Some aspects of Genome Wide Association Studies (GWAS)

Dirk Metzler

Statistical Genetics
Department Biologie II
http://evol.bio.lmu.de/_statgen

July 10, 2014

1. Intro to GWAS

2. Genetic Relationships
   - A simple approach: Genomic Control (GC)
   - Structured Association (SA)
   - Regression Control
   - Principal Component (PC) Adjustment
   - Estimating kinship
   - Mixed Regression Models

W. Astle, D.J. Balding (2009) Population Structure and Cryptic Relatedness in Genetic Association Studies
Statistical Science **24(4)**, 451–471

Outline

1. Intro to GWAS

2. Genetic Relationships
   - A simple approach: Genomic Control (GC)
   - Structured Association (SA)
   - Regression Control
   - Principal Component (PC) Adjustment
   - Estimating kinship
   - Mixed Regression Models

Aim of GWAS

Sample of individuals; given data for all individual:

- Many SNPs spead over the whole genome
- Phenotypic trait of interest

Aim of GWAS

Sample of individuals; given data for all individual:

- Many SNPs spead over the whole genome
- Phenotypic trait of interest
- Maybe information about relatedness of individuals

Aim of GWAS

Sample of individuals; given data for all individual:

- Many SNPs spead over the whole genome
- Phenotypic trait of interest
- Maybe information about relatedness of individuals
- Maybe data on other traits or environmental factors that may influence the trait

Aim of GWAS

Sample of individuals; given data for all individual:

- Many SNPs spead over the whole genome
- Phenotypic trait of interest
- Maybe information about relatedness of individuals
- Maybe data on other traits or environmental factors that may influence the trait

Question: Which SNPs have an influence on the phenotypic trait?

Possible problems

- correlations btw causal factors and (unlinked) non-causal factors

Possible problems

- correlations btw causal factors and (unlinked) non-causal factors
  - population structure (due to large sample sizes even modest structure can lead to false positives)

Possible problems

- correlations btw causal factors and (unlinked) non-causal factors
  - population structure (due to large sample sizes even modest structure can lead to false positives)
  - pleiotropy: e.g. if there is **selection for skin color**, locus A influences skin color, locus B influences skin color and eye color, then **GWAS for eye color** detects both A and B!

Possible problems

- correlations btw causal factors and (unlinked) non-causal factors
    - population structure (due to large sample sizes even modest structure can lead to false positives)
    - pleiotropy: e.g. if there is **selection for skin color**, locus A influences skin color, locus B influences skin color and eye color, then **GWAS for eye color** detects both A and B!
- more than one causal factor

Possible problems

- correlations btw causal factors and (unlinked) non-causal factors
    - population structure (due to large sample sizes even modest structure can lead to false positives)
    - pleiotropy: e.g. if there is **selection for skin color**, locus A influences skin color, locus B influences skin color and eye color, then **GWAS for eye color** detects both A and B!
- more than one causal factor
- ascertainment bias (e.g. cases are sampled from some clinic, controls somewhere else)

Possible problems

- correlations btw causal factors and (unlinked) non-causal factors
  - population structure (due to large sample sizes even modest structure can lead to false positives)
  - pleiotropy: e.g. if there is **selection for skin color**, locus A influences skin color, locus B influences skin color and eye color, then **GWAS for eye color** detects both A and B!
- more than one causal factor
- ascertainment bias (e.g. cases are sampled from some clinic, controls somewhere else)
- more markers than repetitions (“$n \ll p$ problem”)

Some free GWAS software packages

- PLINK
  `http://pngu.mgh.harvard.edu/~purcell/plink/`
- R packages
  - GWASTools
    `http://bioconductor.org/packages/release/bioc/html/GWASTools.html`
    Bioconductor package, install in R with

    ```
    source("http://bioconductor.org/biocLite.R")
    biocLite("GWASTools")
    ```
  - GenABEL etc.
    `http://genabel.org/packages`
    CRAN package, install in R with

    ```
    install.packages("GenABEL")
    ```

Outline

Different scenarios

Whether we have to compensate for relatedness in the data depends on the where the indivuduals come from.

