Structured analysis
    - Rationale: genotype data on a sample of individuals for many markers can give us information about sub populations in the sample
    - Strategy: assign and control for population membership, performing analysis in each population (*reduce statistical power*)
    - Performed in STRUCTURE or STRAT type analyses

(Pritchard JK, Stephens M and Donnelly P (1999) Inference of population structure using multilocus genotype data. Genetics 155: 945-959)

Population stratification

- Methods to detect and control for stratification
  - Principal Component Analysis
    - Rationale: Similar to SA approach in that goal is to use many markers to capture variation that is due to ancestry (*GWAS lot of markers*)
    - Strategy:

1. Get a covariate (or set of covariates) for each individual that represents their genetic ancestry from PC.
2. Adjust phenotype and genotypes for ancestry based on the covariate that represents ancestry (*Test association of PC to detect PS, or visualization*)
3. Adjust phenotype (outcome) and genotype (predictor) for ancestry (PC), in the association test as usual

Price et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38:904-909

Population stratification

- Methods to detect and control for stratification
  - Adjust Genetic Covariance (Kinship matrix).
    - Rationale:Similar to PCA approach, but uses the whole genomic variability in the data. Additionally, it adjust the model by an overall genomic background. (need *lot of markers to be efficient*)
    - Strategy:

1. Calculate a genomic relationship matrix (GRM)

$$A_{jk} = \frac{1}{N}\sum_i A_{ijk} = \begin{cases} \frac{1}{N}\sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}, & j \neq k \\ 1 + \frac{1}{N}\sum_i \frac{x_{ij}^2 - (1+2p_i)x_{ij} + 2p_i^2}{2p_i(1-p_i)}, & j = k \end{cases} \quad (6)$$

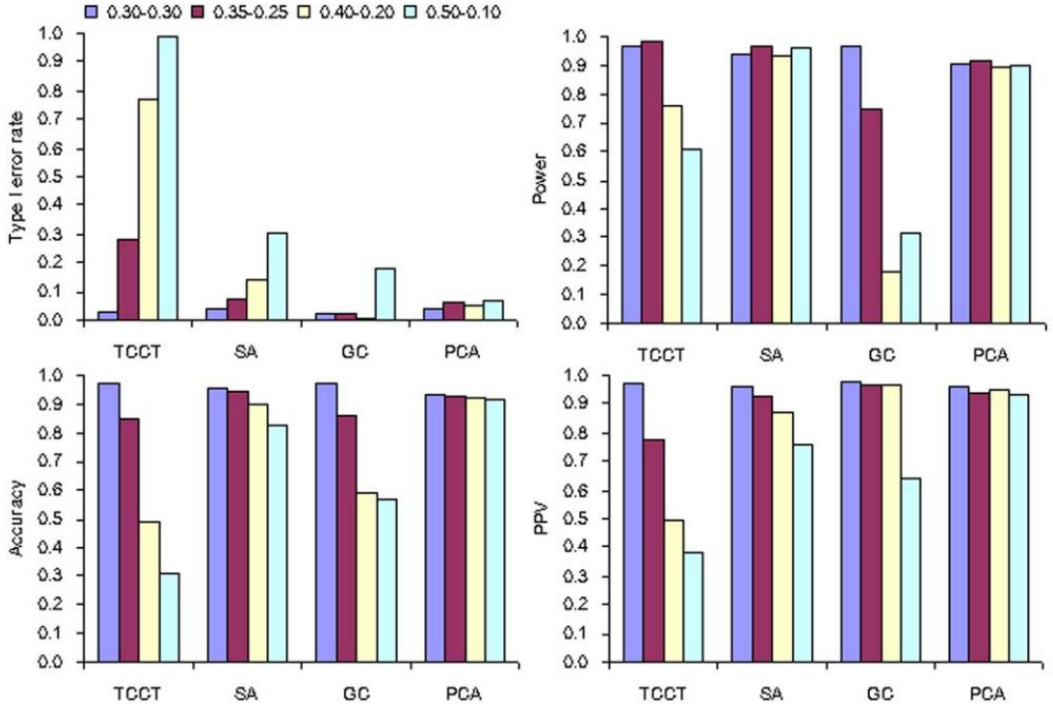
2. Include a polygenic effect in a mixed model taking into account GRM

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad \mathbf{u} \sim N(0, \mathbf{A}\sigma_u^2)$$

3. Alternative, PC from the GRM and implement PCA strategy

Population stratification

- Comparison of methods
  - stratification level



Zhang F, Wang Y, Deng H-W (2008) Comparison of Population-Based Association Study Methods Correcting for Population Stratification. PLoS ONE 3(10): e3392. doi:10.1371/journal.pone.0003392

Figure 1. Performance of the four analytical methods in stratified populations with stratification levels varying from 0.3−0.3 to 0.5−0.1 (sample size = 1200, frequency of disease susceptible allele = 0.20±0.02 and number of AIMs = 40).
doi:10.1371/journal.pone.0003392.g001

