Statistics for genomic data science in health

Cyril Dalmasso
cyril.dalmasso@univ-evry.fr

Université d'Evry Val d'Essonne

2022-2023

1. Introduction to GWAS/WGS/WES
2. Data structure
3. Single-marker analyses
4. Multiple testing
5. Multi-marker analyses

Statistics and genetics/genomics

Historical perspective

- Mendel (1866) and Morgan (1915) $\rightarrow$ genetic heritability concept
- 1953 : DNA structure resolved $\rightarrow$ Molecular genetics
- 1970s : Databases constitution $\rightarrow$ Bioinformatics
- 1990 - : Whole genome sequencing
- 2000 - : High throughput technologies $\rightarrow$ massive genomic data

Genomics

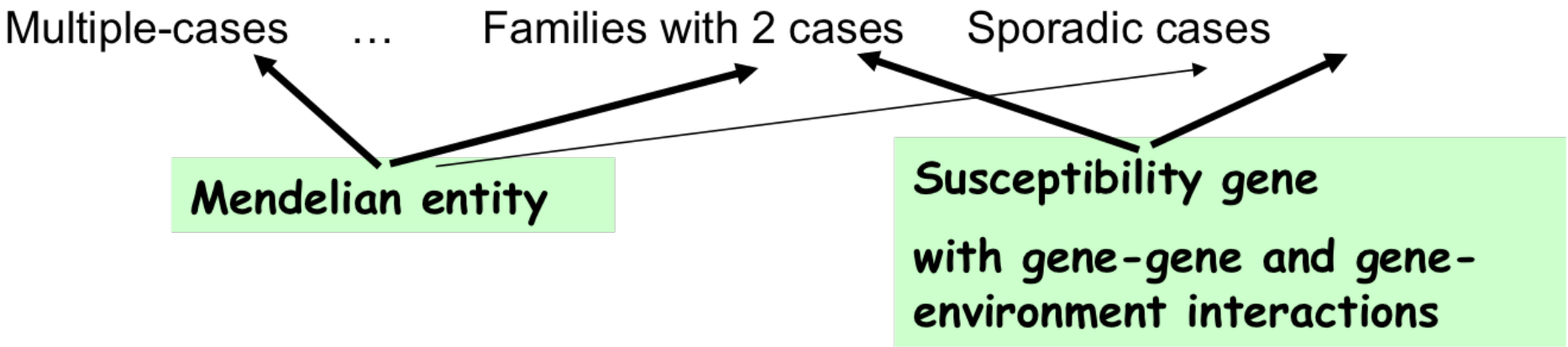
Genomics is the study of genomes

Genetic factors in a medical context

Monogenic diseases

- One causal gene (mendelian entity)
  - Rare mutations / allelic heterogeneity
  - High penetrance ($\mathbb{P}(phenotype|riskgenotype)$) $\Rightarrow$ multiple cases (familial aggregation)
- Environmental factors

Multifactorial diseases



V. Chaudru

Identification of causal genes and gene-environment interactions

- Is there familial aggregation? (epidemiological study)
- Is there a mendelian entity? (segregation analysis)
- In which genome regions can we find susceptibility genes ?
  $\rightarrow$ **linkage analyses** (family based)
  powerful in gene identification of mendelian diseases
- Which are the susceptibility genes?
  $\rightarrow$ **association studies** (population based)
  powerful in gene identification of complex diseases

Linkage analyses and association studies are based on **genetic markers**

Association studies

Objectives of association studies

- to localize regions containing a causal gene
- to test association with potential candidate genes
- to characterize such genes

Association studies

Candidate gene

Use of pre-specified genes

Fine mapping

Specific region (1-10Mb; 100 SNPs)

Genome wide association studies (GWAS)

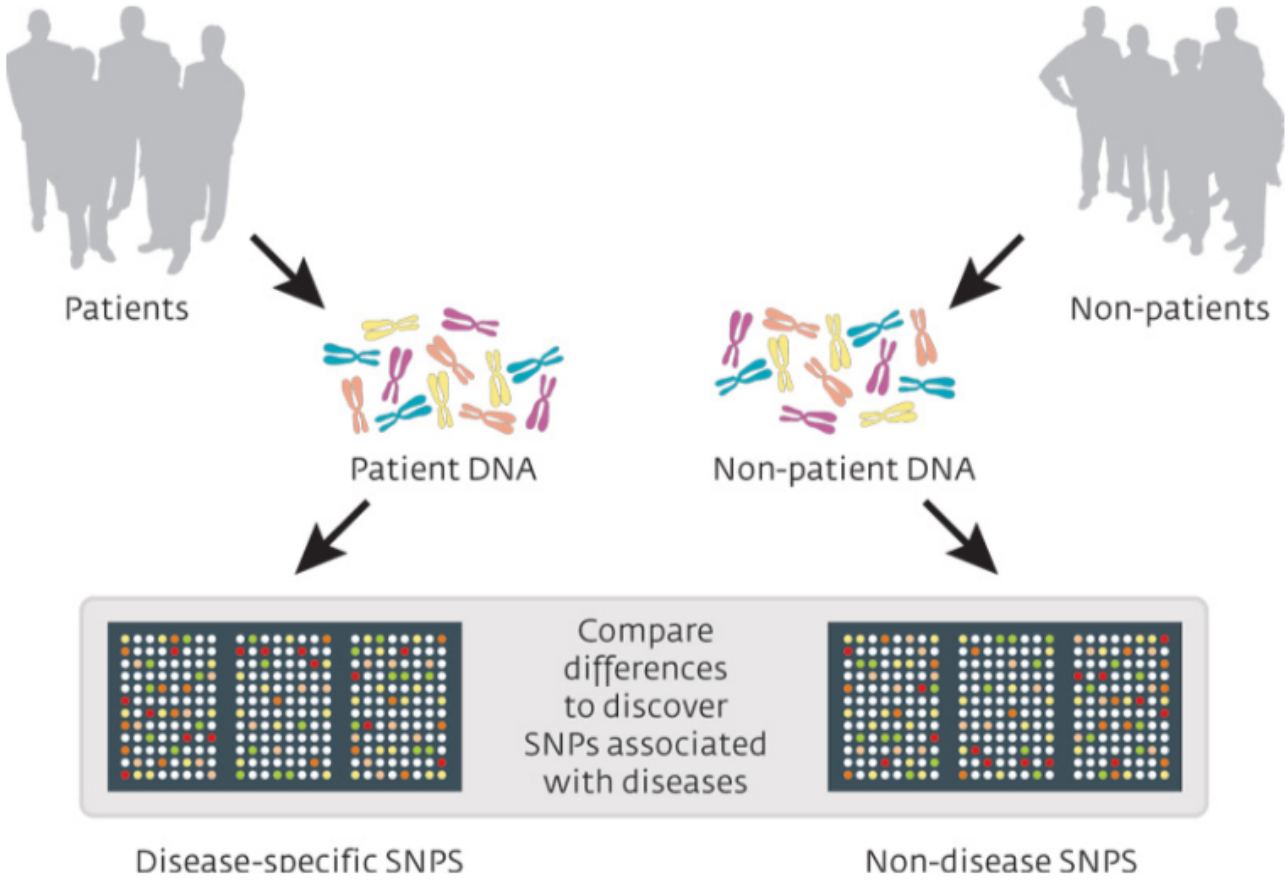
Use of genes all along the genome
Remark : Association with polymorphisms that are not themselves causal risk factors can be used to localize the trait gene

Association studies

Population based studies

Use of unrelated individuals rather than families

Case-control studies



© Pasieka, Science Photo Library

Genome wide association studies

Overall strategy

1. Calculate association statistics with the phenotype of interest
2. Derive p-values
3. Apply a Multiple Testing Procedure
4. Follow-up (report, meta-analysis, auxiliary analysis, ...)

Genetic markers

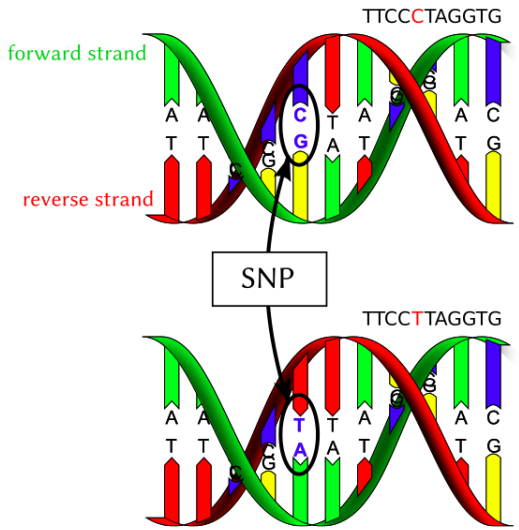
Definition

A genetic marker is a DNA sequence

- with a known location on a chromosome
- easily detectable
- can be described as a variation that can be observed

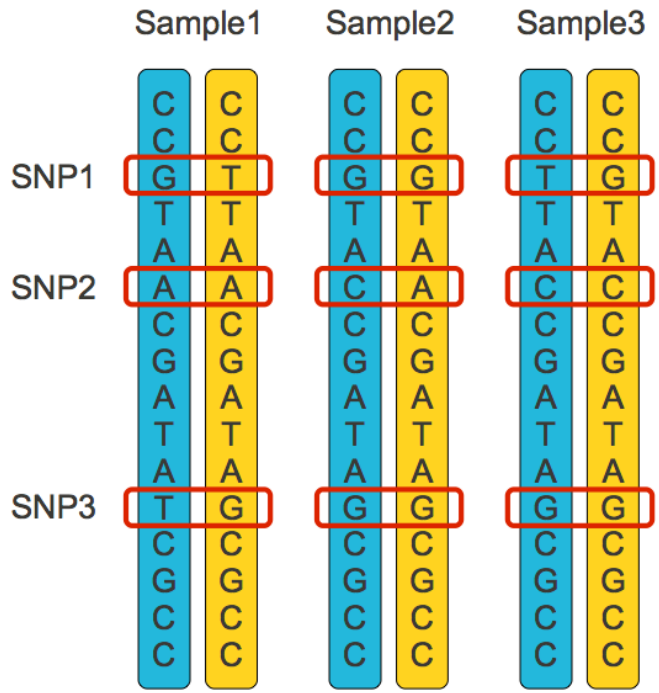
Single nucleotide polymorphism (SNP)

Single nucleotide polymorphisms (SNP) are the most common polymorphisms (approx. 10 millions known SNPs).



If reverse strand is chosen as reference, genotype for this SNP can be CC, CT or TT (GG, GA or AA on direct strand), often recoded in 0, 1 2 in genetic data files.

