Genome-wide association studies (GWAS)

Purpose:

Detection of common genetic variants (minor allele frequency $> 1\%$) being associated with the disease of interest

Genotyping technologies:

SNP chips consisting of 500,000 – 4,000,000 SNPs

(Affymetrix Inc., Illumina Inc.)

Analysis of GWAS: Quality control

Purpose:

Identification (and exclusion) of low-quality SNPs and low-quality DNA

Criteria:

- minor allele frequency (MAF):

  discard SNPs with MAF less than 1% in cases or controls

- SNP call rate:

  The call rate (CR) for a given SNP is the fraction of individuals whose genotypes are called for this SNP.

  SNPs with a CR$< 95\%$ (or $< 97\%$ or $< 98\%$ ...) in cases or controls are excluded

Analysis of GWAS: Quality control

Criteria:

- call rate per individual:

  The call rate (CR) per individual is the fraction of SNPs whose genotypes are called for this individual.

  Individuals with a CR$< 98\%$ (or $< 97\%$ or $< 95\%$ ...) are excluded

- Hardy-Weinberg equilibrium (HWE):

  discard SNPs whose distribution of genotypes deviates from HWE in cases or controls

  (but remember exercise on slide P/10: an associated SNP will show deviations from HWE in cases unless the effect is multiplicative!)

Analysis of GWAS: Relationship inference

Purpose:

Identification of differences between the assumed and the real pedigree structure

(Examples: non-paternity in family studies, close relationship between individuals in case-control studies, . . .)

Method:

IBS, kinship coefficients

Software:

KING (**k**inship-based **in**ference for **G**WAS)

(http://people.virginia.edu/~wc9c/KING/)

Analysis of GWAS:

Correcting for population stratification (PS)

Multidimensional scaling (MDS):

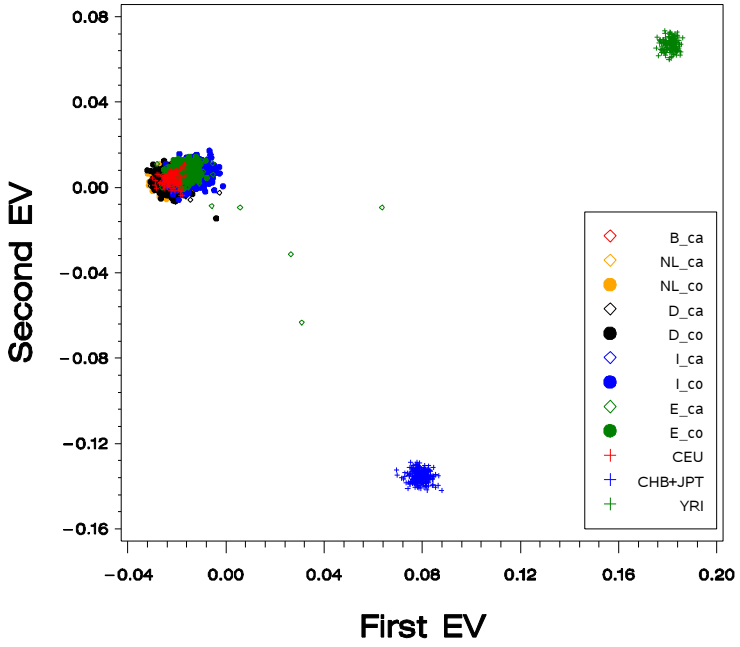
MDS projects the points in a high dimensional space to a lower dimensional space, but preserve the distances between points as much as possible. The coordinates of the points in the lower dimensional space can be included as additional covariates (representing the ancestry of an individual) in the logistic regression analysis.

Software:

PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/)

Analysis of GWAS: Correcting for PS

Example:



Representation of the study sample and HapMap individuals for the first two dimensions obtained by multidimensional scaling analysis of a matrix of pairwise IBS (identical by state) distance values between individuals

Analysis of GWAS: Correcting for PS

- Principal components analysis (PCA):

  Principal components can be included as additional covariates in the logistic regression analysis.

  Software:

  EIGENSTRAT/EIGENSOFT

  (http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html)

- Genomic control (Devlin and Roeder (1999), Biometrics 55: 997–1004): will be discussed later

Analysis of GWAS: Single marker analyis

- Armitage's trend test
- logistic regression (with principal components (PCA) or eigenvectors (MDS) as covariates)
- TDT

Each SNP is tested at a prespecified significance level

(typically: $\alpha = 5 \cdot 10^{-8}$)

Analysis of GWAS: Correcting for PS

Genomic control:

All test statistics listed on GWAS/8 follow a χ^2_1 distribution under the null hypothesis. The 0.5-quantile (i.e., median) of a χ^2_1 distribution is 0.45494.