Population stratification

- Comparison of methods
  - sample size

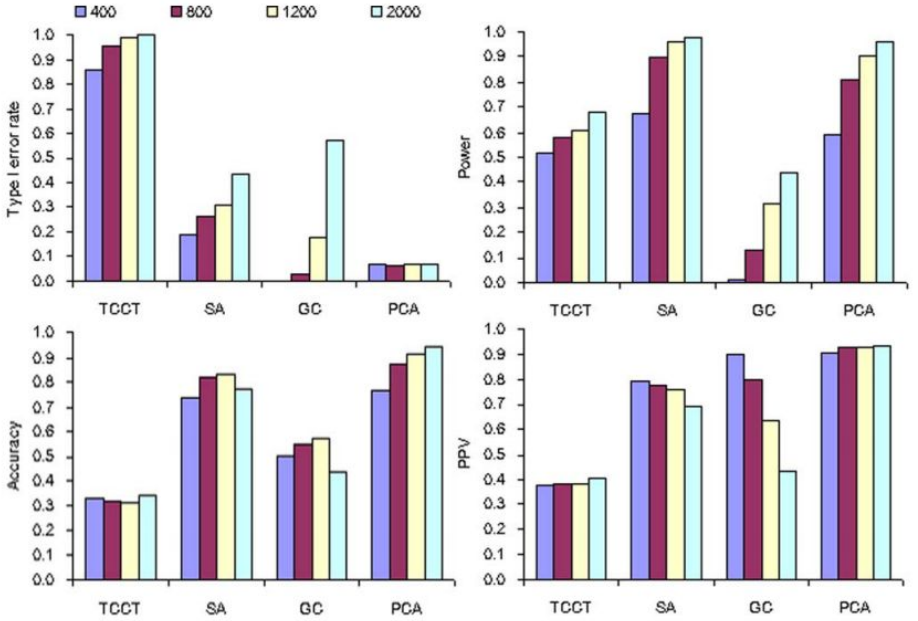
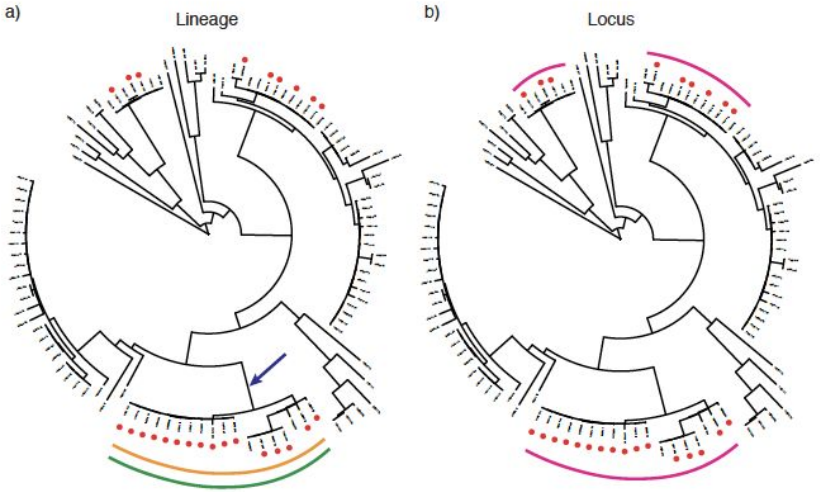


Figure 2. Performance of the four analytical methods in stratified populations with sample sizes varying from 400 to 2000 (stratification level = 0.5−0.1, frequency of disease susceptible allele = 0.20±0.02 and number of AIMs = 40).
doi:10.1371/journal.pone.0003392.g002

Zhang F, Wang Y, Deng H-W (2008) Comparison of Population-Based Association Study Methods Correcting for Population Stratification. PLoS ONE 3(10): e3392. doi:10.1371/journal.pone.0003392

Population stratification -bacterial GWAS

- Haplotype organisms
  - entire chromosome is clonally copied. LD extends across entire genome
  - all sites are perfectly correlated (not possible to know which mutation is causal)
- causal mutation from an ancestral phylogeny branch, is not distinguible of new non-causal mutations (aka lineage association).