Single Nucleotide polymorphism (SNP)



Single Nucleotide polymorphism (SNP)

Some numbers

- Average distance between two SNPs: 600bp
- Total number of SNPs: 10 millions (among 3.2 billions base pairs)

Linkage disequilibrium

Definition

Linkage disequilibrium is the tendency for pairs of alleles at nearby loci to be associated with each other more than expected by chance

Linkage disequilibrium

LD measures

$Marker A \cdot^{Marker B}$	B	b	
A	p_{AB}	p_{Ab}	p_{A+}
a	p_{aB}	p_{ab}	p_{a+}
	p_{+B}	p_{+b}	

- $\mathcal{D} = p_{AB} - p_A p_B$
- $\mathcal{D}' = \frac{\mathcal{D}}{\mathcal{D}_{max}}$ where

$$\mathcal{D}_{max} = \begin{cases} min(p_A p_b; p_a p_B) & \text{if } \mathcal{D} > 0 \\ min(p_a p_b; p_A p_B) & \text{if } \mathcal{D} < 0 \end{cases}$$

- $R^2 = \frac{\mathcal{D}^2}{p_a p_A p_b p_B}$ (-> correlation coefficient)

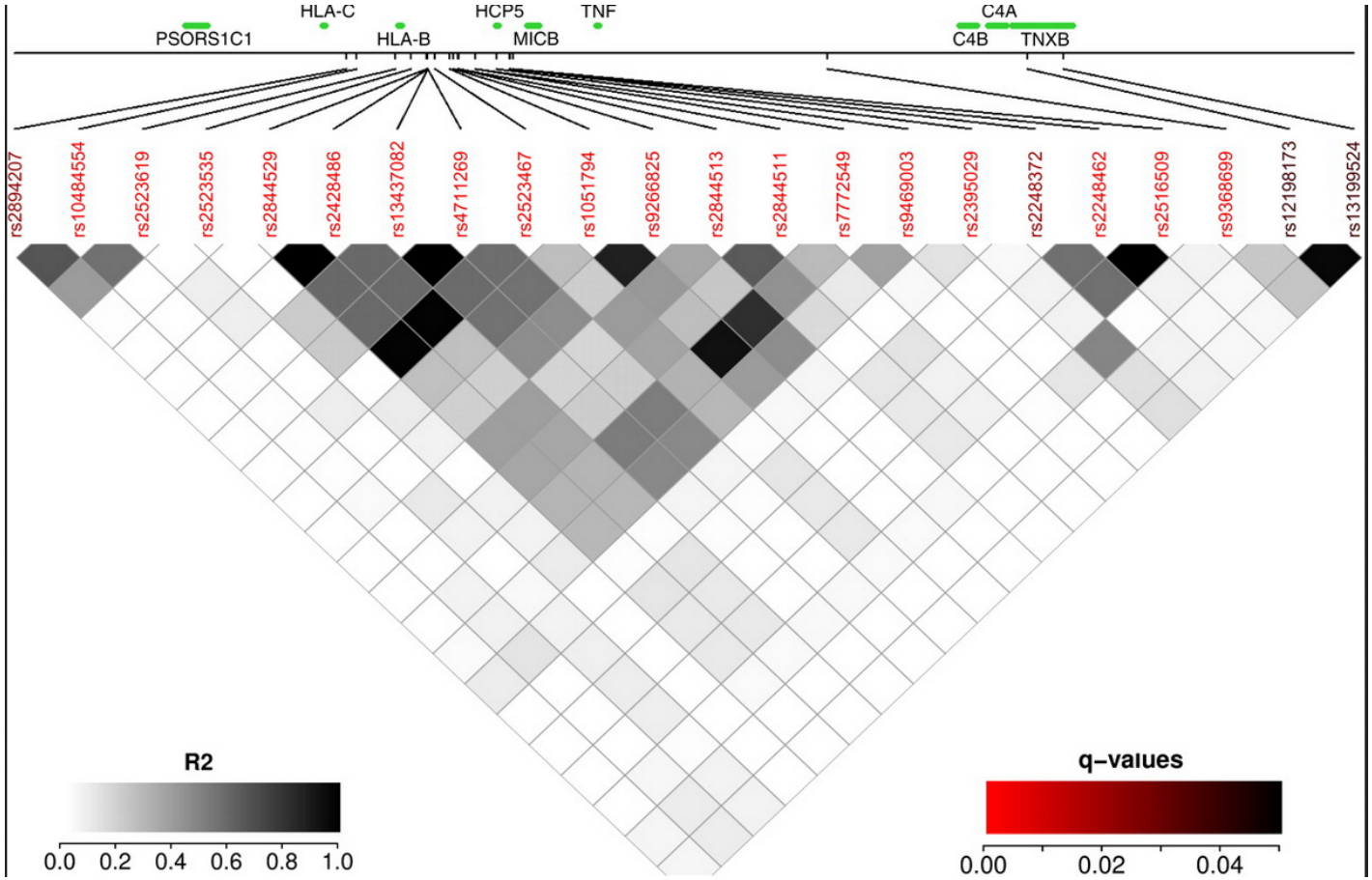
Linkage disequilibrium

Haplotype blocks

- Haplotype: set of SNPs that tend to occur together.
- Haplotype block: Islands of high linkage disequilibrium separated by regions of low linkge disequilibrium
- Recombination rates appear greater between blocks than within blocks
- Blocks exhibit low haplotypic diversity and most of the common haplotypes can be defined by a relatively small number of SNPs (3-5)

Linkage disequilibrium

Haplotype blocks



Guergnon J. et al, *J Infect Dis*, 2012

Genomic technologies

Technologies

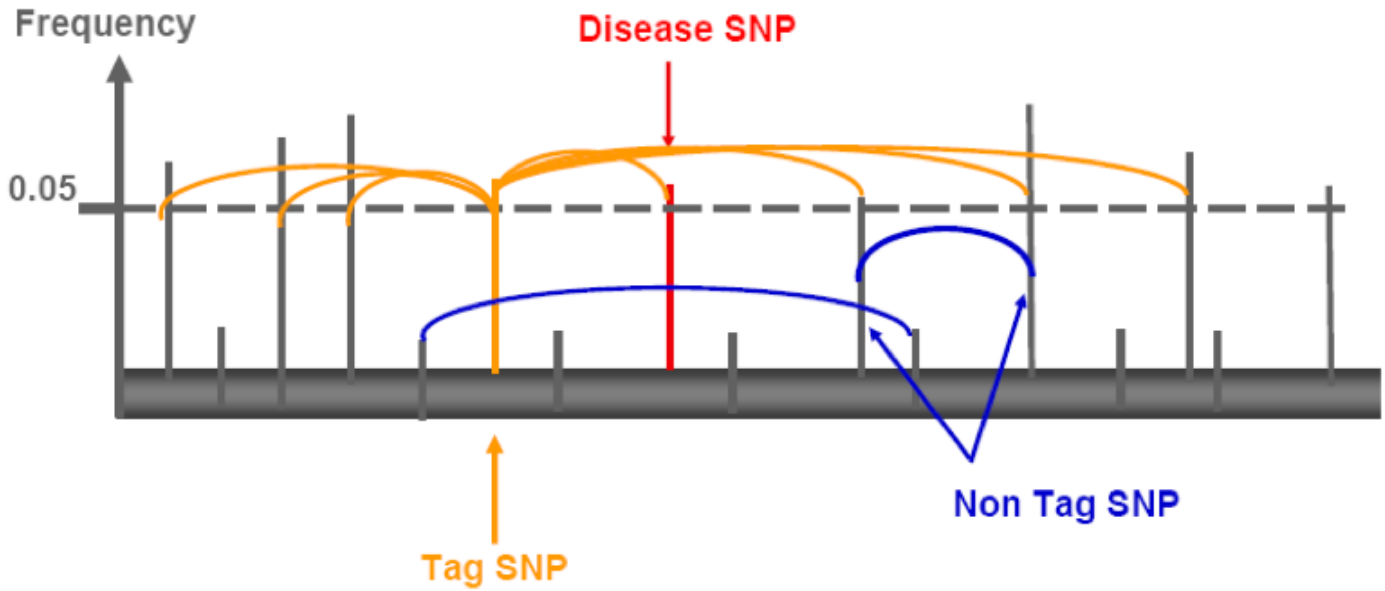
- Whole genome sequencing (WGS)
- Whole exome sequencing (WES)
- SNP genotyping (microarrays)

Microarrays

Key concept for GWAS

Exploiting the correlation structure in the genome to selectively genotype a reduced number of polymorphisms by providing a reasonable coverage of the genome.

