Multivariate Statistics in Ecology and Quantitative Genetics
Genome Wide Association Studies

Dirk Metzler & Noémie Becker

http://evol.bio.lmu.de/_statgen

Summer semester 2018

1. Intro to GWAS
2. Basic statistics
3. Confounding factors
   - A simple approach: Genomic Control (GC)
   - Structured Association (SA)
   - Regression Control
   - Principal Component (PC) Adjustment
   - Estimating kinship
4. GWAS in R
5. Additional topics


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


S. Besenbacher, T. Mailund, M.H. Schierup (2012) Association Mapping and Disease: Evolutionary Perspectives
in M. Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods, Volume 2*.

Contents

Aim of GWAS

Linkage and Association studies have the same aim: finding loci that are associated with a phenotype (commonly a human disease but can be any phenotype in any species). They differ by the study design and methods:

Aim of GWAS

Linkage and Association studies have the same aim: finding loci that are associated with a phenotype (commonly a human disease but can be any phenotype in any species). They differ by the study design and methods:

- Linkage seeks loci at which there is correlation between the phenotype and the pattern of transmission of DNA sequence over generations in aknown pedigree.

Aim of GWAS

Linkage and Association studies have the same aim: finding loci that are associated with a phenotype (commonly a human disease but can be any phenotype in any species). They differ by the study design and methods:

- Linkage seeks loci at which there is correlation between the phenotype and the pattern of transmission of DNA sequence over generations in aknown pedigree.
- Association looks for loci that show differences in allele frequency between unrelated cases and controls.

Aim of GWAS

Linkage and Association studies have the same aim: finding loci that are associated with a phenotype (commonly a human disease but can be any phenotype in any species). They differ by the study design and methods:

- Linkage seeks loci at which there is correlation between the phenotype and the pattern of transmission of DNA sequence over generations in aknown pedigree.
- Association looks for loci that show differences in allele frequency between unrelated cases and controls.

Setting of GWAS

Sample of individuals:

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

Setting of GWAS

Sample of individuals:

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

Setting of GWAS

Sample of individuals:

- Many SNPs spread 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

Type of data

GWAS are based on Genome-wide chip data that genotype up to several million SNPs in the human genome.

Type of data

GWAS are based on Genome-wide chip data that genotype up to several million SNPs in the human genome.

Of course, not all variants in the genome are genotyped but if LD is strong enough, we expect to have genotyped a variant close enough to the causal variant.

Type of data

GWAS are based on Genome-wide chip data that genotype up to several million SNPs in the human genome.

Of course, not all variants in the genome are genotyped but if LD is strong enough, we expect to have genotyped a variant close enough to the causal variant.

High LD is thus good to identify a causal region but too high LD will make fine-mapping of the identified region more difficult.

Importance of quality control

After genotyping quality control is performed on individuals and on SNPs: On individuals:

- Low DNA quality: remove if more than 2-3% of SNPs not called
- excess heterozygosity: could be a sign of contamination
- control if recorded sex matches with genotype on sexual chromosomes
- control ancestry on PCA

Importance of quality control

After genotyping quality control is performed on individuals and on SNPs: On individuals:

- Low DNA quality: remove if more than 2-3% of SNPs not called
- excess heterozygosity: could be a sign of contamination
- control if recorded sex matches with genotype on sexual chromosomes
- control ancestry on PCA

On SNPs:

- remove SNPs with more than 5% missing data
- remove SNPs with large deviations from Hardy-Weinberg: could be misscalls of heterozygotes

Contents

Contigency tables and χ^2

For each SNP one can count the number of alleles in cases and controls...

	Allele A	Allele B	
Case	$N_{Case,A}$	$N_{Case,B}$	N_{Cases}
Control	$N_{Control,A}$	$N_{Control,B}$	$N_{Controls}$
	N_A	N_B	N

Contigency tables and χ^2

For each SNP one can count the number of alleles in cases and controls...