Population stratification -bacterial GWAS

Traditional solutions

- Focus on variants within a relevant region
- Detect homoplasic variants (appeared in several lineges)
- Find recombinants that are independent of genetic background (lineages)

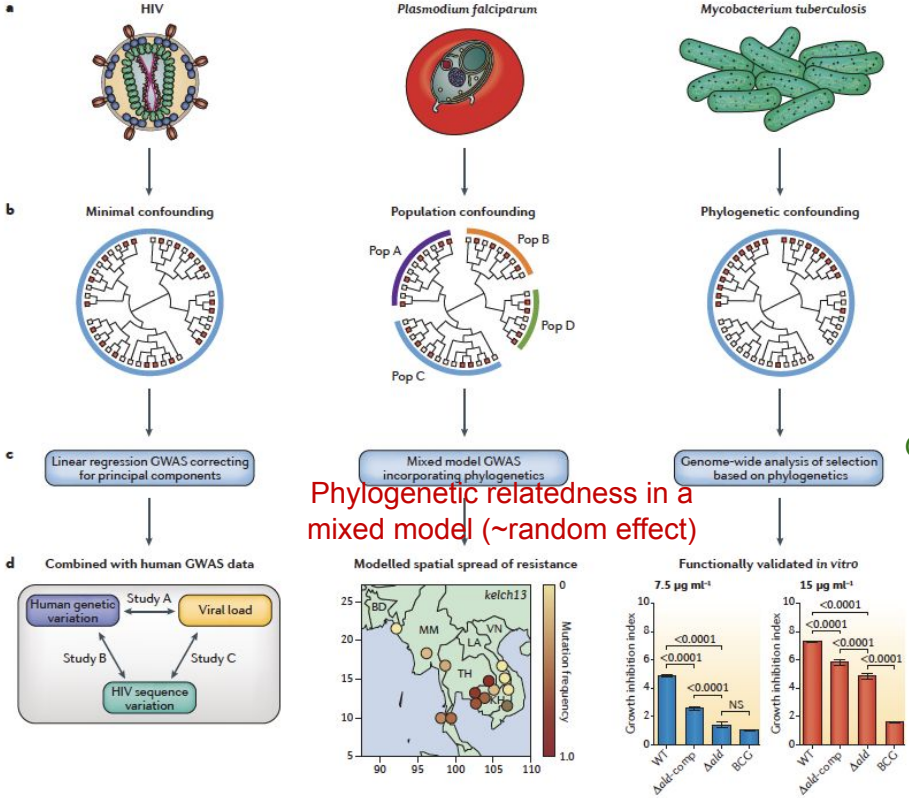
Recommendations for GWAS

- Genomic control or Cochran-Mantel-Haenszel (CMH) test
- use PCA as covariates (may not work for recombinants from distant ancestry)
  - Still confounders that require additional methods

Population stratification -bacterial GWAS

1. Homologous recombination
   a. Long range LD may still exist even after adjusting for PCA
   b. Admixture methods that make use of recombination patterns may help
2. Selection
   a. Microb population structure may be due to adaptation (e.g. AMR)
   b. Can lead to panmictic populations (with lots of recombinants)
   c. Use longitudinal samples (unless the phenotype of interest is longitudinal itself, e.g. time to disease sympthoms, because it's adding confounders)
   d. Use mixed models (GRM) that account for relatedness.