TAG SNPs

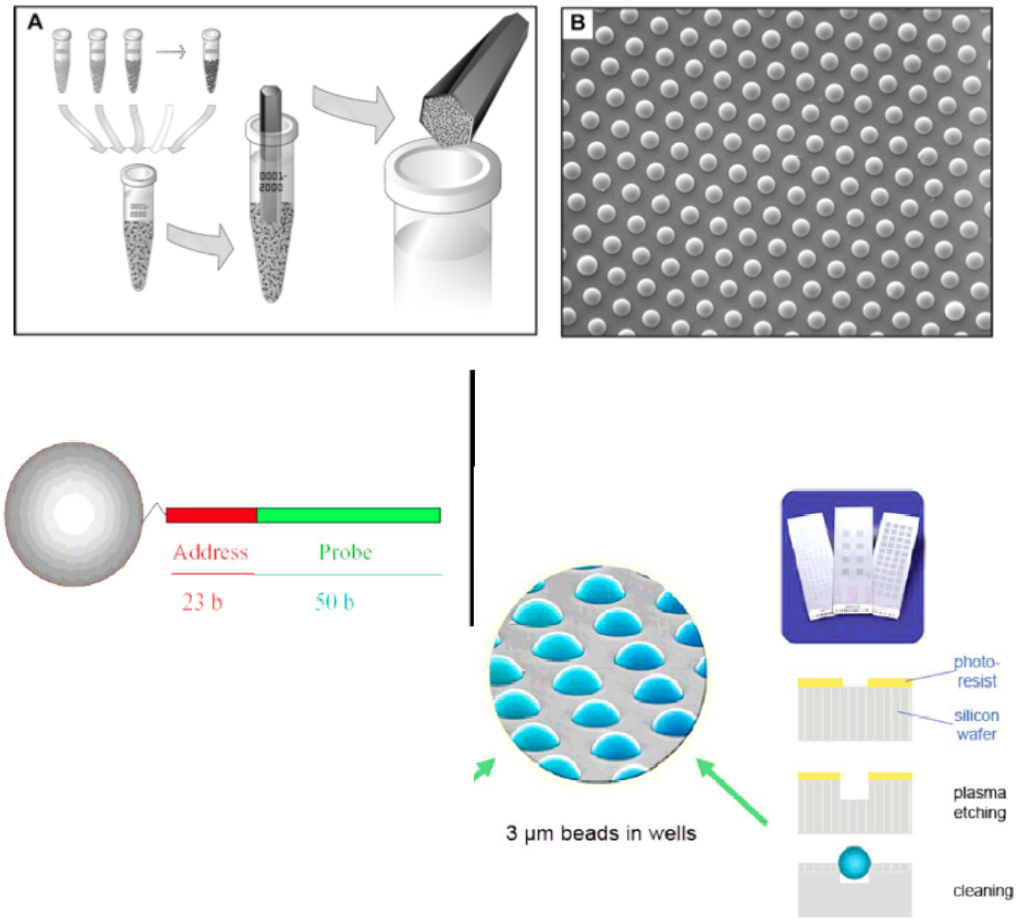


MAF> 0.05 - Common SNP

15

illumina

Illumina SNP arrays



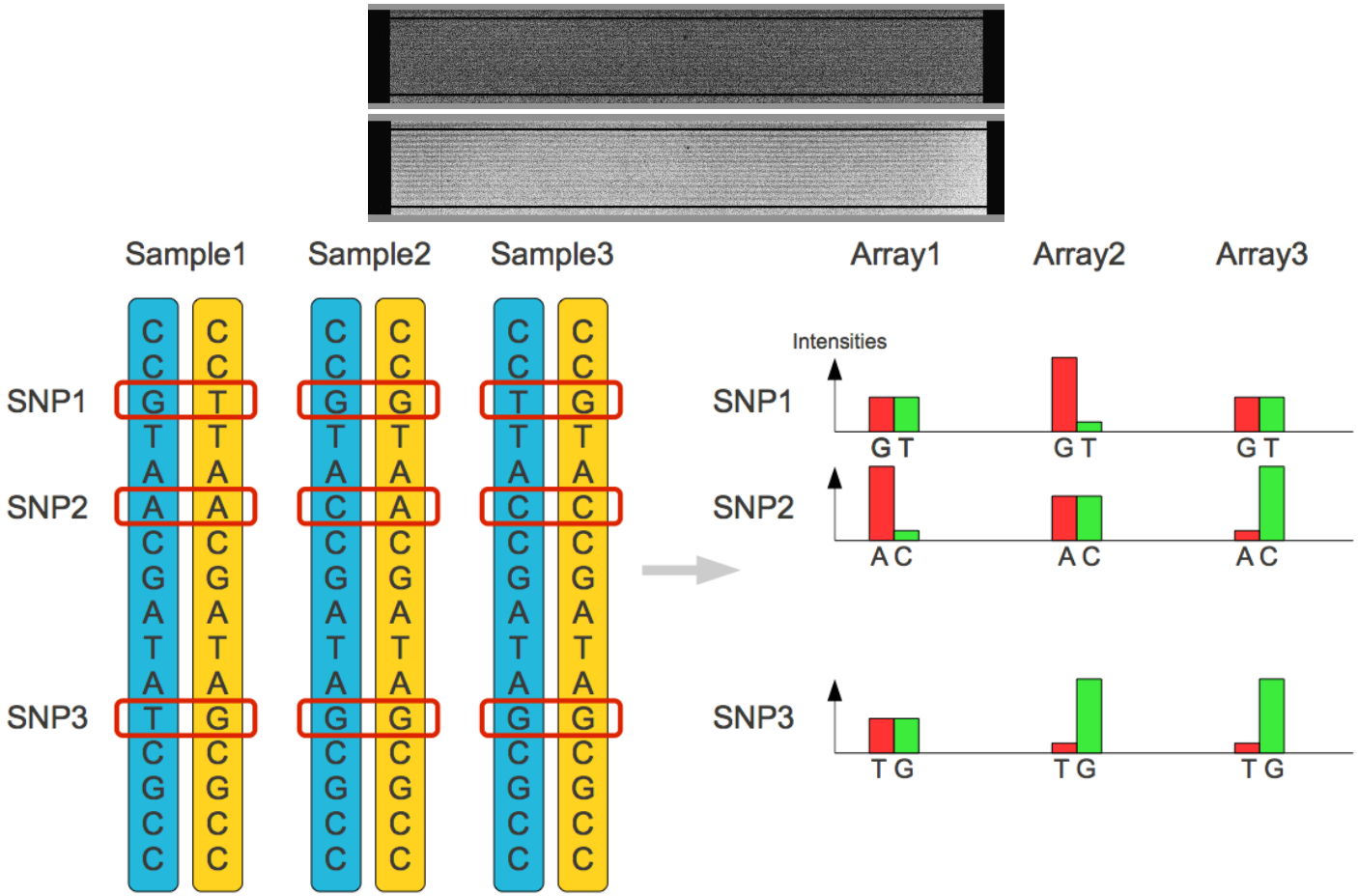
Oliphant et al. Biotechniques. 2002.

Affymetrix SNP arrays 6.0

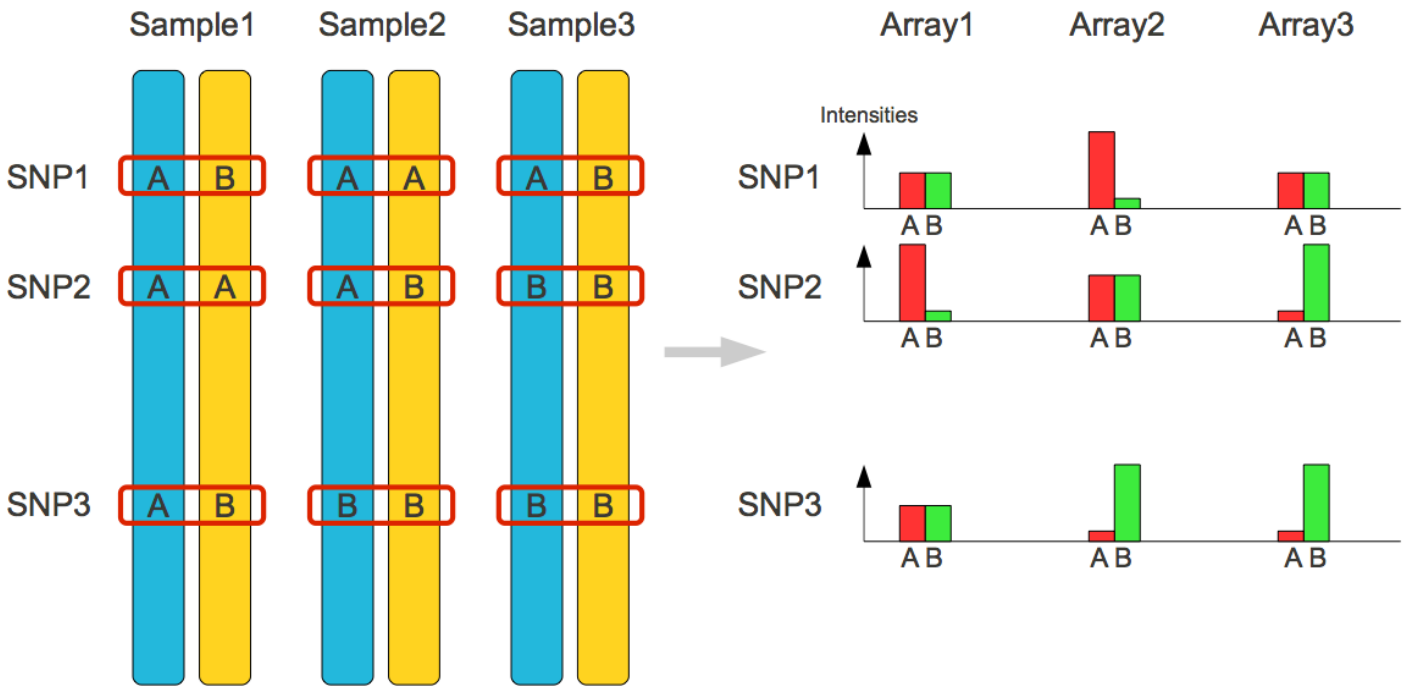
Affymetrix SNP Array 6.0

- 906,600 SNP
  - 482 000 SNP from SNP Array 5.0
  - 424 000 new tag-SNP
- 946,000 CNV
  - 202,000 probes targeting 5 677 regions from the 'Toronto Database of Genomic Variants'
  - 744,000 probes, evenly spaced along the genome