- Crossing scheme (e.g. in plant breeding): Individuals are F1 (or Fn) generation of two homozygous individuals
- Pedigree is known (up to possible errors)
- Individuals are somehow related but pedigree is unknown
- Individuals are sampled from large population, but there may be some population structure

Outline

1. Intro to GWAS

2. Genetic Relationships
   - A simple approach: Genomic Control (GC)
   - Structured Association (SA)
   - Regression Control
   - Principal Component (PC) Adjustment
   - Estimating kinship
   - Mixed Regression Models

Genomic Control (GC): Fast and simple method to compensate for population structure or cryptic relatedness.

Genomic Control (GC): Fast and simple method to compensate for population structure or cryptic relatedness.

Main idea is to multiply test statistic with constant λ to make it fit χ^2_1 distribution.

Genomic Control (GC): Fast and simple method to compensate for population structure or cryptic relatedness.

Main idea is to multiply test statistic with constant λ to make it fit χ_1^2 distribution.

The test statistic is T^2/V, where T measures for a locus the difference in allele frequencies between cases and controls, and V approximates the variance of T for the case of neutrality an unrelated samples. Under the latter conditions, T^2/V is approximately χ_1^2-distributed.

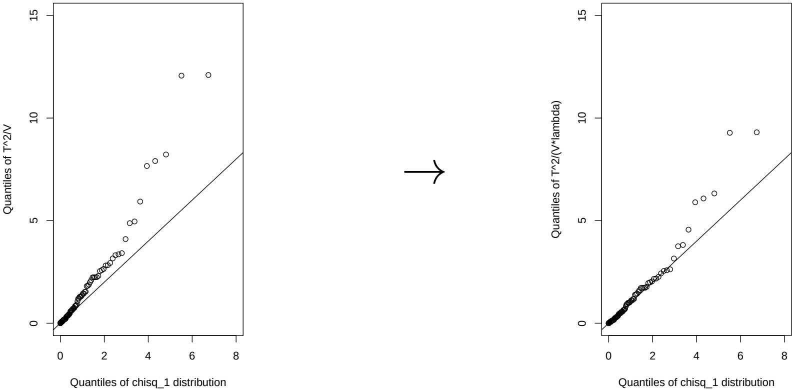
Genomic Control (GC): Fast and simple method to compensate for population structure or cryptic relatedness.

Main idea is to multiply test statistic with constant λ to make it fit χ^2_1 distribution.

The test statistic is T^2/V, where T measures for a locus the difference in allele frequencies between cases and controls, and V approximates the variance of T for the case of neutrality an unrelated samples. Under the latter conditions, T^2/V is approximately χ^2_1-distributed.

Fitting λ is based on the assumption that only few SNPs are in strong causal association with the test statistic.

Instead of T^2/V use $T^2/(\lambda \cdot V)$, where λ is chosen to make the distribution fit χ^2_1 (up to outliers).



The outliers are candidate loci to be associated with the trait.

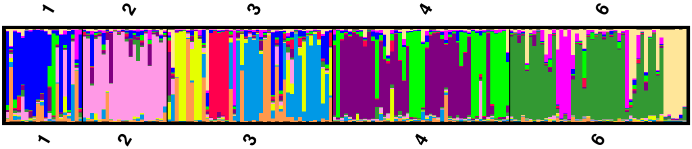
Outline

1. Intro to GWAS

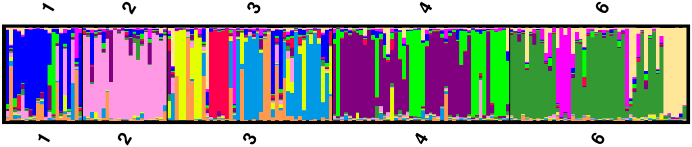
2. Genetic Relationships
   - A simple approach: Genomic Control (GC)
   - Structured Association (SA)
   - Regression Control
   - Principal Component (PC) Adjustment
   - Estimating kinship
   - Mixed Regression Models

- Software: e.g. PLINK
- SA assumes that population consists of subpopulations (“islands”)

- Software: e.g. PLINK
- SA assumes that population consists of subpopulations (“islands”)
- Population structure can be estimated from $\sim$ 100 SNPs e.g. with software STRUCTURE, assuming that each island is in Hardy-Weinberg equilibrium