Population stratification -bacterial GWAS

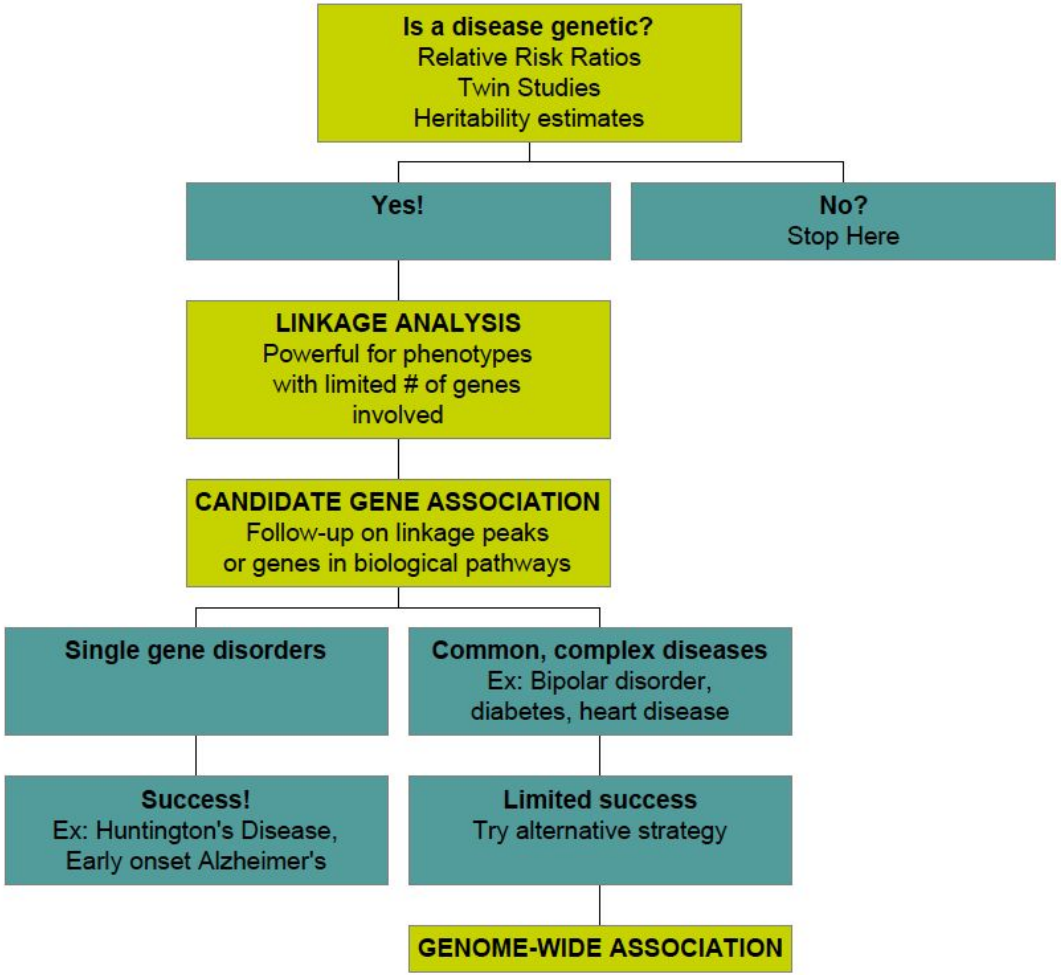


PCA

Phylogenetic relatedness in a mixed model (~random effect)

GWAS based on phylogenetics (as a random effect or GRM)