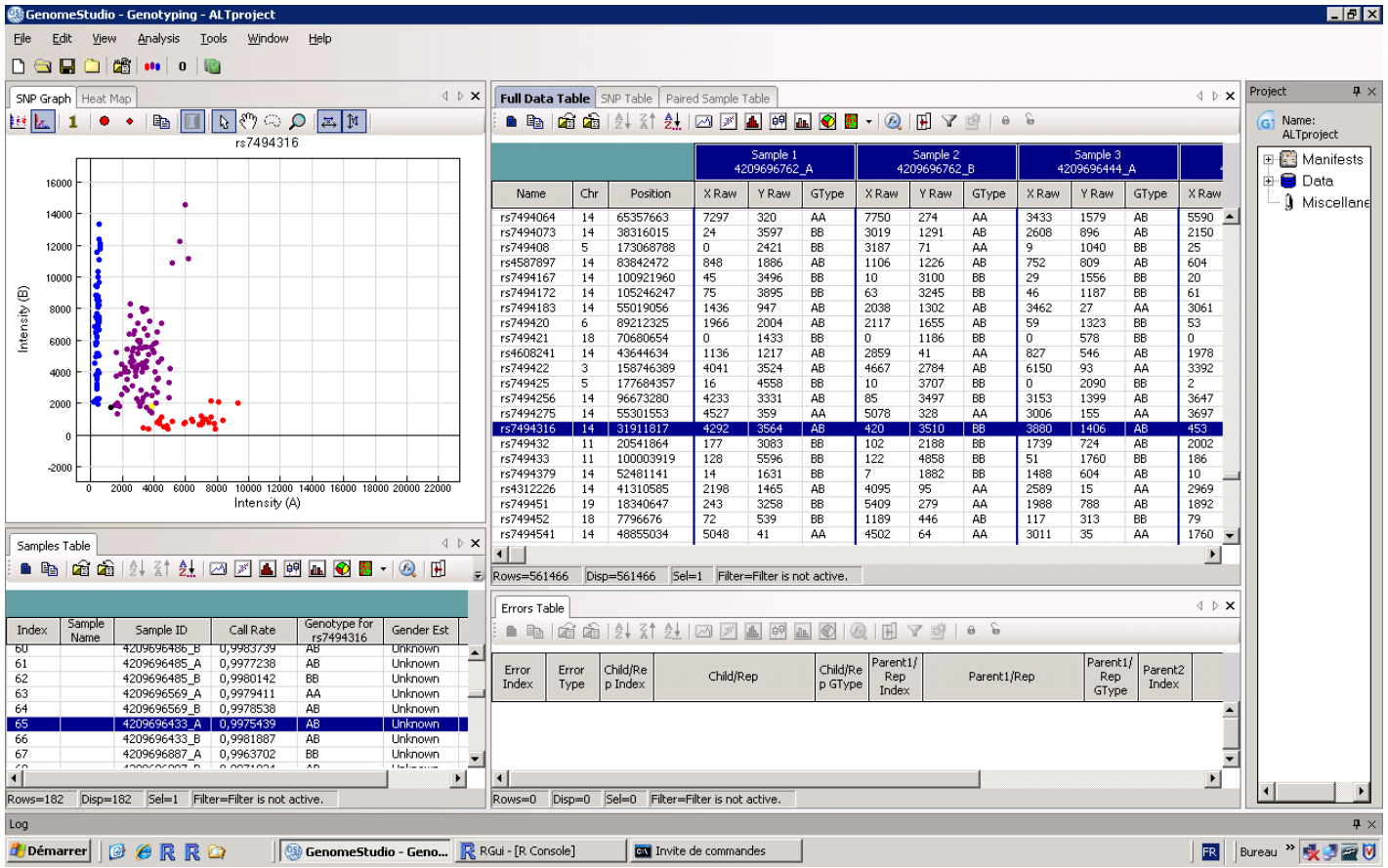
Intensity values for both alleles



Intensity values for both alleles



Genome Studio (Illumina)

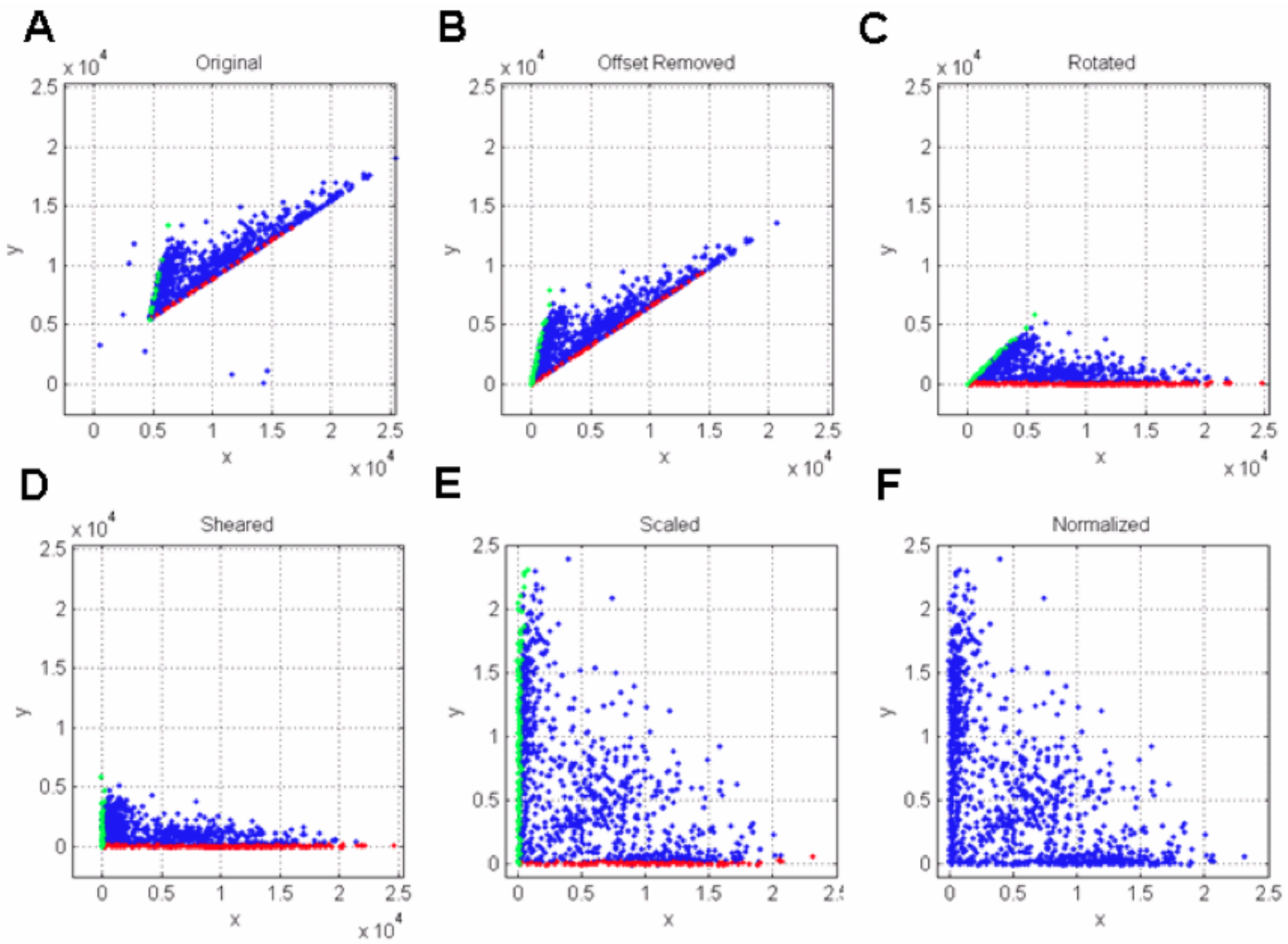


Preprocessing / normalization

Sources of variability

- Preparing the samples
  - MRNA preparation
  - Reverse transcription to cDNA
  - Dye labeling
- Spotting the chips
  - PCR amplification
  - Pin geometry and surface features
  - Amount of cDNA transported by pins
  - Amount of cDNA fixated on slide
- Hybridization process
  - Hybridization parameters (temperature, time, amount of sample)
  - Spatial dis-homogeneity of hybridization on the slide
  - Non-specific hybridization Image production and processing:
  - Non-linear transmission, saturation effects, variations in spot shape
  - Global background shining, local overshining from neighboring spots

Normalization - Example (Illumina)



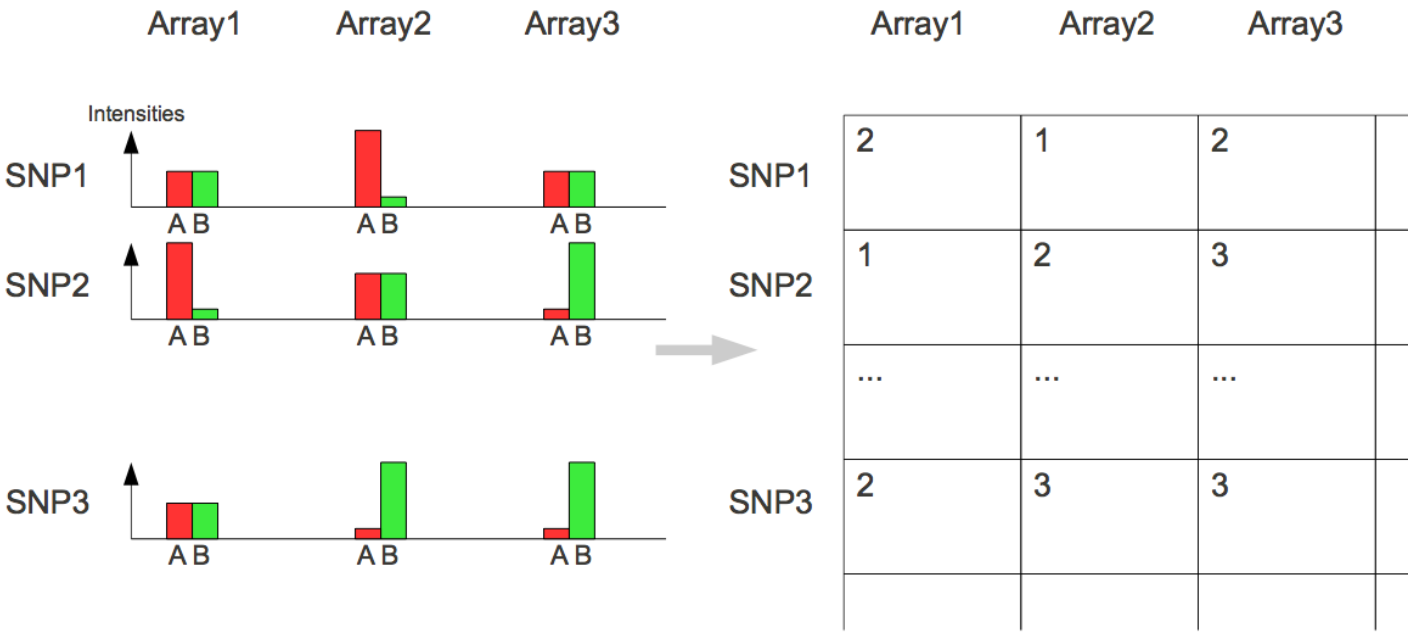
www.illumina.com

Normalization - Example (Illumina)