	Allele A	Allele B	
Case	$N_{Case,A}$	$N_{Case,B}$	N_{Cases}
Control	$N_{Control,A}$	$N_{Control,B}$	$N_{Controls}$
	N_A	N_B	N

... and calculate the expected number in each category.

	Allele A	Allele B	
Case	$(N_{Cases} * N_A)/N$	$(N_{Cases} * N_B)/N$	N_{Cases}
Control	$(N_{Controls} * N_A)/N$	$(N_{Controls} * N_B)/N$	$N_{Controls}$
	N_A	N_B	N

Contigency tables and χ^2

We recognize a χ^2 setting and compute:
$\chi^2 = \sum((Observed - Expected)^2/Observed)$

Contigency tables and χ^2

We recognize a χ^2 setting and compute:
$\chi^2 = \sum((Observed - Expected)^2/Observed)$

Under the null hypothesis that there is no association between the SNP and the disease, the statistic should be χ^2 distributed with 1 degree of freedom.

Contigency tables and χ^2

We recognize a χ^2 setting and compute:
$\chi^2 = \sum((Observed - Expected)^2/Observed)$

Under the null hypothesis that there is no association between the SNP and the disease, the statistic should be χ^2 distributed with 1 degree of freedom.

In case of low MAF (minor allele frequency) or low number of individuals, the χ^2 approximation might not be good and one could use a Fisher exact test instead.
If one wants to allow for associations following a dominant or recessive pattern, one can count genotypes and perform a χ^2 with 2 degrees of freedom.

Odds Ratio and Relative Risk

To measure the effect size of the association one can use the Odds Ratio (OR) or the Relative Risk (RR).

Odds Ratio and Relative Risk

To measure the effect size of the association one can use the Odds Ratio (OR) or the Relative Risk (RR).

$$OR = \frac{N_{Case,A}/N_{Control,A}}{N_{Case,B}/N_{Control,B}}$$

Odds Ratio and Relative Risk

To measure the effect size of the association one can use the Odds Ratio (OR) or the Relative Risk (RR).

$$OR = \frac{N_{Case,A}/N_{Control,A}}{N_{Case,B}/N_{Control,B}}$$

$$RR = \frac{N_{Case,A}/N_A}{N_{Case,B}/N_B}$$

Multiple testing correction

You will perform as many tests as SNPs so you should correct for multiple testing.

Multiple testing correction

You will perform as many tests as SNPs so you should correct for multiple testing.

- Bonferroni correction counting the number of SNPs
- For whole genome in humans a commonly used Pvalue threshold is $5 * 10^{-8}$

Some free GWAS software packages

- PLINK

  `http://pngu.mgh.harvard.edu/~purcell/plink/`
- EMMAX

  `https://genome.sph.umich.edu/wiki/EMMAX`
- 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`

    1.7-6) CRAN package, but seems not to be supported anymore. I installed it via the archive downloaded from genabel.org (version 1.7-6)

Contents

Possible problems

Confounding factors are all differences between cases and controls unrelated to the phenotype.

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

Possible problems

Confounding factors are all differences between cases and controls unrelated to the phenotype.

- 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

Confounding factors are all differences between cases and controls unrelated to the phenotype.

- 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

Confounding factors are all differences between cases and controls unrelated to the phenotype.

- 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!
  - Cryptic relatedness

Possible problems

Confounding factors are all differences between cases and controls unrelated to the phenotype.

- 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!
  - Cryptic relatedness
- more than one causal factor

Possible problems

Confounding factors are all differences between cases and controls unrelated to the phenotype.

- 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!
  - Cryptic relatedness
- more than one causal factor
- ascertainment bias (e.g. cases are sampled from some clinic, controls somewhere else)

Different scenarios

Whether we have to compensate for relatedness in the data depends on where the individuals 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

Contents

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^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 and unrelated samples. Under the latter conditions, T^2/V is approximately χ^2_1-distributed.