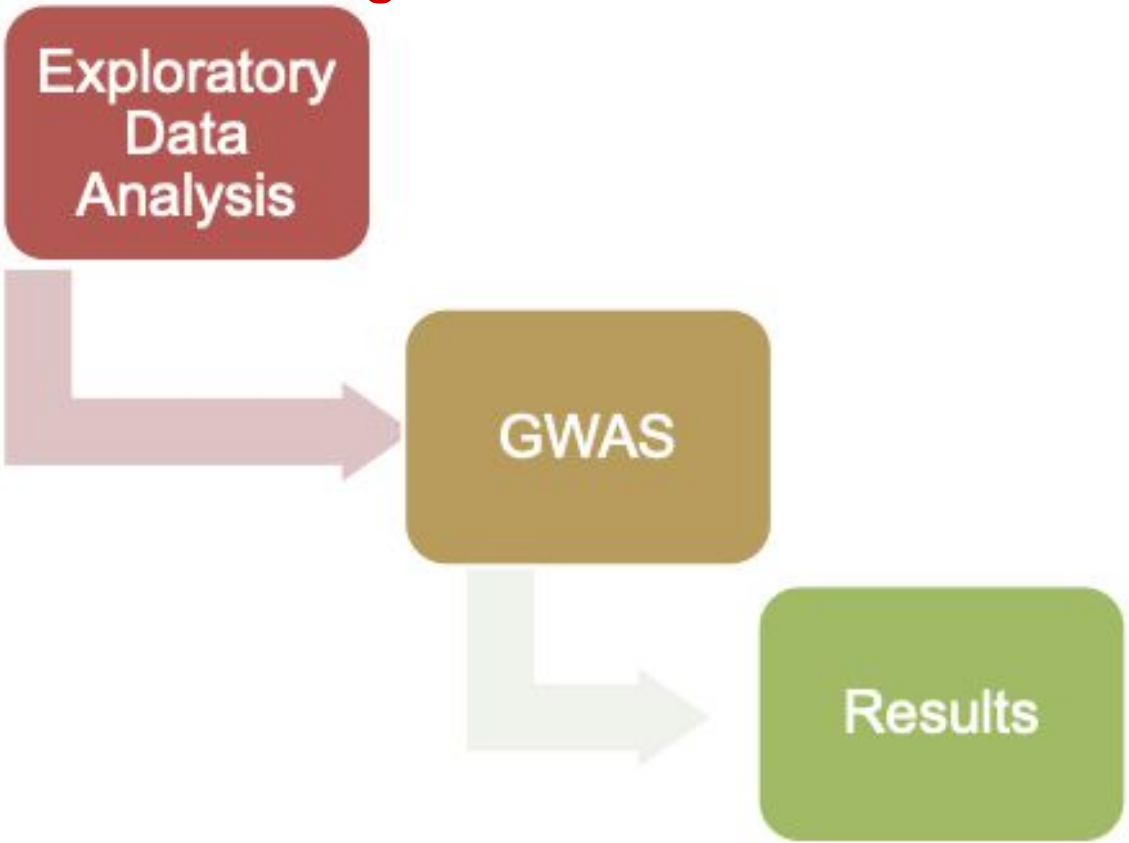
Experimental design



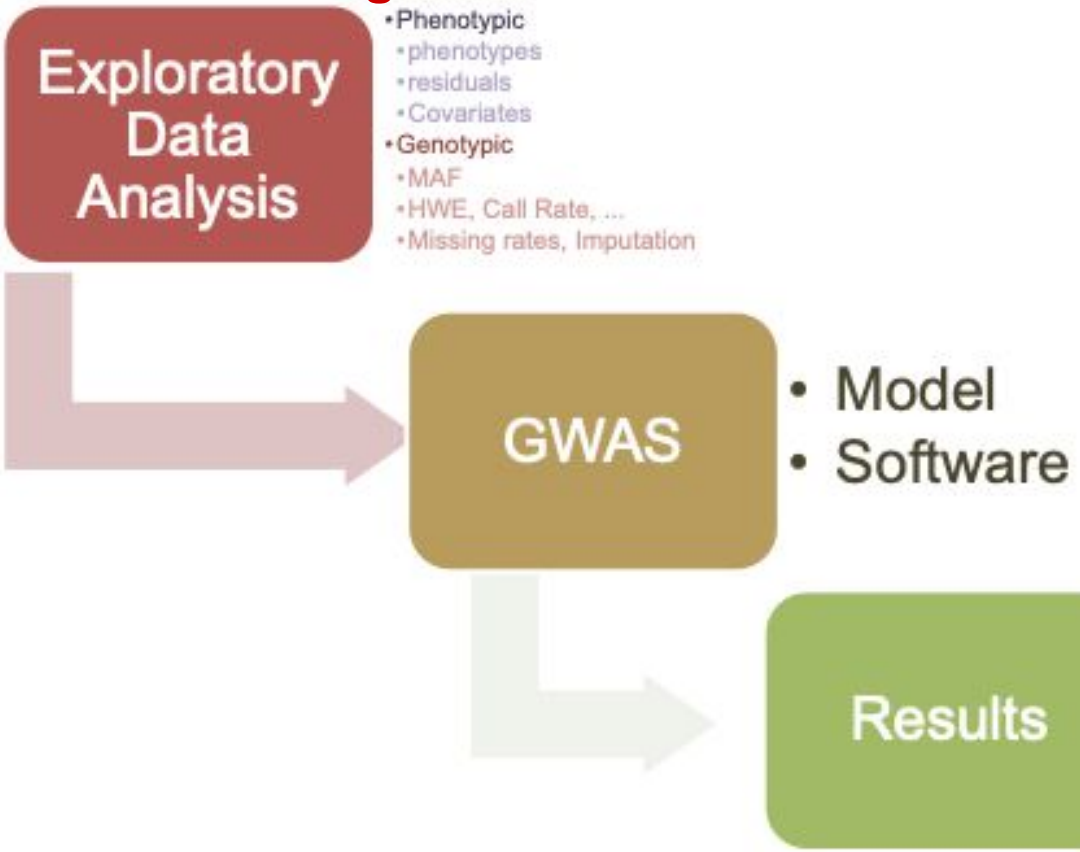
Experimental design

- Collaboration, collaboration, collaboration
- Collaboration with many disciplines epidemiologists, biostatisticians, bioinformatics, clinicians, geneticists, etc.
- Collaboration across funded research centers to pool and share results
- Analysis methods and techniques are constantly changing
- High dimensional data (p >>>>> n)
  - Large p, small n problem