1. Outlier removal
2. Background estimation
3. Rotational estimation
4. Shear estimation
5. Scaling estimation

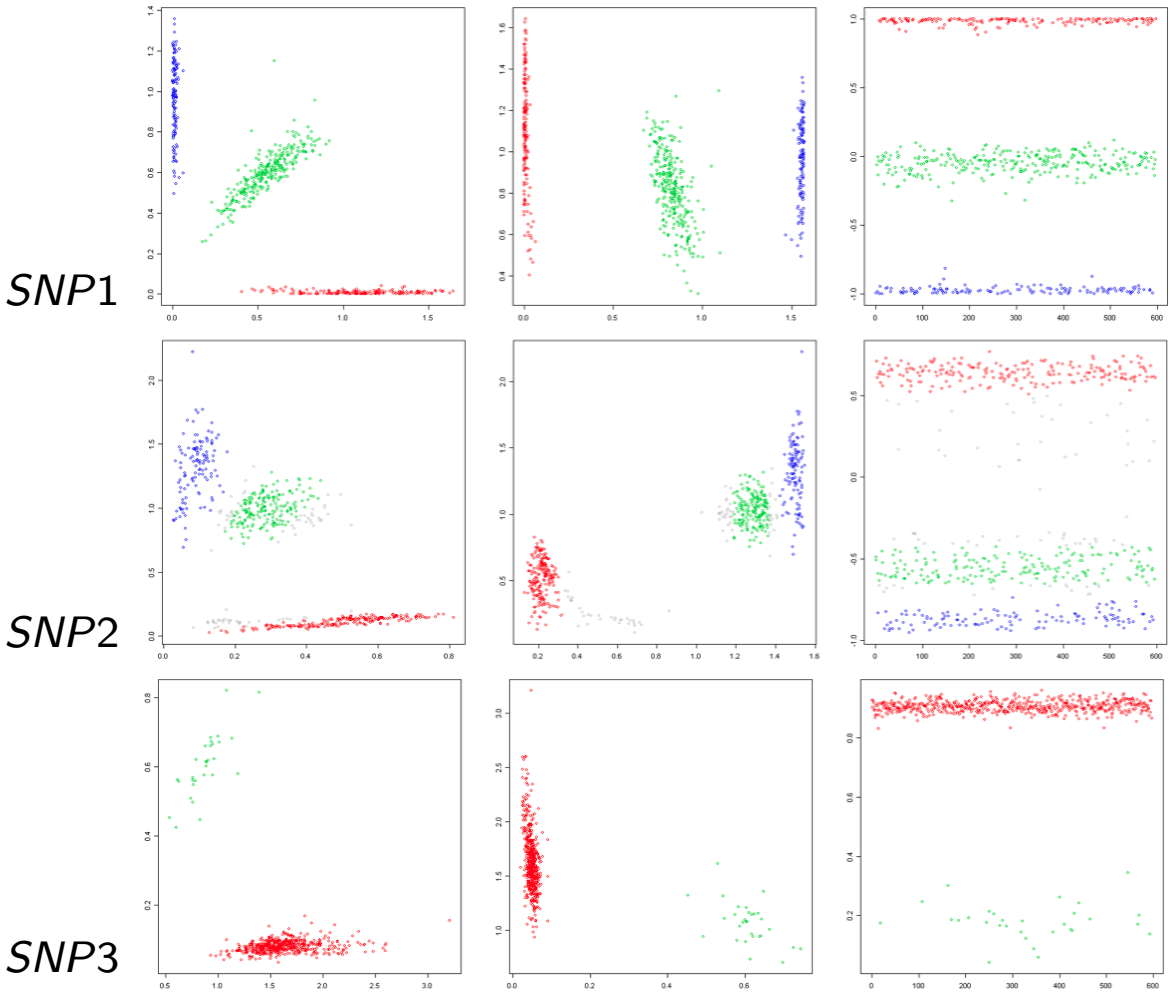
Genotyping

Objective



Genotyping

Summary indexes



Methods

Classification

- K-means, K-medoids
  Limits: sensitive to initial values, need for class number specification, similar group sizes, ...
- Mixture models
  - EM algorithm
  - Bayesian framework

  Limits: sensitive to the model choice, need for class number specification, ...
- ...

Comparison of genotying algorithms for Illumina's SNP arrays

Ritchie et al. BMC Bioinformatics. 2011.

Data structure

	$marker_1$	$marker_2$	...	$marker_m$	phenotype	age	sex	...
$sample_1$	0	2		0	y_1	42	M	...
$sample_2$	1	1		0	y_2	63	F	...
...								
$sample_n$	0	1		2	y_n	27	F	...

Whole genome sequencing (WGS) and whole exome sequancing (WES)

A run (=realization of a full process by the machine) products a large number of reads (=strings of bases), corresponding to DNA/RNA sequences.

+ long read length
- low throughput: only 10^6 reads per run

Roche/454 GS FLX
read length: 700bp
read number: 1M
run time: 23 hours

Ion Torrent Proton
read length: ~200bp
read number: 80M
run time: 2-4 hours

+ very high throughput: 10^9 reads /run
- short read length

Illumina HiSeq
read length: 100bp
read number: 6G
run time: 11 days

Life SOLiD 3
read length: 75bp
read number: 3G
run time: 14 days

+ absence of amplification of the input genomic material to sequence
- low throughput

Helicos HeliScope
read length: 35bp
read number: 1G
run time: 8 days

Pacific Biosciences RS
read length: ~3000bp
read number: 150K/smrt cell
run time: 10 hours

from Smahane CHALABI (CNRGH)

Whole genome sequencing (WGS) and whole exome sequancing (WES)

Data preprocessing

- Raw reads
- Quality check of raw reads
- Mapping

Variant calling

Call SNPs, indels and some SVs (separately or simultaneously)

Microarrays vs. Sequencing

Microarrays

- Data easily stored and analyzed
- Allele calling is standardized
- Experiment well understood
- Number of statistical tests known and carefully considered
- SNP interrogated directly and indirectly

Sequencing

- Requires massive storage capacity
- Allele and Structural Variation calling still in flux
- Experiment not clearly defined
- SNPs interogated at different depths

Preprocessing

Phenotypes Quality Controls

The phenotype is critical to good genetic studies

- Precise
- The closest to a gene product

Preprocessing

Phenotypes Quality Controls

In practice

- Create standard report with descriptive statistics
- Check distribution of quantitative traits
- Look for outliers
- If needed, impute missing phenotype

Preprocessing

Genotypes Quality Controls

- Call rates
- Sex inconsistencies
- Hardy Weinberg Equilibrium test
- Minor alele frequencies
- Population stratification

Data filtering

Call rates

No consensual threshold. Typically:

- Individuals with more than 10% of missing SNPs are removed
- SNPs with more than 5% of mising samples are removed (depends on the sample size)

Preprocessing

Sex inconsistency

Comparison between the reported sex and the predicted sex by from X-chromosome markers heterozygosity.

Data filtering

Hardy Weinberg Equilibrium test

HWE test is used to detect genotyping errors (usually at level 10^{-7}, 10^{-5}, 10^{-3}, ...).

Hardy Weinberg disequilibrium test

Hardy-Weinberg principle

Both allele and genotype frequencies in a population remain constant

$$p^2 + 2pq + q^2 = 1$$

χ^2 test for deviation

$$\frac{(N_{AA} - n\hat{p}^2)^2}{n\hat{p}^2} + \frac{(N_{AB} - n2\hat{p}(1-\hat{p}))^2}{n2\hat{p}(1-\hat{p})} + \frac{(N_{BB} - n(1-\hat{p})^2)^2}{n(1-\hat{p})^2} \xrightarrow{\mathcal{L}} \chi^2_1$$

Data filtering

Minor Allele Frequency

Most GWAS studies (particularly microarrays based studies) are powered to detect a disease association with common SNPs (MAF$\geq$ 0.05).
Depending on the sample size, SNPs with MAF$<$0.01 or 0.05 are removed.

Case-Control association tests

Allelic tests

- Sampling unit: allele
- Hardy Weinberg equilibrium assumption

Genotypic tests

- Sampling unit: Individual
- Additive / dominant / recessive models

Allelic tests

Pearson's χ^2 test for association

Test for independance between trait and allele

- Table for a diallelic locus

	Cases	Controls	Total
Allele A	n_{11}	n_{12}	n_{1+}
Allele a	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n_{++}

- Tested hypotheses:
  - H_0: There is no association between trait and allele
  - H_1: There is an association between trait and allele
- Test statistic:

$$X^2 = \sum_{ij} \frac{(n_{ij} - \frac{n_{i+}n_{+j}}{n_{++}})^2}{\frac{n_{i+}n_{+j}}{n_{++}}} \underset{H_0}{\rightarrow} \chi^2_1$$

Example

Leber's Hereditary Optic Neuropathy (LHON) disease and marker rs6767450 (Phasukijwattana et al., 2010)

- Table for genotypes

	AA	Aa	aa
Cases	6	8	75
Controls	10	66	163

- Corresponding table for alleles

	Cases	Controls	Total
Allele a	158	392	550
Allele A	20	86	106
Total	178	478	656

from T. Thornton

Example

Pearson's χ^2 test for association

Table for alleles

	Cases	Controls	Total
Allele A	158	392	550
Allele a	20	86	106
Total	178	478	656

Expected counts

	Cases	Controls	Total
Allele A	149.2378	400.7622	550
Allele a	28.7622	77.2378	106
Total	178	478	656

Example

Pearson's χ^2 test for association

Table for alleles

	Cases	Controls	Total
Allele A	158	392	550
Allele a	20	86	106
Total	178	478	656

- Test statistic:

$$X^2 = \frac{(158 - 149.2378)^2}{149.2378} + ... + \frac{(86 - 77.2378)^2}{77.2378} = 4.369$$

- p-value:

$$p = \mathbb{P}(X^2 \geq 4.369) = 0.037$$

Allelic tests

Fisher's exact test for association

For contingency tables that have cells with small expected counts

- Table for a diallelic locus

	Cases	Controls	Total
Allele A	21	14	35
Allele a	3	10	13
Total	24	24	48

- Assumption: Marginal counts of the table are fixed
- Tested hypotheses:
  - H_0: There is no association between trait and allele
  - H_1: There is an association between trait and allele
- Test statistic: X the number of cas alleles of type A

$$X \underset{H_0}{\sim} \mathcal{H}(N, m, n)$$

Allelic tests

Fisher's exact test for association

- Table for a diallelic locus

	Cases	Controls	Total
Allele A	21	14	35
Allele a	3	10	13
Total	24	24	48

- Probability distribution for X:

x	11	12	13	14	15	16	17	18	19	20	21	22	23
P_x	10^{-5}	3.10^{-4}	.004	.021	.072	.162	.241	.241	.162	.072	.021	.004	3.10

- Rejection region at level $\alpha = 5\%$:

$$\Gamma = \{11, 12, 13, 14, 21, 22, 23, 24\}$$

- Conclusion: $21 \in \Gamma$

A Fast Unbiased and Exact Allelic Test (fueatest)

- Classical allelic test are biased if the Hardy Weinberg assumption is not true (for both cases and controls)
- Table for genotypes

	AA	Aa	aa	Total
Cases	$D0$	$D1$	$D2$	n_D
Controls	$C0$	$C1$	$C2$	n_C

- Corresponding table for alleles

	Cases	Controls	Total
Allele A	$2D_0 + D_1$	$2C_0 + C_1$	$2n_0 + n_1$
Allele a	$2D_2 + D_1$	$2C_2 + C_1$	$2n_2 + n_1$
Total	$2n_D$	$2n_C$	$2n$

- The unbiased allelic test is based on the same statistic as the χ^2 allelic test but on the multinomial sampling of genotypes instead of alleles taken independently:

$$(D_0, D_1, D_2) \underset{H_0}{\sim} \mathcal{M}(n_D, p_{D_0}, p_{D_1}, p_{D_2})$$
$$(C_0, C_1, C_2) \underset{H_0}{\sim} \mathcal{M}(n_C, p_{C_0}, p_{C_1}, p_{C_2})$$