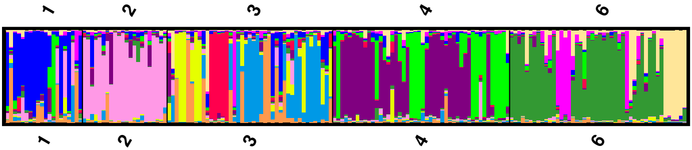


- Software: e.g. PLINK
- SA assumes that population consists of subpopulations (“islands”)
- Population structure can be estimated from $\sim$ 100 SNPs e.g. with software STRUCTURE, assuming that each island is in Hardy-Weinberg equilibrium



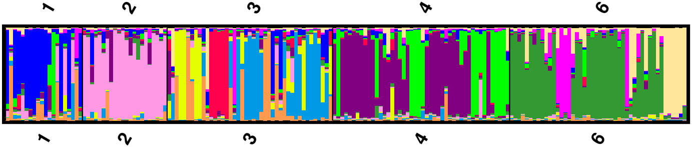
- with “admiture” option, individual genomes are admixed from different island

- Software: e.g. PLINK
- SA assumes that population consists of subpopulations ("islands")
- Population structure can be estimated from $\sim$ 100 SNPs e.g. with software STRUCTURE, assuming that each island is in Hardy-Weinberg equilibrium



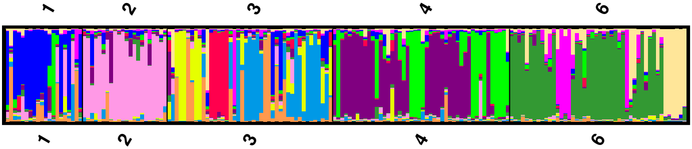
- with "admiture" option, individual genomes are admixed from different island
- stratified tests are applied, i.e. search for significant associations of trait and loci within the islands

- Software: e.g. PLINK
- SA assumes that population consists of subpopulations (“islands”)
- Population structure can be estimated from $\sim$ 100 SNPs e.g. with software STRUCTURE, assuming that each island is in Hardy-Weinberg equilibrium



- with “admiture” option, individual genomes are admixed from different island
- stratified tests are applied, i.e. search for significant associations of trait and loci within the islands
- island model is not always suitable for human populations

- Software: e.g. PLINK
- SA assumes that population consists of subpopulations (“islands”)
- Population structure can be estimated from $\sim$ 100 SNPs e.g. with software STRUCTURE, assuming that each island is in Hardy-Weinberg equilibrium



- with “admiture” option, individual genomes are admixed from different island
- stratified tests are applied, i.e. search for significant associations of trait and loci within the islands
- island model is not always suitable for human populations
- SA does not explicitely account for pedigree-level relationships

Outline

1. Intro to GWAS

2. Genetic Relationships
   - A simple approach: Genomic Control (GC)
   - Structured Association (SA)
   - Regression Control
   - Principal Component (PC) Adjustment
   - Estimating kinship
   - Mixed Regression Models

- GLM with phenotypic trait as target variable
- use $\sim$ 100 widely spaced, putatively neutral SNPs as regression covariates
- these covariates are informative about the underlying pedigree and are supposed to eliminate its effect in regression-based test with locus of interest

- GLM with phenotypic trait as target variable
- use $\sim$ 100 widely spaced, putatively neutral SNPs as regression covariates
- these covariates are informative about the underlying pedigree and are supposed to eliminate its effect in regression-based test with locus of interest
- to avoid overfitting apply backward selection and regularization (shrinkage) to these covariates

- GLM with phenotypic trait as target variable
- use $\sim$ 100 widely spaced, putatively neutral SNPs as regression covariates
- these covariates are informative about the underlying pedigree and are supposed to eliminate its effect in regression-based test with locus of interest
- to avoid overfitting apply backward selection and regularization (shrinkage) to these covariates
- in absence of ascertainment bias similar performance as GC and SA

- GLM with phenotypic trait as target variable
- use $\sim$ 100 widely spaced, putatively neutral SNPs as regression covariates
- these covariates are informative about the underlying pedigree and are supposed to eliminate its effect in regression-based test with locus of interest
- to avoid overfitting apply backward selection and regularization (shrinkage) to these covariates
- in absence of ascertainment bias similar performance as GC and SA
- computationally faster than SA
- more robust to ascertainment bias than GC