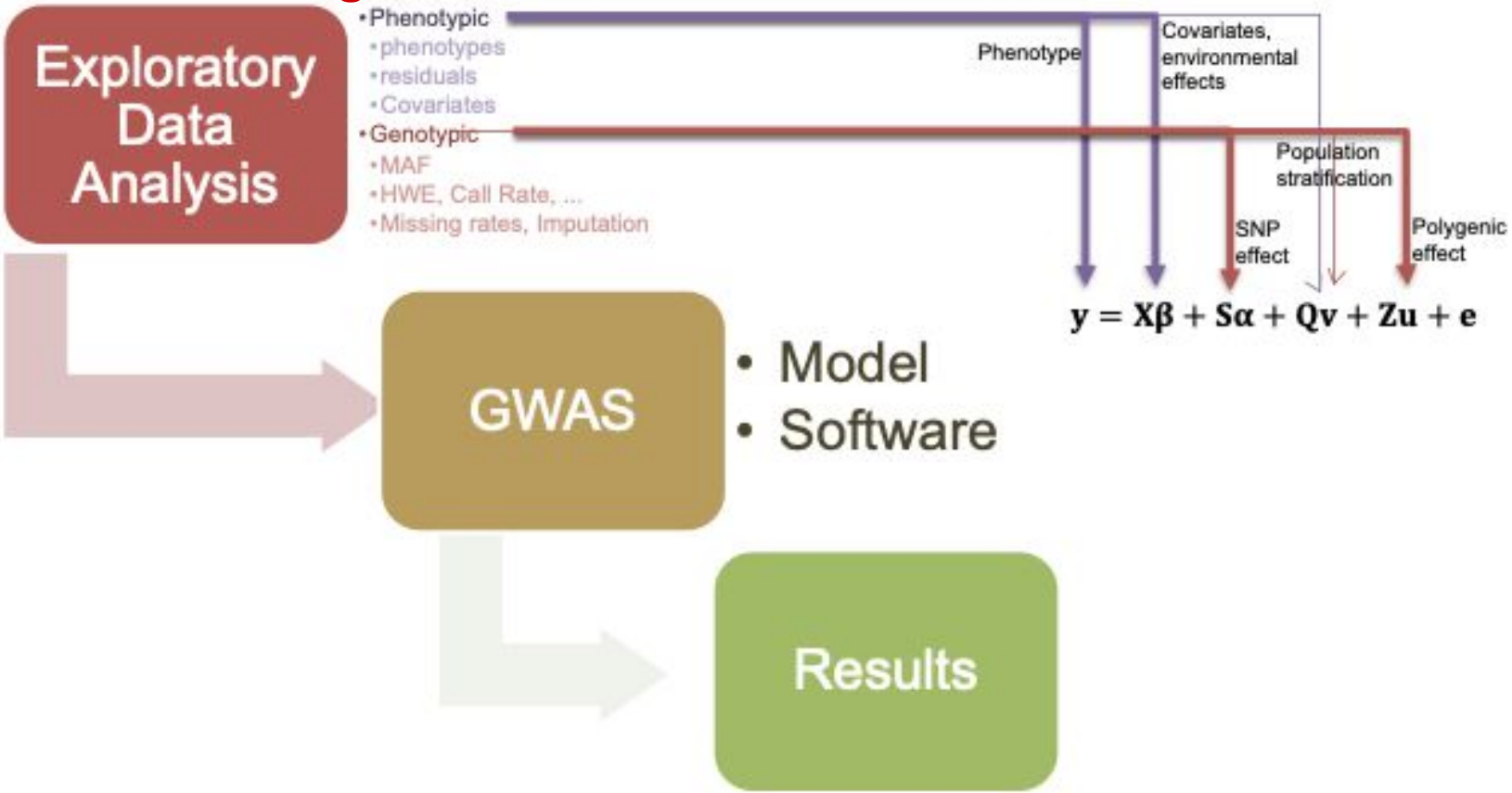
Experimental design



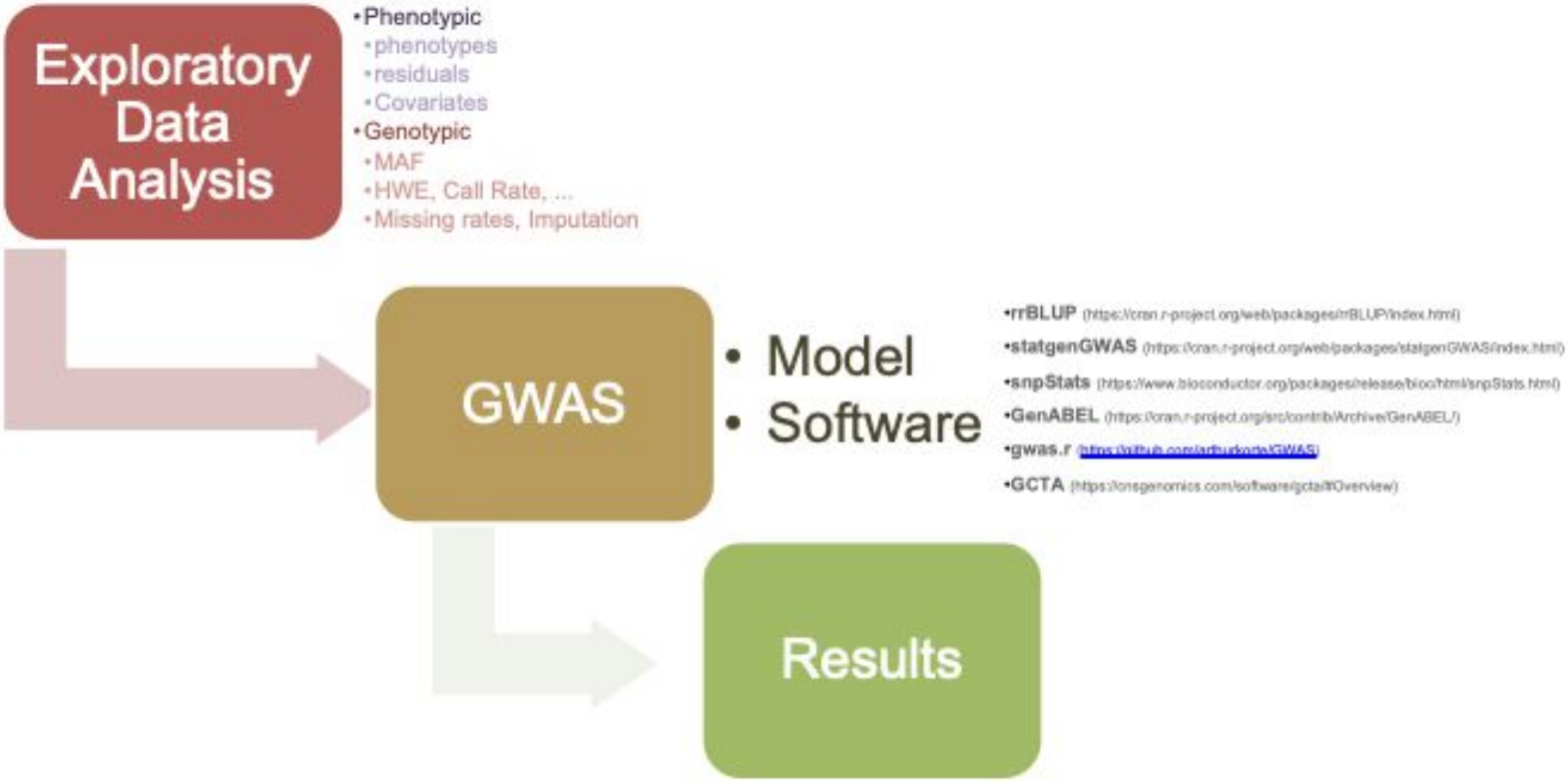
Experimental design



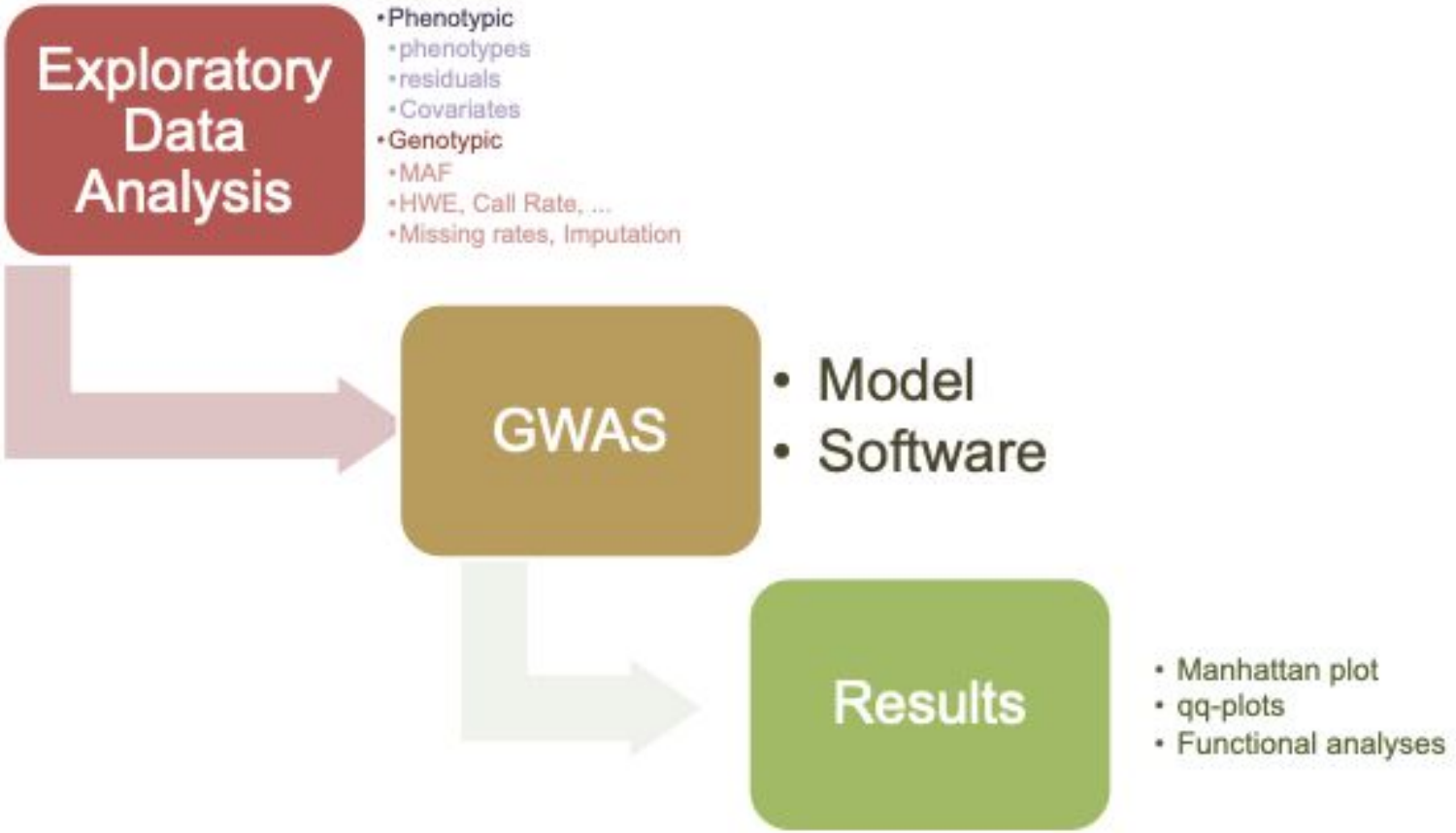
Experimental design



Experimental design



Experimental design



Experimental design

1. SNP QUALITY CONTROL

Filter process to end up with the highest quality set of SNPs

- SNP genotyping rates: usually >95%
- Sample call rates: usually >95%
- Minor allele frequency: ranges >0.002 - 0.05
- Hardy Weinberg Equilibrium: *p*-value threshold ranges $>10^{-3}$ - 10^{-6}