Genotypic test

Pearson's χ^2 test

	AA	Aa	aa	Total
Cases	$D0$	$D1$	$D2$	n_D
Controls	$C0$	$C1$	$C2$	n_C
Total	n_0	n_1	n_2	

Test statistic:

$$X^2 = \sum_{ij} \frac{(n_{ij} - \frac{n_{i+}n_{+j}}{n_{++}})^2}{\frac{n_{i+}n_{+j}}{n_{++}}} \underset{H_0}{\rightarrow} \chi^2_2$$

Example

Pearson's χ^2 test

	AA	Aa	aa
Cases	6	8	75
Controls	10	66	163

- Test statistic: $X^2 = 13.15$
- p-value $p = 0.001395$

Genotypic test

Cochran Armitage trend test for association

- The most used genotypic test for unrelated individuals
- Let
  - $Y_i = 1$ if i is a case (0 if i is a control)
  - X_i the genotype (coded 0,1,2)
- Linear probability model :

$$\pi_i = \alpha + \beta X_i \text{ with } \pi_i = \mathbb{P}(Y = 1|X = i)$$

- Tested hypotheses :

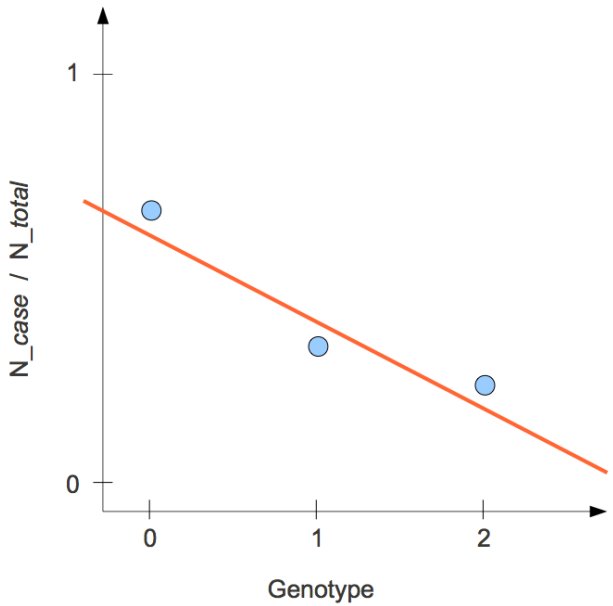
$$H_0 : \pi_0 = \pi_1 = \pi_2 \text{ vs } H_1 : \pi_0 < \pi_1 < \pi_2$$

- Test statistic :

$$\frac{\hat{\beta}}{Var(\hat{\beta})} \underset{H_0}{\rightarrow} \chi^2_1$$

Genotypic test

Cochran Armitage trend test for association



Remarks

- The Cochran Armitage trend test has a better power than the Pearson's χ^2 test if the suspected trend is correct
- The test can be shown to be valid when the HWE does not hold

Example

Cochran Armitage trend test for association

	AA	Aa	aa
Cases	6	8	75
Controls	10	66	163

- Test statistic: $X^2 = 3.74$
- p-value: $p = 0.053$

Genotypic test

Logistic regression

- Let X_{1i} the genotype for the SNP of interest
- Let X_{ji} $(j \geq 2)$ adjustmnt variables
- Logistic model:

$$logit(\mathbb{P}(Y = 1|X)) = ln\frac{\mathbb{P}(Y = 1|X)}{1 - \mathbb{P}(Y = 1|X)}$$

$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

$$\Leftrightarrow \mathbb{P}(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_k X_k}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_k X_k}}$$

- Tested hypotheses:
  - $H_0 : \beta_1 = 0$
  - $H_1 : \beta_1 \neq 0$

Genotypic test

Logistic regression

- Let $\hat{\beta}_1$ the maximum likelihood estimator of β_1
- Classical tests
  - Wald test:

$$T = \frac{\widehat{\beta_1}}{\sqrt{\hat{V}(\widehat{\beta_1})}} \underset{H_0}{\rightarrow} N(0,1)$$

  - Likelihood ratio test:

$$LR = -2ln(\frac{sup(\mathcal{L}(\beta_1 = 0))}{sup(\mathcal{L}(\beta_1 \in]-\infty;\infty[))}) \underset{H_0}{\rightarrow} \chi^2_1$$

  - Score test:

$$S = \frac{\frac{\partial log\mathcal{L}(\beta_1)}{\partial \beta_1}(\beta_1 = 0)}{-\mathbb{E}(\frac{\partial^2}{\partial \beta_1^2} log\mathcal{L}(\beta_1 = 0)|\beta_1 = 0)} \underset{H_0}{\rightarrow} \chi^2_1$$

Odds ratios

Genotypes

	AA	Aa	aa	Total
Cases	$D0$	$D1$	$D2$	n_D
Controls	$C0$	$C1$	$C2$	n_C

Typically choose a reference genotype (eg *aa*).

$$OR_{AA} = \frac{\text{odds of disease for an individual with the AA genotype}}{\text{odds of disease for an individual with the aa genotype}}$$

$$OR_{Aa} = \frac{\text{odds of disease for an individual with the Aa genotype}}{\text{odds of disease for an individual with the aa genotype}}$$

where

$$\text{"odd"} = \frac{\pi}{1-\pi}$$

Odds ratios

Genotypes

	AA	Aa	aa	Total
Cases	$D0$	$D1$	$D2$	n_D
Controls	$C0$	$C1$	$C2$	n_C

For the logistic model:

- $OR = exp(\beta_1)$ (proportional odds assumption)
- $1 - \alpha$ confidence interval :

$$IC_{1-\alpha} = [exp(\hat{\beta}) \pm q_{1-\alpha/2}\sqrt{\hat{V}(\hat{\beta}_1)})]$$

Quantitative trait

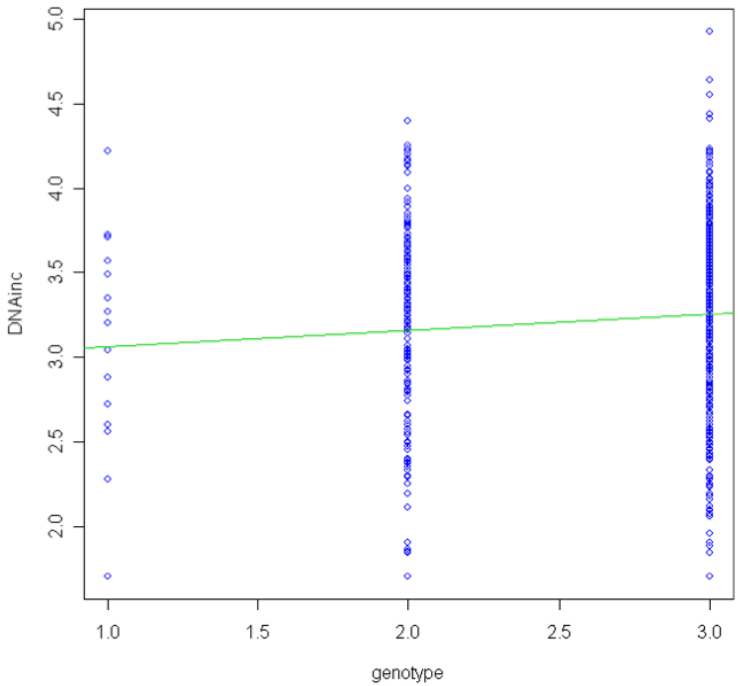
Quantitative Trait Loci (QTL) mapping aim at identifying genetic loci that influence the phenotypic variation of a quantitative trait

Genetic models

- Dominant
- Recessive
- Additive
- Multiplicative

Quantitative trait

Linear regression model

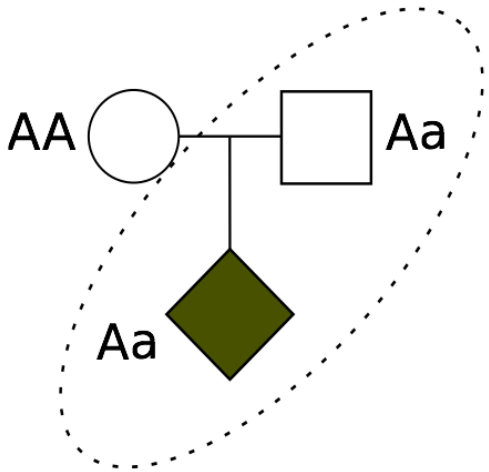


$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Family based association tests

Transmission Disequilibrium Test (TDT)

- Based on trio families (two parents and an affected offspring)
- All are genotypes for a diallelic marker A/a
- Only heterozygous parents are used (homozygous parents are not informative)
- Under the null hypothesis, A is transmited as often as a



Family based association tests

Transmission Disequilibrium Test (TDT)

Combination of transmitted and non-transmitted marker alleles A and a among $2n$ parents of n affected children.

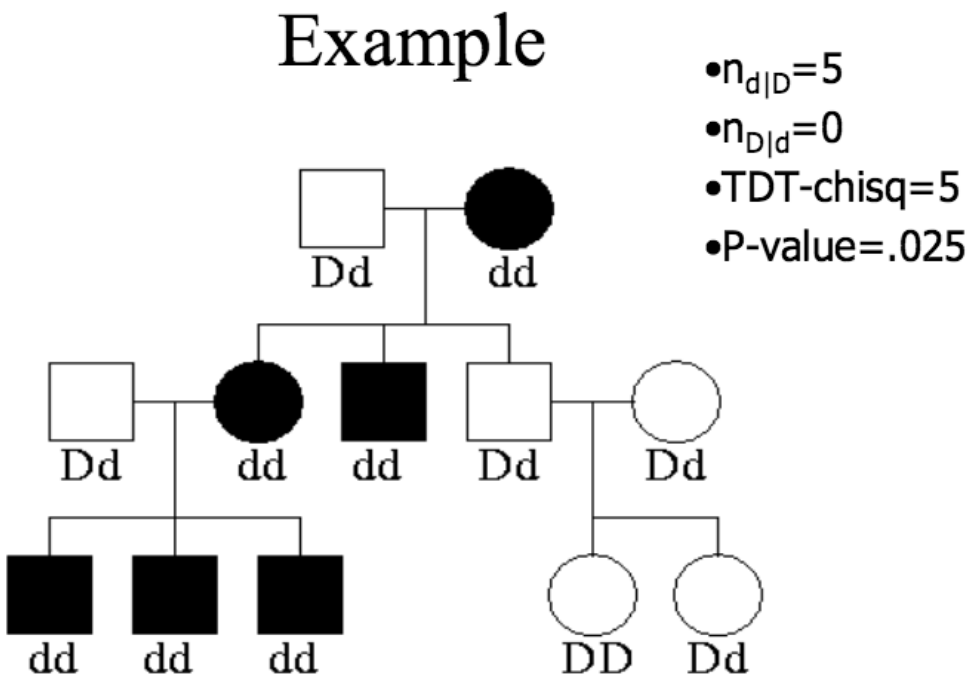
Transmitted allele \ Non-transmitted allele	A	a	Total
A	a	b	a+b
a	c	d	c+d
Total	a+c	b+d	2n

