General aspects of genome-wide association studies

Abstract number 20201
Session 04
Correctly reporting statistical genetics results in the genomic era

Pekka Uimari

University of Helsinki

Quality check of the SNPs

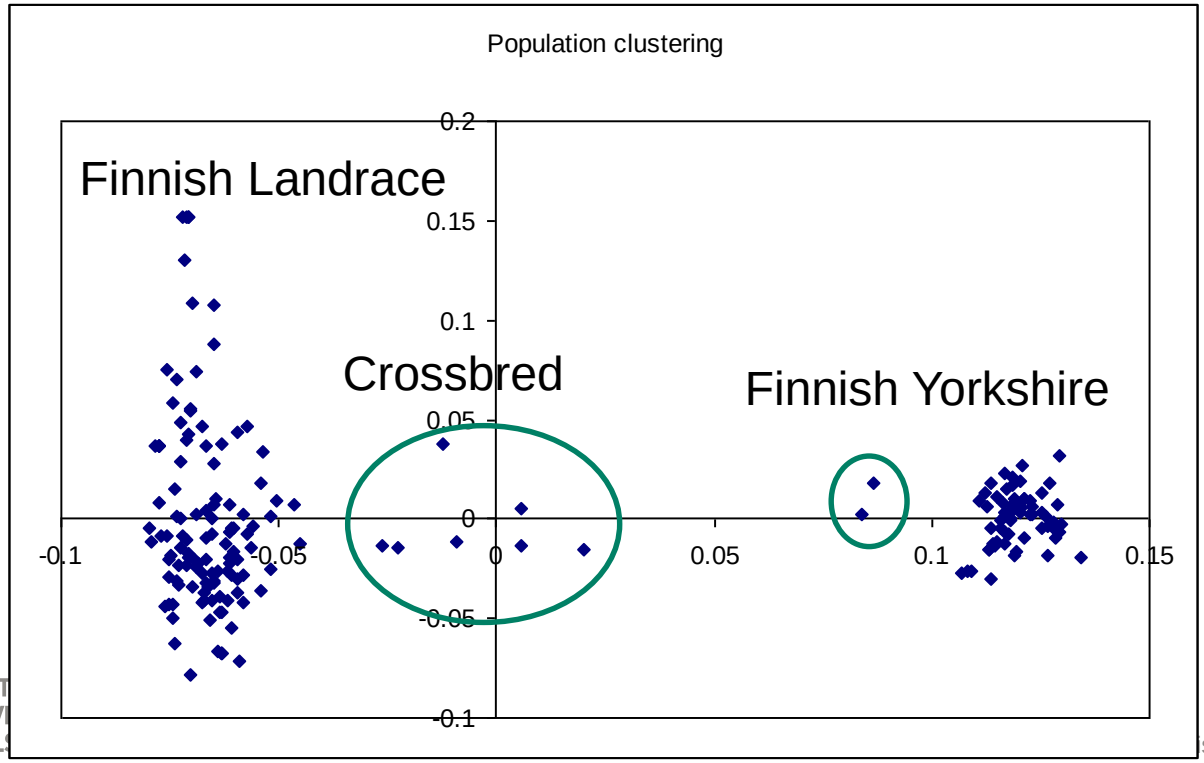
- Call rate
  - Low call rate indicates genotyping problems
  - All SNPs with CR < 90% are usually excluded
- Minor allele frequency
  - Eliminates non-polymorphic SNPs
  - MAF limit of 1% or 5% are commonly used
  - Number of animals in each genotyping class
- Hardy-Weinberg equilibrium
  - May indicate genotyping problems
  - Other causes are selection, recent mutation, random drift, small population size etc.
  - Some variation in practice exist e.g. P- value limit from 0.05 to 10^{-6}
- CR, MAF and HWE of the best SNPs (the smallest P-value in the GWA) are checked with more detail
- Also Illumina quality control parameter can be used
- Illumina Beadstudio allele calling scatter plots

Quality check of the samples

- Call rate
  - Commonly used limit for call rate is 90%
- Duplication
  - IBS/IBD is appr. 100%
- Know relatedness, parentage test
  - IBS/IBD estimation
    - Parent-offspring: proportion of SNPs with IBD=1 should be close to 100% depending on the level of inbreeding
    - Full-sibs expectation is IBD=0 25% IBD=1 50% IBD=2 25%
  - Population test