- GLM with phenotypic trait as target variable
- use $\sim$ 100 widely spaced, putatively neutral SNPs as regression covariates
- these covariates are informative about the underlying pedigree and are supposed to eliminate its effect in regression-based test with locus of interest
- to avoid overfitting apply backward selection and regularization (shrinkage) to these covariates
- in absence of ascertainment bias similar performance as GC and SA
- computationally faster than SA
- more robust to ascertainment bias than GC
- allow flexibility of regression methods

Outline

1. Intro to GWAS

2. Genetic Relationships
   - A simple approach: Genomic Control (GC)
   - Structured Association (SA)
   - Regression Control
   - Principal Component (PC) Adjustment
   - Estimating kinship
   - Mixed Regression Models

Principal Component Adjustment

- similar to regression control, but uses PCA (instead of backward selection and regularization) to avoid overfitting
- well-founded for island models
- not clear how well it works for more complex cryptic relatedness

Outline

1. Intro to GWAS

2. Genetic Relationships
   - A simple approach: Genomic Control (GC)
   - Structured Association (SA)
   - Regression Control
   - Principal Component (PC) Adjustment
   - Estimating kinship
   - Mixed Regression Models

Kinship coefficients based on marker data

Kinship coefficient K_{ij} of two individuals i and j: probability of two alleles, one drawn from i and the other drawn from j are identical by descent (IBD), i.e. both stem from the same *recent* ancestor.

Kinship coefficients based on marker data

Kinship coefficient K_{ij} of two individuals *i* and *j*: probability of two alleles, one drawn from *i* and the other drawn from *j* are identical by descent (IBD), i.e. both stem from the same *recent* ancestor.

If *p* is the frequency of allele *A* and x_i and x_j count the A alleles (0,1, or 2) of *i* and *j*, then

$$\mathrm{Cov}(x_i, x_j) = 4p(1-p)K_{ij}.$$

Kinship coefficients based on marker data

Kinship coefficient K_{ij} of two individuals i and j: probability of two alleles, one drawn from i and the other drawn from j are identical by descent (IBD), i.e. both stem from the same *recent* ancestor.

If p is the frequency of allele A and x_i and x_j count the A alleles (0,1, or 2) of i and j, then

$$\mathrm{Cov}(x_i, x_j) = 4p(1-p)K_{ij}.$$

Thus, K_{ij} can be estimated from genome-wide covariances of allele counts:

$$\widehat{K}_{ij} = \frac{1}{L}\sum_{\ell=1}^{L}\frac{(x_{i\ell}-2p_\ell)\cdot(x_{j\ell}-2p_\ell)}{4p_\ell(1-p_\ell)}$$

where L is the number of loci and p_ℓ is the frequency of allele A at locus ℓ. (At each locus we choose one allele and call it A).

To refine the estimates of p_ℓ and K we can iteratively apply the formulas

$$\widehat{p}_\ell = \frac{\sum_{ij} \left(\widehat{K}^{-1}\right)_{ij} x_{j\ell}}{\sum_{ij} \left(\widehat{K}^{-1}\right)_{ij}}$$

and

$$\widehat{K}_{ij} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{(x_{i\ell} - 2\widehat{p}_\ell) \cdot (x_{j\ell} - 2\widehat{p}_\ell)}{4\widehat{p}_\ell(1 - \widehat{p}_\ell)}.$$

To refine the estimates of p_ℓ and K we can iteratively apply the formulas

$$\widehat{p}_\ell = \frac{\sum_{ij} \left(\widehat{K}^{-1}\right)_{ij} x_{j\ell}}{\sum_{ij} \left(\widehat{K}^{-1}\right)_{ij}}$$

and

$$\widehat{K}_{ij} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{(x_{i\ell} - 2\widehat{p}_\ell) \cdot (x_{j\ell} - 2\widehat{p}_\ell)}{4\widehat{p}_\ell(1 - \widehat{p}_\ell)}.$$

For human populations $\sim$ 100.000 SNPs are usually required to obtain reasonable estimates of K.

So far we have not accounted for LD btw. markers. This can be done with hidden-Markov models (HMMs).