Test statistic:

$$X^2 = \frac{(b - \frac{b+c}{2})^2}{\frac{b+c}{2}} + \frac{(c - \frac{b+c}{2})^2}{\frac{b+c}{2}} = \frac{(b-c)^2}{b+c} \underset{H_0}{\rightarrow} \chi^2_1$$

Family based association tests

Transmission Disequilibrium Test (TDT)



Family based association tests

FBAT

Generalization of the TDT that can deal with

- general trait
- multi-allelic markers
- missing parents

Family-based vs. Case-control

Family based methods

- robust to population substructure
- robust to HWE failure
- more powerful for rare highly penetrant diseases

Case Control

- Test for HWE in controls
- More powerful in most other situations

Multiple testing

Problem

Under the complete null hypothesis (H_{0i} true for all i) selecting SNPs based on the usual 5% threshold would lead to a large number of false positives:

$$\mathbb{E}(\text{number of false positives}) = 10^6 \times 0.05 = 50,000$$

Cost

- False positives $\Rightarrow$ laboratory cost
- False negatives $\Rightarrow$ discovery/publication cost

FWER procedures

Strategy

1. Choose an error criterion
2. Apply a procedure targeting the criterion

Remark

Most procedures mainly focuse on false positives related error criteria (FWER, FDR, ...)

FWER procedures

'Effective' number of independent tests

Due to the correlations among test statistics induced by linkage disequilibrium, the 'effective' number of independent tests is expected to be smaller than m ('genome wide significance' concept).

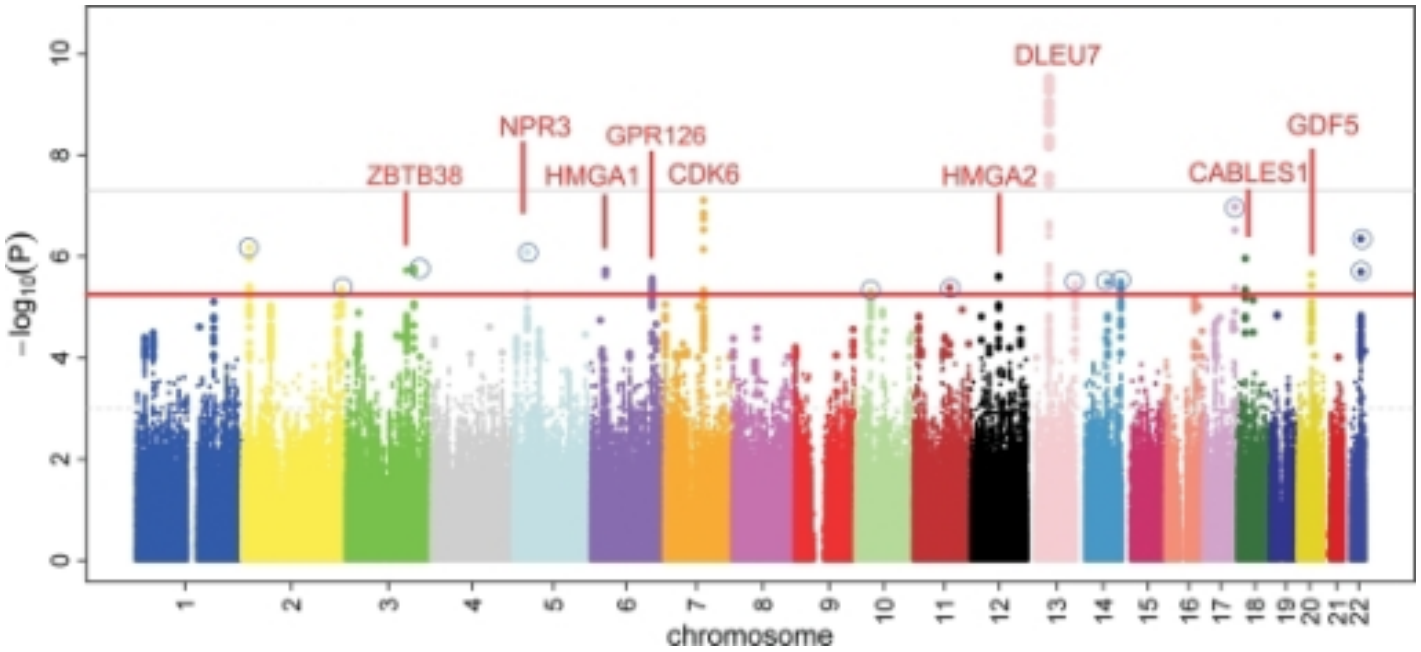
Classes of relaxation methods

- Permutation testing
- Principal component analysis
- Analysis of blocks of LD

To be used with caution!

Results presentation

Manhattan plot



Estrada et al, Hum Mol Genet, 2009 .

Population stratification

Population stratification occur if the sample consists of different populations.

Population stratification

False positives due to admixture

- Population 1: $p = 1$

	Allele A	Allele B	Total
Affected	64	16	80
Unaffected	16	4	20
Total	80	20	

- Population 2: $p = 1$

	Allele A	Allele B	Total
Affected	4	16	20
Unaffected	16	64	80
Total	20	80	

- Populations combination: $p = 6.6 \times 10^{-7}$

	Allele A	Allele B	Total
Affected	68	32	100
Unaffected	16	4	100
Total	100	100	

Population stratification

False negatives due to admixture

- Population 1: $p = 4.4 \times 10^{-14}$

	Allele A	Allele B	Total
Affected	20	80	100
Unaffected	80	20	100
Total	100	100	

- Population 2: $p = 4.4 \times 10^{-14}$

	Allele A	Allele B	Total
Affected	80	20	100
Unaffected	20	80	100
Total	100	100	

- Populations combination: $p = 1$

	Allele A	Allele B	Total
Affected	100	100	200
Unaffected	100	100	200
Total	200	200	

Population stratification

How to detect stratification - QQ plot

Inflation factor

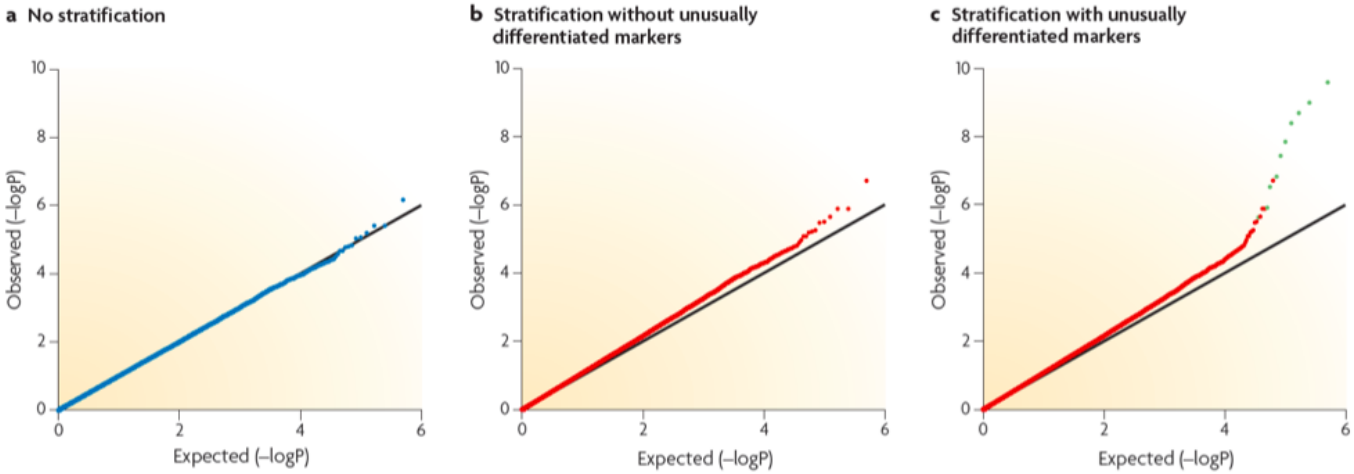


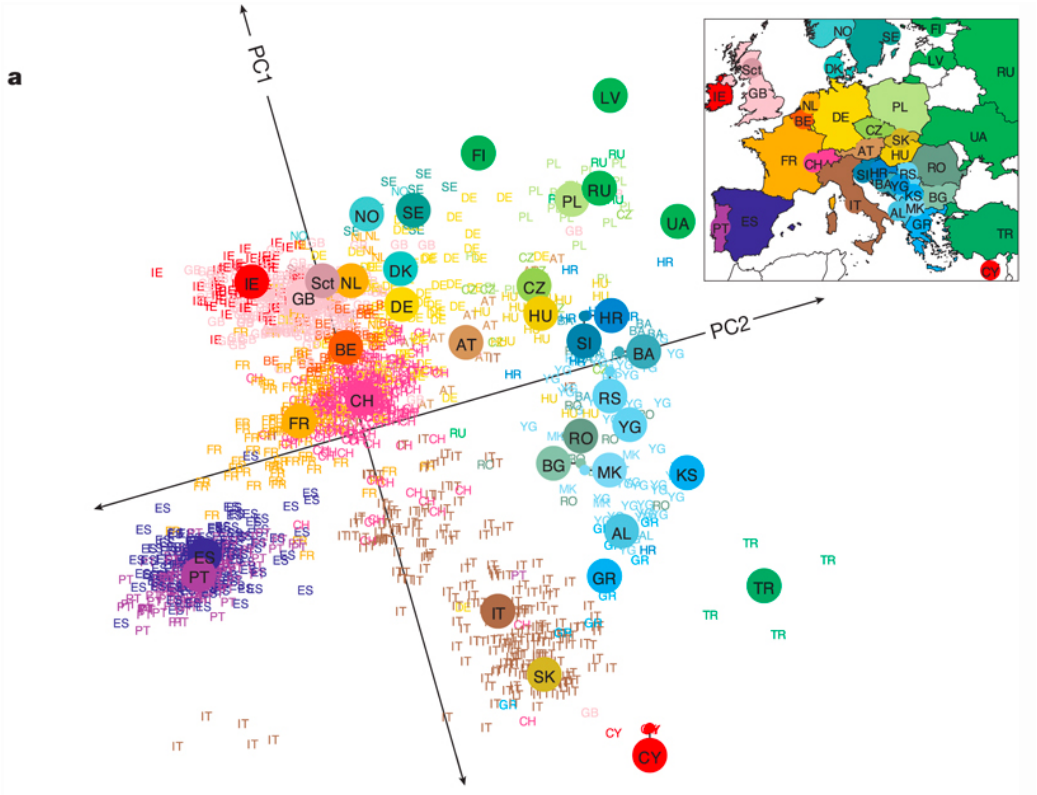
Figure 1 | **P–P plots for the visualization of stratification or other confounders.** The figure shows simulated P–P plots under three scenarios for genome-wide scans with no causal markers. **a** | No stratification: *p*-values fit the expected distribution. **b** | Stratification without unusually differentiated markers: *p*-values exhibit modest genome-wide inflation. **c** | Stratification with unusually differentiated markers: *p*-values exhibit modest genome-wide inflation and severe inflation at a small number of markers.