Population structure

- Sample structure can be studies using e.g. PLINK multi-dimensional scaling, Eigenstrat by Patterson, PLOS 2006 or other methods and software

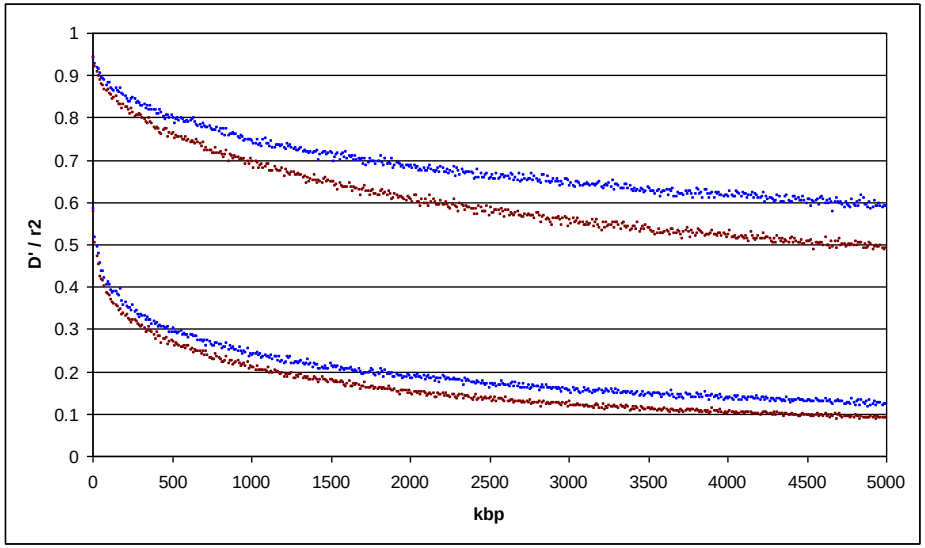


Imputation of genotypes

- Estimates missing genotypes, the most probably is usually used as an imputed genotype
- Increases power of association analysis
- Is done simultaneously together with haplotyping
- In animal genetics the most commonly used software seems to be Beagle (Browning and Browning, 2009)
  - Faster than for example older fastPHASE
  - Imputation error is usually very low appr. 2-3%

Choice of Animals

- Purebred animals commonly used
  - Genetic homogeneity
  - Usually high LD
- For populations with high LD less markers are needed compared to populations with low LD
- Synthetic breeds / breed crosses
  - Allele and locus heterogeneity



Number of Animals

- In human genetic studies the number of studied / genotyped individuals runs from 100 → 1000 → 10000 or more
- Effects the power of the study
  - Power of the study design can be estimated before hand but usually several assumptions about the mode of inheritance must be made, thus these estimates are seldom for any use
  - For simple Mendelian traits 10 cases and 10 controls can be enough
  - More realistic picture about the power can be achieved comparing the results of GWAS of similar or bigger sample size
- Remember that even with small sample sizes you can achieve P-values < 10^{-7}, some of those are false positives

Choice of Phenotypes

- The most common “phenotype” is based on estimated breeding values of males used in AI (AI-bulls, AI-boars)
- Possible only for traits that are measured in a national recording scheme (or similar type of data recording from performance testing stations etc.)
- For other traits observations of the genotyped animals are used
- Usually EBV’s are deregressed prior to GWAS (Garrick et al. 2009)
  - Removes the effect of parents on EBVs
  - Deregression
  - Calculation of weight of observations for GWAS