Idea: Consider the ratio λ (genomic inflation factor) of the (empirical) median of the test statistics $T_1, \ldots, T_m$ for all m SNPs and 0.45494. Adjust T_i by dividing by λ, i.e., $T_i^{\text{adj}} = T_i/\lambda$, $(i = 1, \ldots, m)$.

Analysis of GWAS: Imputation

Statistical inference of genotypes at unobserved marker loci by using data from a reference panel (e. g. HapMap data)

(Marchini and Howie (2010), Nat Rev Genet 11: 499–511)

Software:

- IMPUTE2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html)
- BEAGLE (http://faculty.washington.edu/browning/beagle/beagle.html)
- MACH (http://csg.sph.umich.edu/abecasis/mach/tour/imputation.html)
- SNP2HLA (http://www.broadinstitute.org/mpg/snp2hla/)

Analysis of GWAS: Haplotype analysis

Methods:

see slides PC/6 and PC/9

Software:

- FBAT (http://www.biostat.harvard.edu/fbat/fbat.htm)
- UNPHASED (https://sites.google.com/site/fdudbridge/software/)
- FAMHAP (http://famhap.meb.uni-bonn.de/)

SNP-SNP interaction analysis

500,000 SNPs $\Rightarrow 1.25 \cdot 10^{11}$ combinations of two SNPs

Challenges:

- efficient storage of data
- calculation of the 3×3 contingency tables $(D_{i,j})_{i,j=0,1,2}$, $(C_{i,j})_{i,j=0,1,2}$
- calculation of the likelihood ratio test statistic for H_0^I

SNP-SNP interaction analysis

Wan et al. (2010), Am J Hum Genet 87: 325–340

Storage of data:

	Case1	Case2	Case3	Case4	...
SNP1	TC	TT	CC	TT	
SNP2	AG	AG	00	GG	

	Case1	Case2	Case3	Case4	...
$SNP1_1$	0	1	0	1	
$SNP1_2$	1	0	0	0	
$SNP1_3$	0	0	1	0	
$SNP2_1$	0	0	0	1	
$SNP2_2$	1	1	0	0	
$SNP2_3$	0	0	0	0	...

Three bits are required to store a single genotype, but ...

SNP-SNP interaction analysis

Calculation of $(D_{i,j})_{i,j=0,1,2}$, $(C_{i,j})_{i,j=0,1,2}$:

$$D_{ij} = \mathrm{SNP1}_i \cdot \mathrm{SNP2}_j^T$$

- bitwise “AND” operation
- counting of “1” bits (Hamming weight)

SNP-SNP interaction analysis

Calculation of the likelihood ratio test statistic for H_0^I:

$T = 2\left(\ln L(\widehat{\beta}) - \ln L(\tilde{\beta})\right)$ with $\widehat{\beta}$ denoting the unrestricted MLE of β and with $\tilde{\beta}$ denoting the MLE of β under the null hypothesis

- closed analytical solution of $\widehat{\beta}$ is available ($\Rightarrow$ calculation of $\ln L(\widehat{\beta})$ is inexpensive)
- calculation of $\tilde{\beta}$ requires iterations ($\Rightarrow$ expensive)

  Idea: use a "guess" $\beta^\star$ of $\tilde{\beta}$ and calculate $T^\star = 2\left(\ln L(\widehat{\beta}) - \ln L(\beta^\star)\right)$. Obviously, $L(\beta^\star) < L(\tilde{\beta})$ and, therefore, $T < T^\star$. Only in case that this "screening step" results in a sufficiently large $T^\star$, calculate the expensive $\tilde{\beta}$ and T.

SNP-SNP interaction analysis

Example:

- 499,593 SNPs
- 124,796,333,028 combinations of two SNPs
- 399 Cases
- 1318 Controls

total running time: $<$100 h

Natural parallelisation:

All combinations of SNPs on chromosome i and SNPs of chromosome j $(i \leq j)$ are considered in one job ($\Rightarrow$ 253 jobs)

Genome-wide association studies: Replication

Confirmation of the findings of the initial study by using independent samples.

It is not straightforward to calculate the power of the replication study because estimates of effect size obtained from significant association samples are biased upwards ("winner's curse", Zöllner and Pritchard (2007), Am J Hum Genet 80: 605–615).

Genome-wide association studies: Successes