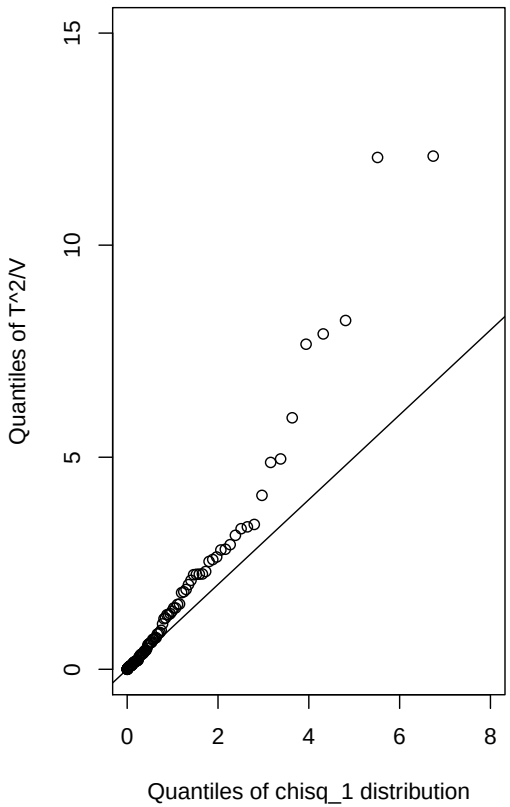
Genomic Control (GC): Fast and simple method to compensate for population structure or cryptic relatedness.

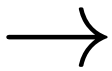
Main idea is to multiply test statistic with constant λ to make it fit χ^2_1 distribution.

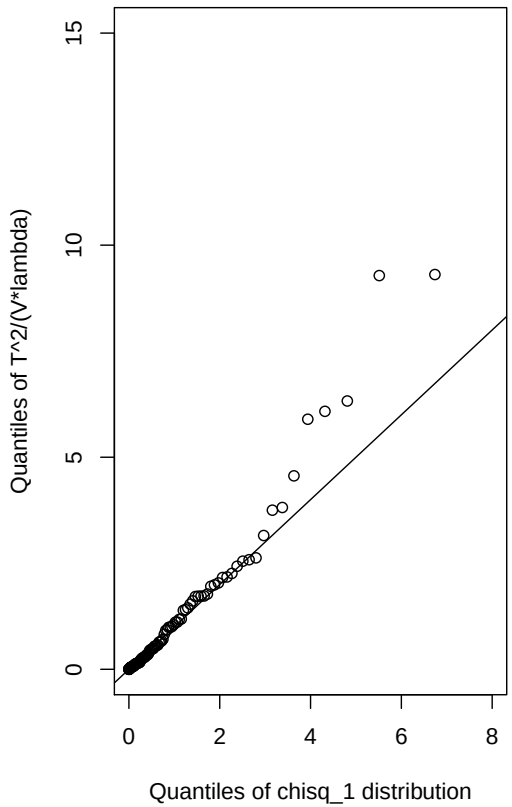
The test statistic is T^2/V, where T measures for a locus the difference in allele frequencies between cases and controls, and V approximates the variance of T for the case of neutrality and unrelated samples. Under the latter conditions, T^2/V is approximately χ^2_1-distributed.

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).