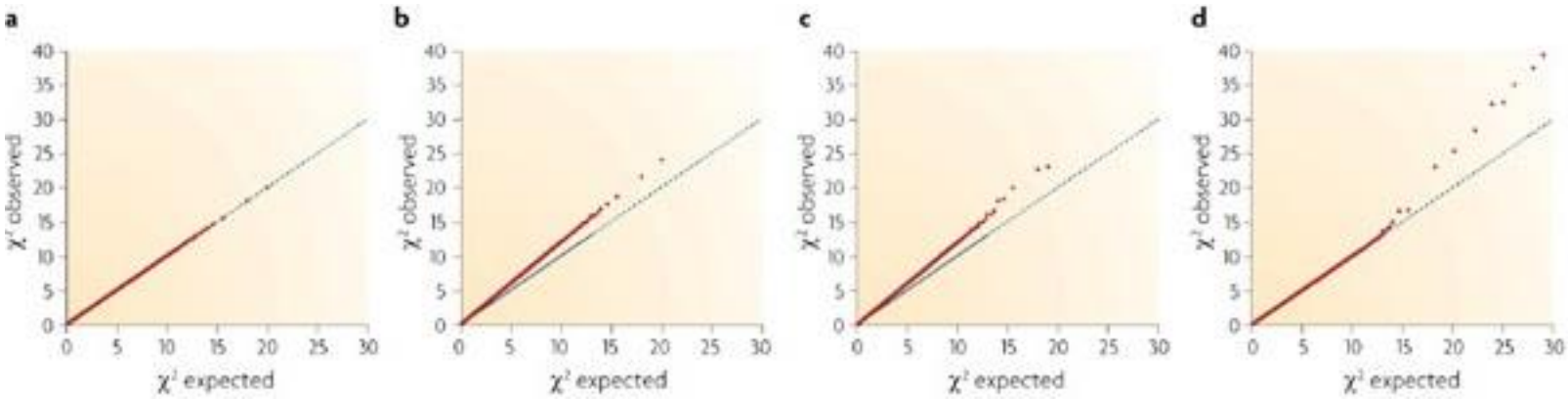
Statistical methods

- Single SNP model
  - The most commonly used model
  - Test for association one SNP at the time
- Multiple SNP model
  - Several or all SNPs analysed simultaneously
  - Oversaturation (number of markers >> number of individuals) needs to be handled
    - Selecting a subset of variables
    - Shrinking the estimates towards zero

Population stratification / relationship between the samples

- Genomic control (GC) is based on the idea a majority of the markers are not associated with the trait and the test statistics should follow the null hypothesis distribution
- Q-Q plot

Q-Q plot



a) No association no population stratification or relatedness

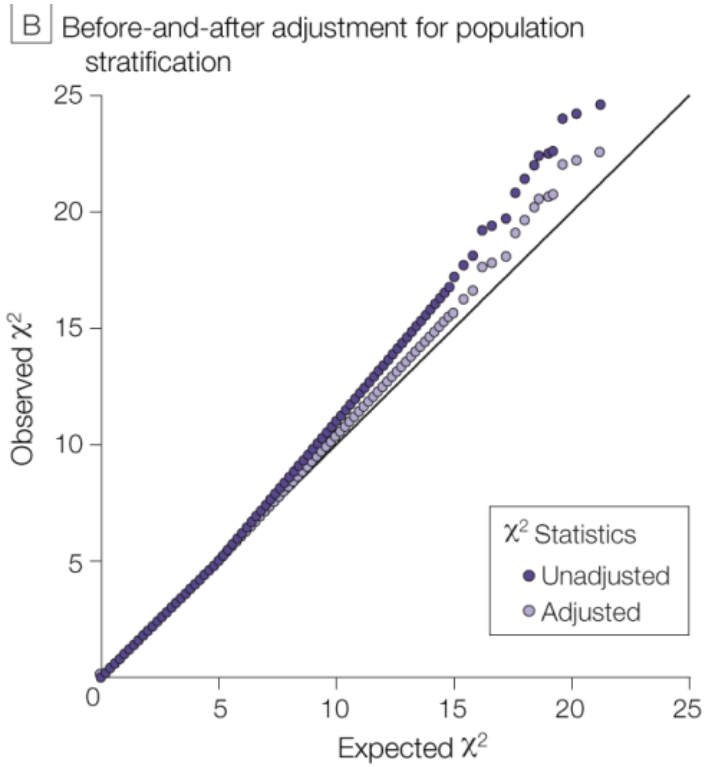
b) No association but indication of population stratification or relatedness

c) Evidence for association and population stratification or relatedness

d) Evidence for association but no population stratification or relatedness

McCarthy et al, Nature Reviews Genetic, 2008, 9:356-369

Genomic control



Pearson and Manolio, 2008, JAMA 299:1335-2150

- If there are indications (based on Q-Q plot) of population stratification or relatedness of samples test statistics can be adjusted using genomic control λ (Devlin & Roeder, 1999, Biometrics 55:997-1004)
- A simple estimate of λ is the mean of the obtained tests statistics or the median divided by 0.456 (0.456 is the expected median for chi-square distribution with df=1)