Outline

1. Intro to GWAS

2. Genetic Relationships
   - A simple approach: Genomic Control (GC)
   - Structured Association (SA)
   - Regression Control
   - Principal Component (PC) Adjustment
   - Estimating kinship
   - Mixed Regression Models

$$\mathbb{E}[y_i|\delta_i] = \alpha + x_i\beta + \delta_i$$

- y_i is the trait of interest for individual i
- x_i genotype of individual i at loci of interest
- δ_i is the polygenetic contribution of all other loci ("small, additive, genetic effects distributed across the genome").

$$\mathbb{E}[y_i|\delta_i] = \alpha + x_i\beta + \delta_i$$

- y_i is the trait of interest for individual i
- x_i genotype of individual i at loci of interest
- δ_i is the polygenetic contribution of all other loci ("small, additive, genetic effects distributed across the genome").

$$\begin{pmatrix} \delta_1 \\ \vdots \\ \delta_n \end{pmatrix} =: \delta \sim \mathcal{N}\left(\vec{0}, 2\sigma^2 h^2 K\right)$$

$$\mathbb{E}[y_i|\delta_i] = \alpha + x_i\beta + \delta_i$$

y_i is the trait of interest for individual i

x_i genotype of individual i at loci of interest

δ_i is the polygenetic contribution of all other loci ("small, additive, genetic effects distributed across the genome").

$$\begin{pmatrix} \delta_1 \\ \vdots \\ \delta_n \end{pmatrix} =: \delta \sim \mathcal{N}\left(\vec{0}, 2\sigma^2 h^2 K\right)$$

K is the kinship matrix

h^2 is the *narrow sense heritability* of the trait (proportion of variation due to additive polygenetic effects)

$$\mathbb{E}[y_i|\delta_i] = \alpha + x_i\beta + \delta_i$$

y_i is the trait of interest for individual i

x_i genotype of individual i at loci of interest

δ_i is the polygenetic contribution of all other loci (“small, additive, genetic effects distributed across the genome”).

$$\begin{pmatrix} \delta_1 \\ \vdots \\ \delta_n \end{pmatrix} =: \delta \sim \mathcal{N}\left(\vec{0}, 2\sigma^2 h^2 K\right)$$

K is the kinship matrix

h^2 is the *narrow sense heritability* of the trait (proportion of variation due to additive polygenetic effects)

$$y_i - (\alpha + x_i\beta + \delta_i) \sim \mathcal{N}\left(\vec{0}, \sigma^2(1 - h^2)I\right)$$

Software

EMMA allows fast likelihood-ratio tests with linear mixed models

H.M. Kang *et al.* (2008) Efficient control of population structure in model organism association mapping.
Genetics **178**, 1709–1723

GenABEL contains the command GRAMMAR, which uses an even faster approximative method and may thus have reduced power.

Y.S. Aulchenko,D.-J. de Koning, C. Haley (2007)
Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis
Genetics **177**, 577–585

"[. . .] *the common apporoach of 'correcting for population structure' may be misguided*".

A. Platt, B.J. Vilhámsson, M. Nordborg (2010) Conditions under which genome-wide association studies will be positively misleading
Genetics **186(3)**, 1045–1052

"[. . .] *the common apporoach of 'correcting for population structure' may be misguided*".

A. Platt, B.J. Vilhámsson, M. Nordborg (2010) Conditions under which genome-wide association studies will be positively misleading
Genetics **186(3)**, 1045–1052

Accounting for population structure and kinship does not avoid false positives due to **pleiotropy**, **multiple causal factors** or **epistasis**.

"[...] *the common apporoach of 'correcting for population structure' may be misguided*".

A. Platt, B.J. Vilhámsson, M. Nordborg (2010) Conditions under which genome-wide association studies will be positively misleading
Genetics **186(3)**, 1045–1052

Accounting for population structure and kinship does not avoid false positives due to **pleiotropy**, **multiple causal factors** or **epistasis**. Suggest to rather correct for confounding effects in general.

"[. . .] *the common apporoach of 'correcting for population structure' may be misguided*".

A. Platt, B.J. Vilhámsson, M. Nordborg (2010) Conditions under which genome-wide association studies will be positively misleading
Genetics **186(3)**, 1045–1052

Accounting for population structure and kinship does not avoid false positives due to **pleiotropy**, **multiple causal factors** or **epistasis**.
Suggest to rather correct for confounding effects in general.
Among methods based on the idea that effects of K should be corrected, those are more robust that don't infer K from island model but estimate confounding effects K directly from data, e.g.

J. Yu *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.
Nat. Genet. **38**, 203–208