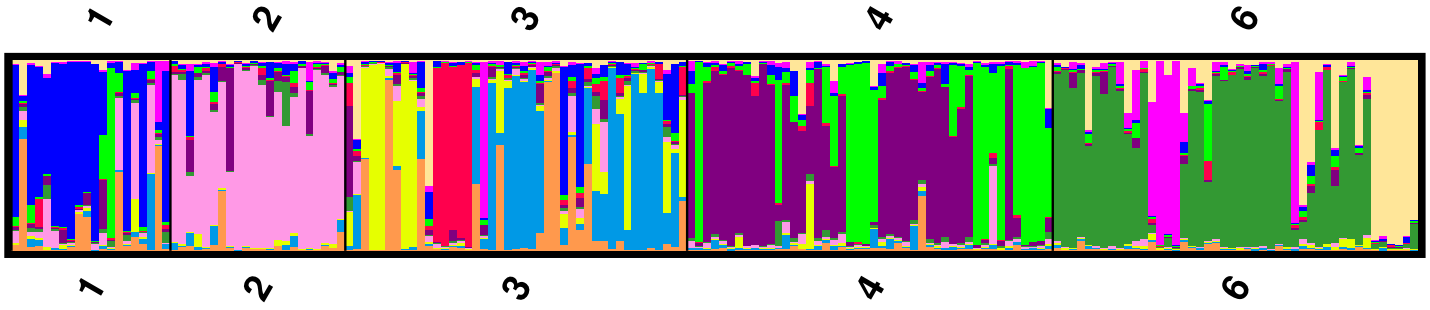


The outliers are candidate loci to be associated with the trait.

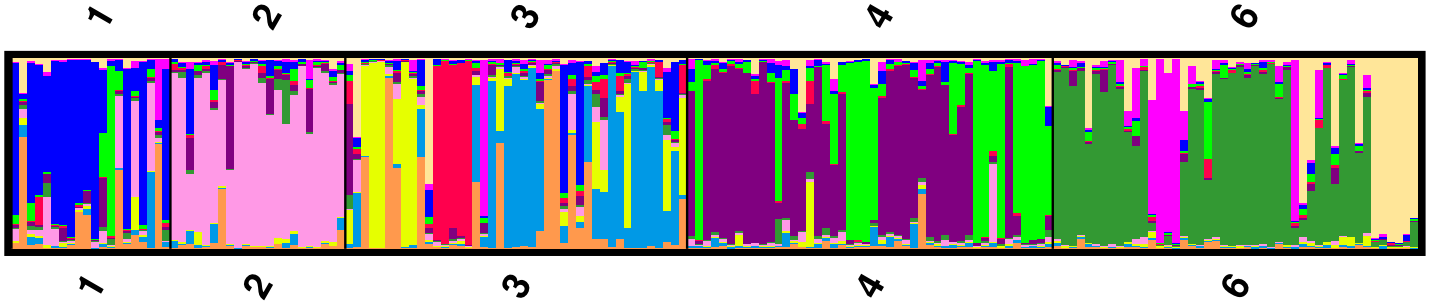
Contents

- 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

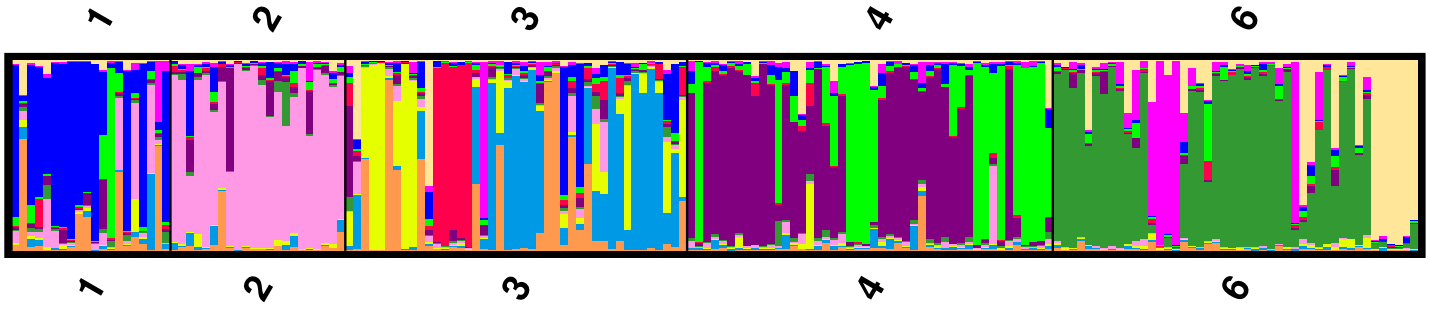


- Software: e.g. PLINK
- SA assumes that population consists of subpopulations (“islands”)
- Population structure can be estimated from $\sim$ 100 SNPs e.g. with software STRUCTURE, assuming that each island is in Hardy-Weinberg equilibrium