Population stratification / relationship between the samples

- The most commonly used way to control relatedness is to include the pedigree structure into a single marker mixed model
- Mixed linear model:

$$y_i = \mu + b*x_i + a_i + e_i,$$

- y_i is the deregressed EBV
- x_i is the number of minor alleles (0, 1, or 2) of the tested SNP
- b is the corresponding regression coefficient
- a_i is a random polygenic effect with $a_i \sim N(0, \mathbf{A}\sigma^2_a)$, where **A** is the additive relationship matrix and σ^2_a is the polygenic variance
- e_i is a random residual effect with $e_i \sim N(0, \mathbf{I}\sigma^2_e/w_i)$, where **I** is an identity matrix, σ^2_e is the residual variance, and w_i is the weight

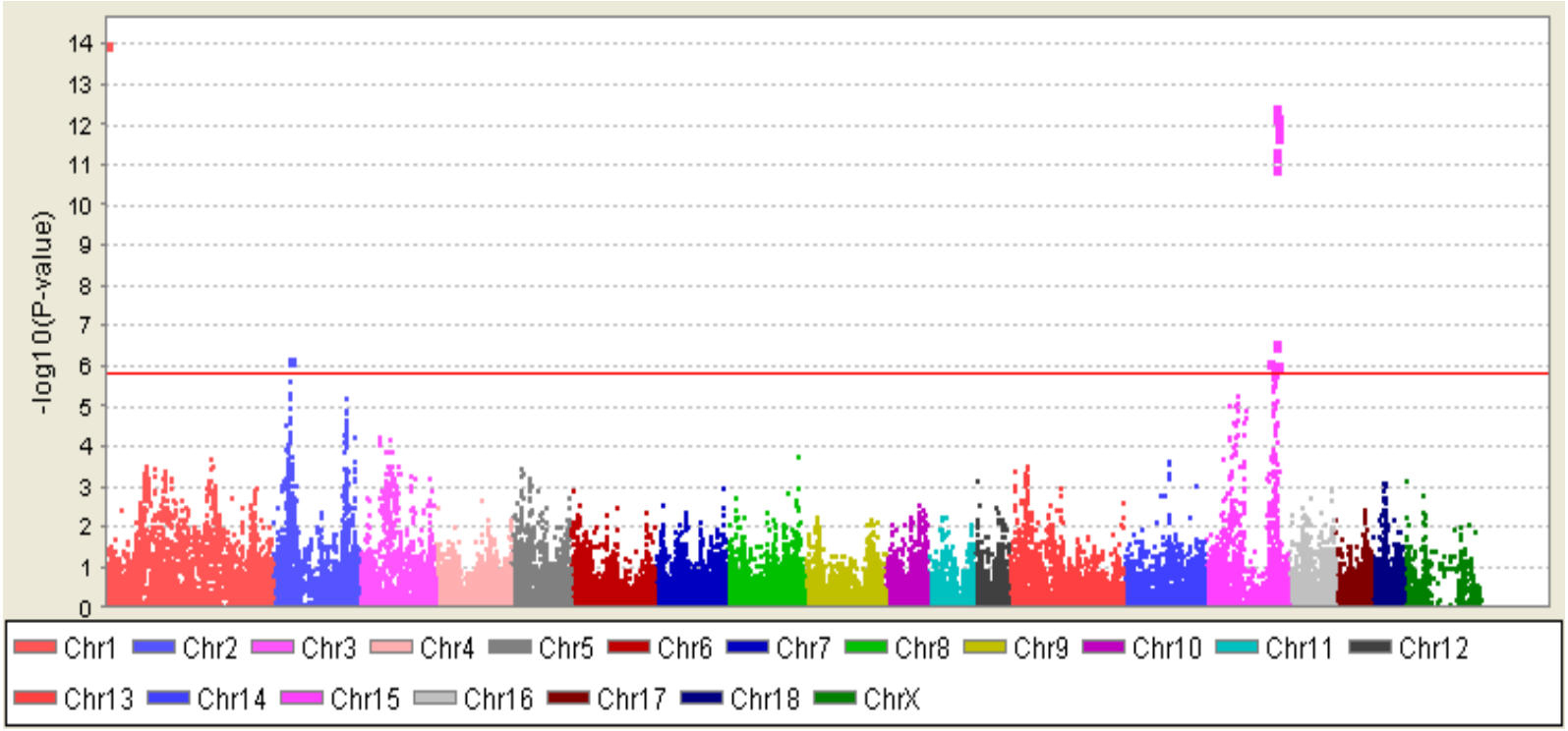
Relationship matrix

- Based either on pedigree (**A**) or genotypes (**G**)
- Genomic relationship matrix (**G**)
- Most commonly used the method presented by VanRaden (2008, J. Dairy Sci. 91, 4414-4423)
  - **G** = **ZZ**'/*k*, where $k=2\sum p_i(1-p_i)$
  - Or weight markers by reciprocals of their expected variance
  - **G** = **ZDZ**', where $D_{ii} = 1/(mk_i)$ where $k_i = 2p_i(1-p_i)$ and m=number of markers

P-values

- Test statistic of the SNP-effect from the mixed linear model:
  - t-test
  - F-test
  - Wald test
  - Squared t-test statistic has an exact F(1,n−1) –distribution
  - The Wald statistic can be used to test a simple hypothesis
    $H_0 : \theta = \theta_0$ on the entire parameter vector,
    $(\hat{\theta} - \theta)^T I(\hat{\theta})(\hat{\theta} - \theta)$ has x^2-distribution with 1 df
  - When $n \to \infty$ Wald-test with 1 df $\approx$ the square of the t -test statistic

Manhattan plot



Commonly used programs

- Variance-covariance estimation packages
  - DMU, ASREML etc
  - Fits mixed model equation and estimates variance components for each SNP → takes long time to analyze all markers
  - Use either A or G-matrix
- Methods and programs that take into account population and family structure (approximations)
  - GenABEL (GRAMMAR), Aulchenko et al, 2007, Genetics
  - EMMAX, Kang et al, 2010, Nature Genetics
  - Tassel, Zhang et al, 2010 Nature Genetics
- GEMMA, Zhou and Stephens 2012 Nature Genetics
  - Should be faster than DMU and ASREML