- Mendelian inconsistencies - need “trios”

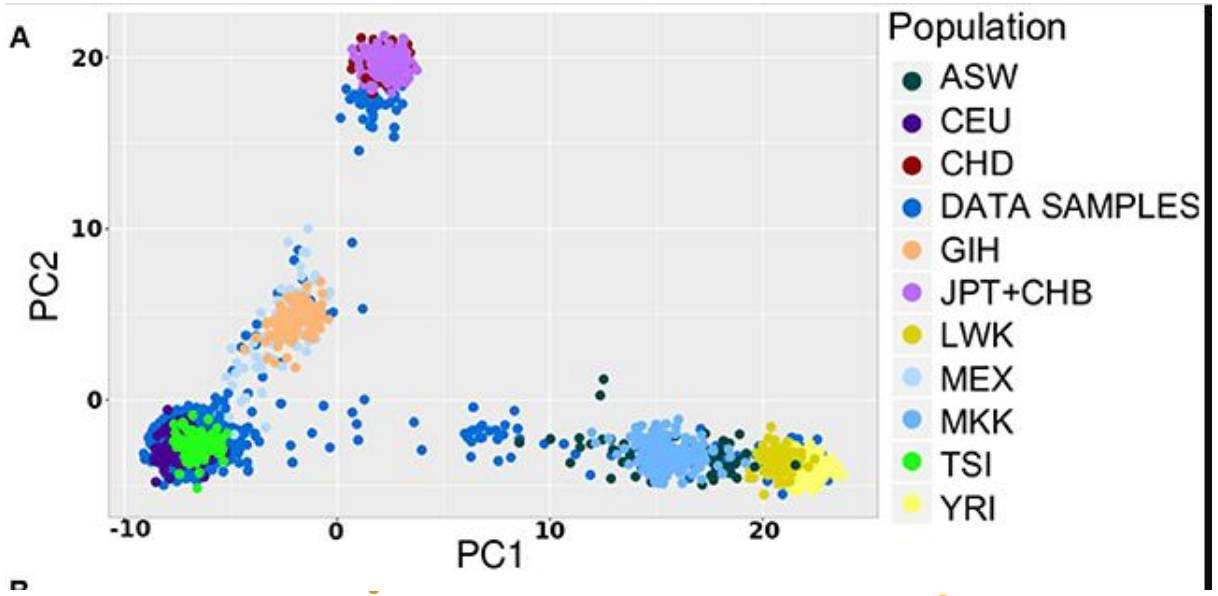
Retain approximately 75-80% (or more) of the SNPs

Experimental design

2. POPULATION STRATIFICATION

Classic issue of confounding

- Test for PS
- Correct for PS



Experimental design

3. ADJUSTMENT OF COVARIATES

Rationale for adjustment in GWAS

1. Confounding

2. Increase ability to predict outcome or explain more variation in the trait -- “what can genetics provide beyond established risk factors?”

What did everyone else do?

- Replication or comparing results across studies requires a similar analysis strategy and adjustment model.

Experimental design

4. MULTIPLE TESTING

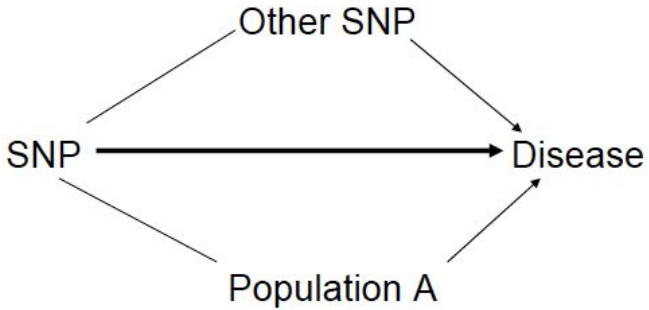
- Choose significance level
- Perform a correction test (Bonferroni, FDR,..)

Experimental design

5. INTERPRETING THE RESULTS

So you think you have a significant SNP?

- Direct causal relationship
- Indirect association - linkage disequilibrium
- Spurious association - population stratification or false positive



Experimental **design**

6. REPLICATION