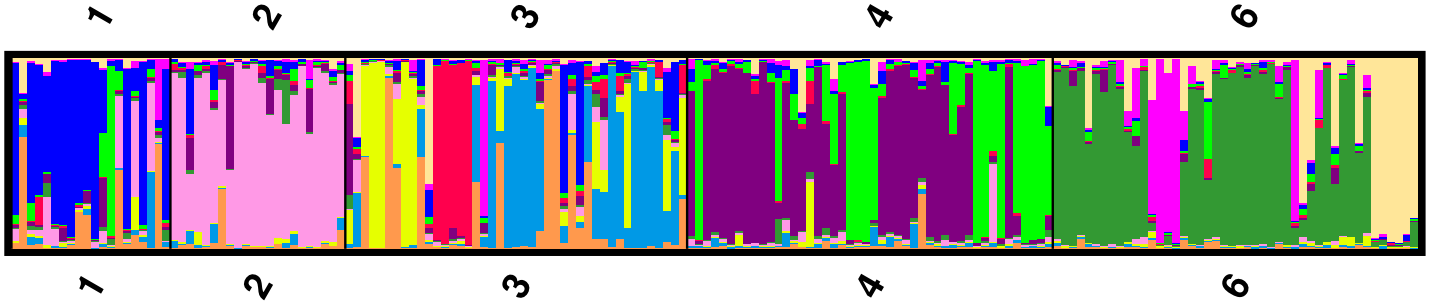
- with “admixture” option, individual genomes are admixed from different island

- 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



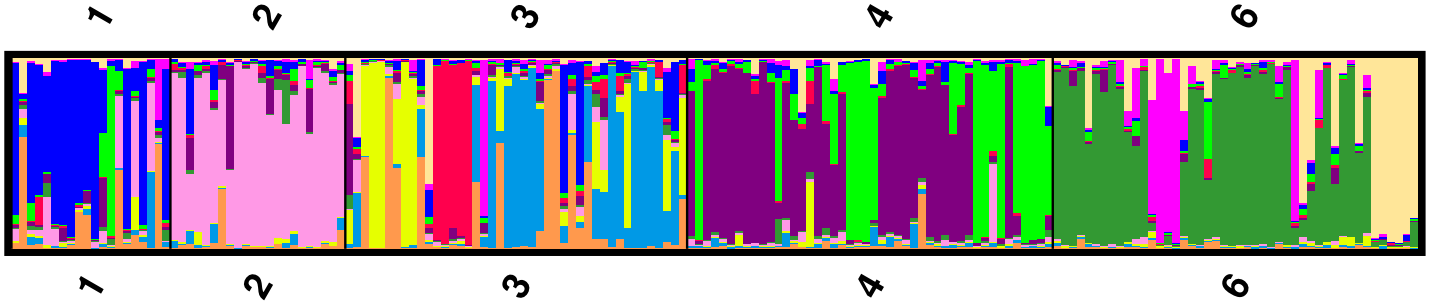
- with “admixture” option, individual genomes are admixed from different island
- stratified tests are applied, i.e. search for significant associations of trait and loci within the islands

- 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 “admixture” 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 “admixture” 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

Contents

- 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

Contents

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

Contents

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).

Contents

Contents

Some additional topics

- Knowing patterns of LD in the population of interest is important
- Imputation can be used to increase the number of loci covered
- GWAS need to be replicated to be considered true