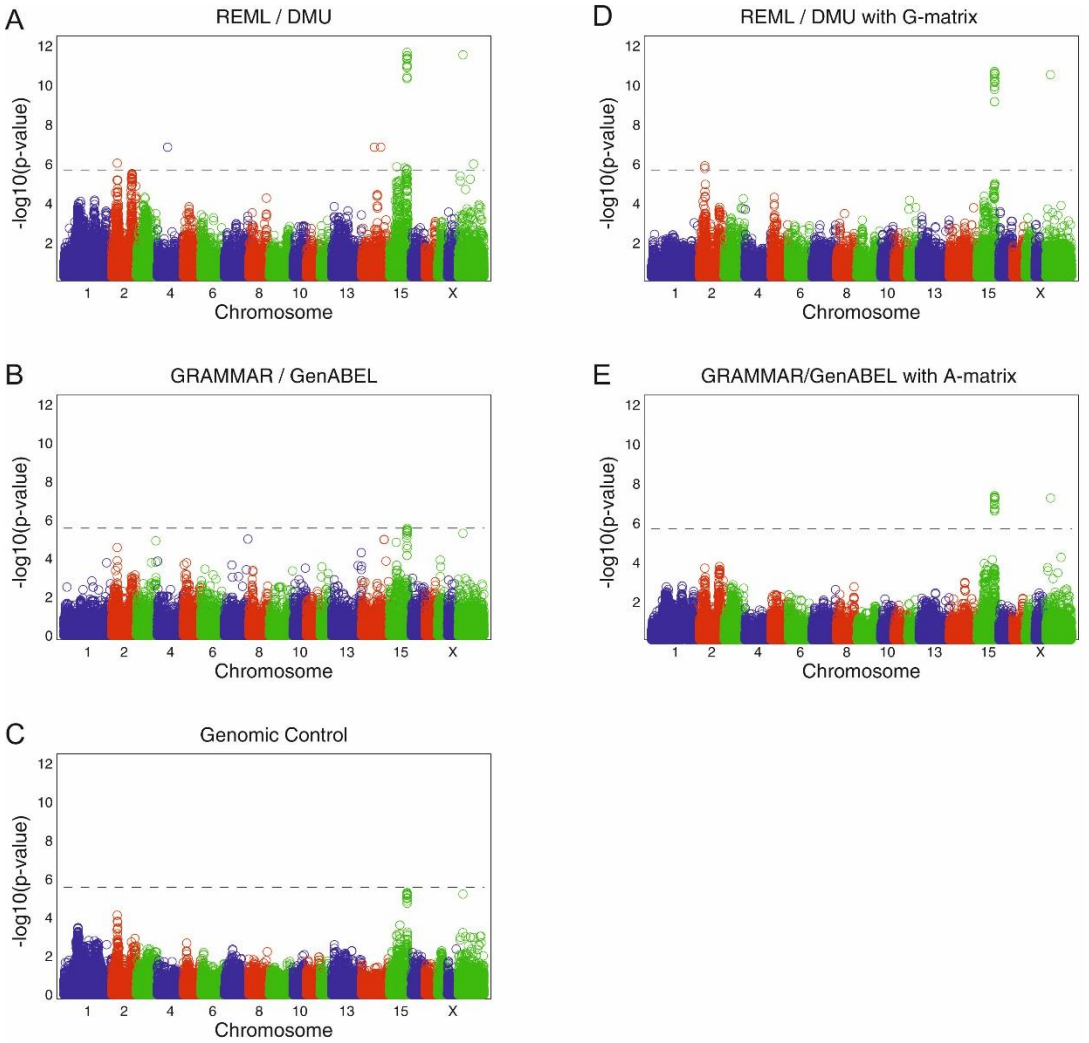
Multiple SNP model

- number of markers >> number of individuals
- Bayesian LASSO
  - $\boldsymbol{y} = \mu + \boldsymbol{Xb} + \boldsymbol{u} + \boldsymbol{e}$
  - $\boldsymbol{X}$ is the matrix on m most correlated marker genotypes
  - $\boldsymbol{b}$ is a vector of marker effects
  - $\boldsymbol{u}$ polygenic effect with G genomic relationship matrix computed from the rest of the markers
  - $b_j|\sigma_j^2 \sim N(0, \sigma_j^2)$
  - $\sigma_j^2|\lambda \sim Exp(\lambda^2/2)$
  - $\lambda \sim Gamma(\kappa, \xi)$
- Kärkkäinen & Sillanpää (2012, Genetics)

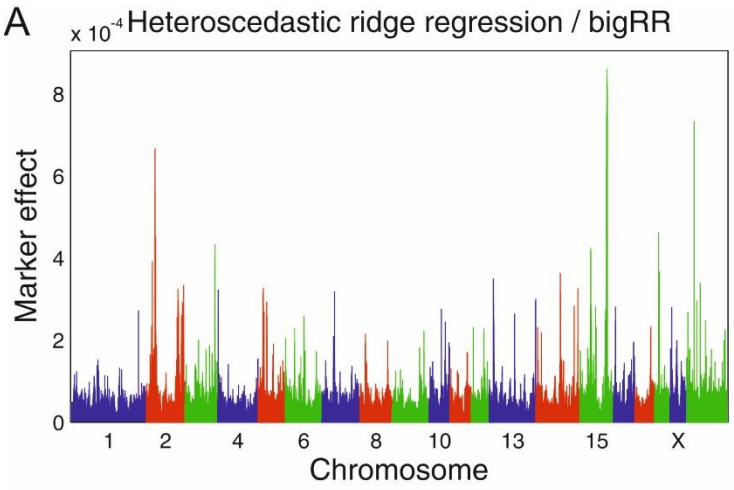
Multiple SNP model

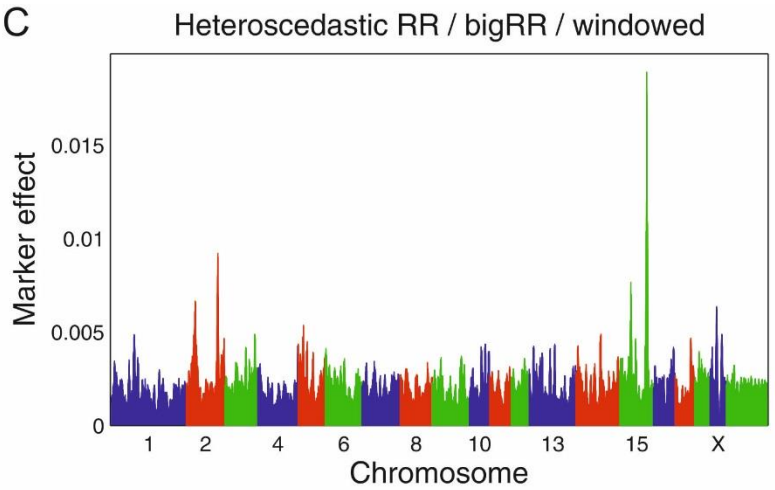
- number of markers >> number of individuals
- Heteroscedastic Ridge Regression
  - $\boldsymbol{y} = \mu + \boldsymbol{Xb} + \boldsymbol{e}$
  - $\boldsymbol{X}$ is the matrix on marker genotypes
  - $\boldsymbol{b}$ is a vector of random marker effects
  - First round: Shrinkage factor $\lambda \sim \sigma_e^2 / \sigma_b^2$
  - Second round: Shrinkage factor $\lambda \sim \sigma_e^2 / \sigma_{bj}^2$ where σ_{bj}^2 is calculated based on an estimate of a marker effect b_j from the first round
- bigRR R-package (Shen et al. 2013, Genetics)

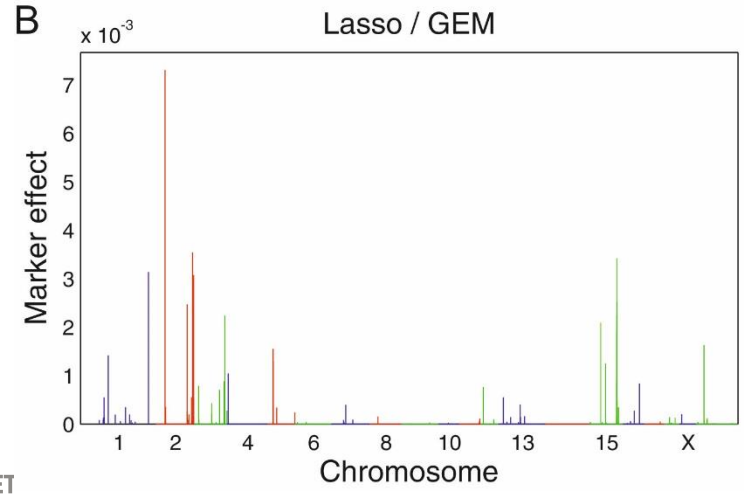
Comparison of the methods
Single SNP methods

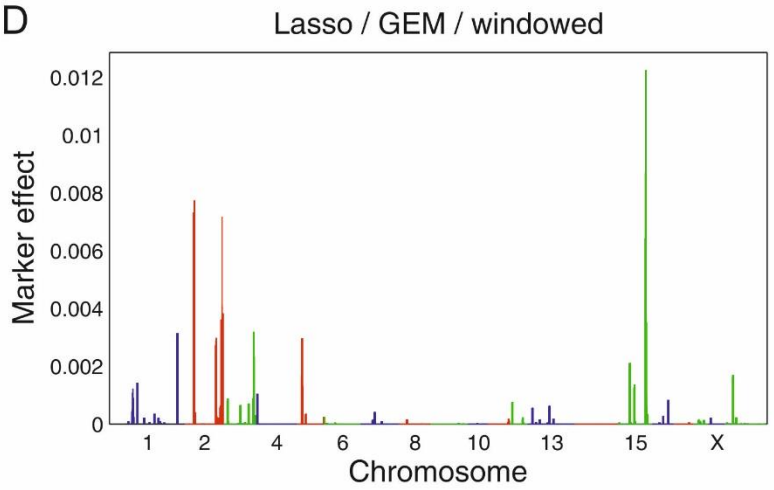


Comparison of the methods
Multiple SNP methods









Post GWAS

(see review by Wojcik et al. 2015 BMC Genetics)

- Aggregate markers into biologically relevant units, gene or pathway
- Increase power: combine multiple weak or moderate signals
- Allow for allelic or locus heterogeneity
- Gene-level analyses
  - Combines independent signals within a gene
  - Should take LD into account
  - E.g. VEGAS (Liu et al. AJHG 2010)
- Pathway-level (gene-set) analyses
  - Related collection of genes with similar biological function
  - Assess if strong associations cluster within a gene set compared to genes outside of the pathway (or gene set)