Price et al. New approaches to population stratification in genome-wide association studies. Nat Rev Genet 2010.

Population stratification

How to detect stratification - PCA

Population structure within Europe

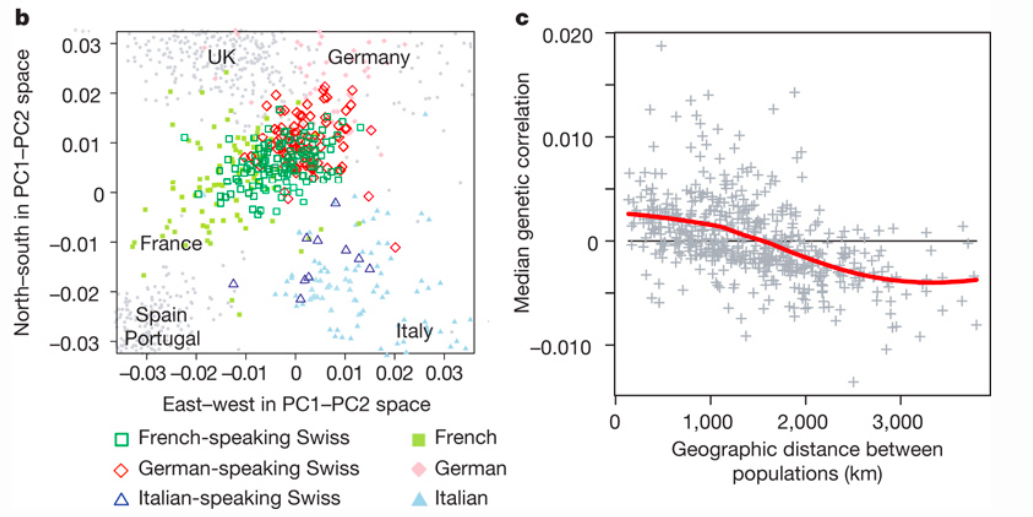


Novembre J et al. Genes mirror geography within Europe. Nature. 2008

Population stratification

How to detect stratification - PCA

Population structure within Europe



Novembre J et al. Genes mirror geography within Europe. Nature. 2008

Population stratification

How to correct for stratification

- Family-based design :
  - TDT
- Population-based design :
  - Structured association testing
  - Genomic control
  - Regional admixture mapping
  - PCA
  - Multivariate regression models

Population stratification

Structured association

- Trim high quality SNPs to be in linkage equilibrium (eg $r^2 < 0.2$)
- Using the genotype data in a Bayesian clustering approach, assign each individual to a subgroup
- Number of subpopulations and their allele frequencies are estimated using a Markov Chain Monte-Carlo method

Pritchard et al, Am J Hum Genet, 2000

Population stratification

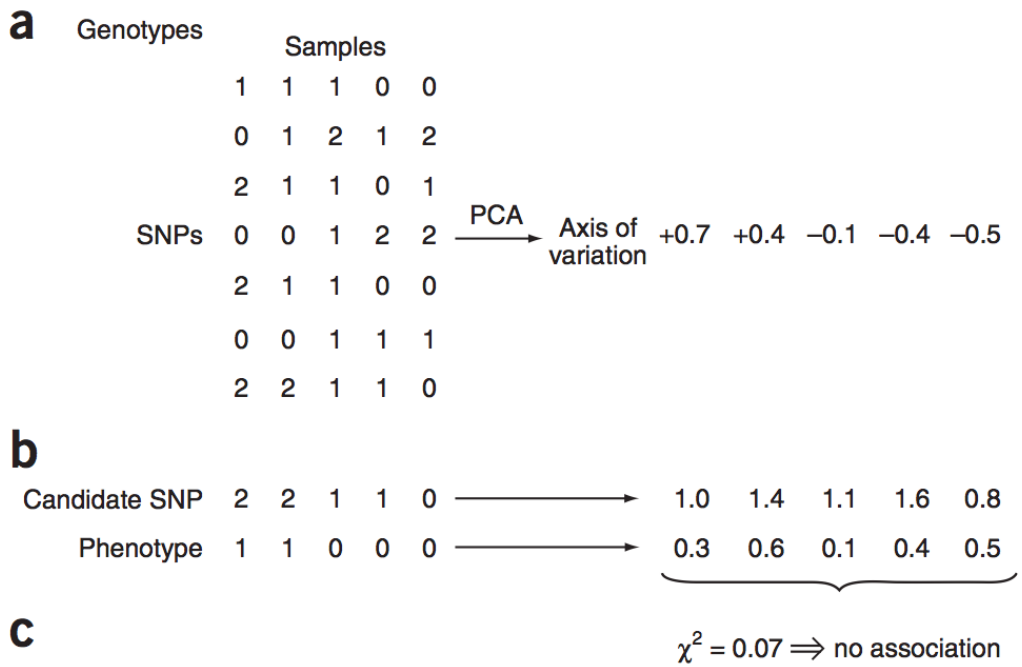
Genomic control

- Assumption: $Y^2 = \lambda\chi^2$
- Inflation factor estimation: $\hat{\lambda} = \frac{median(X_1^2,...,X_M^2)}{0.456}$ where M is the number of unlinked markers

Devlin et al., Theor Popul Biol. 2001

Population stratification

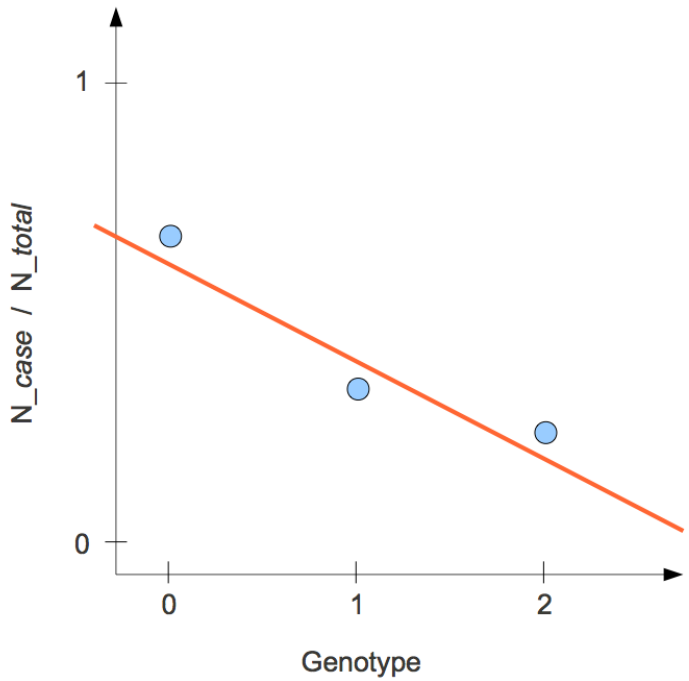
Eigenstrat - PCA



Price et al. Nature Genetics. 2006.

Population stratification

Eigenstrat - Cochran Armitage trend test



Generalization

$$(n-k-1) \times [Corr(G^*, P^*)]^2 \xrightarrow{\mathcal{L}} \chi^2_1$$

Population stratification

PCA

Warning: not adapted to familial data

General statistical setting

Statistical model

- Let $(\mathcal{X}, \mathcal{F}, \mathcal{P})$ a statistical model where $\mathcal{X}$ is the sample space, $\mathcal{F}$ a σ-field and $\mathcal{P}$ a family of probability measures.
- Let x an observation and X the corresponding random variable such that $X \sim P \in \mathcal{P}$
- Often $\mathcal{P} = (P_\theta)_{\theta \in \Theta}$ where θ is the target of the inference

Types of multiplicity

One or two samples problems with multiple endpoints

- Sample space: $\mathcal{X} = \mathbb{R}^{m \times n}$
  The same n observational units are measured with respect to m different endpoints
- Number of tests: m
- Examples: Differential gene expression analysis - Genome Wide Association Studies

Types of multiplicity

k-samples problems with localized comparisons

- Sample space: $\mathcal{X} = \mathbb{R}^{\sum_i n_i}$
  k groups of observational units are considered.
- Number of tests:
  - $k - 1$ if there is one control group
  - $\frac{k(k-1)}{2}$ is all pairwise comparisons are considered
- Example: ANOVA

Tested hypotheses

Multiple testing problem

Let $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta \in \Theta}, \mathcal{H}_0)$ a multiple testing problem.

Family of tested hypotheses

$$\mathcal{H}_0 = \{H_{0i} : i \in I\}$$

defines a family of null hypotheses for an arbitrary index set I such that

$$|I| = m \in \mathbb{N}$$

Null and alternative hypotheses

- $H_{0i} : \theta \in \Theta_{0i}$
- $H_{1i} : \theta \in \Theta_i \backslash \Theta_{0i}$

Tested hypotheses

Complete null hypothesis

The "complete null hypothesis" (or "global null hypothesis") is defined by:

$$H_0^c = \cap_{i \in I} H_{0i}$$

Remark

In the following, we will assume that H_0^c is non empty (ie: null hypotheses are compatible)

Multiple testing procedure

Definition

A multiple testing procedure for $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta \in \Theta}, \mathcal{H}_0)$ is a measurable mapping:

$$\varphi : \mathcal{X} \to \{0,1\}^m$$

such that

$$\varphi(x_i) = \begin{cases} 1 & \text{if } H_{0i} \text{ is rejected} \\ 0 & \text{if } H_{0i} \text{ is not rejected} \end{cases}$$

Rejection regions

Let $T = (T_i)_{i \in I}$ a m-vector of test statistics and let $(\alpha_1, \alpha_2) \in [0,1]^2$. We consider multiple testing procedures based on nested rejection regions such that:

$$\forall i \in I, \alpha_1 \leq \alpha_2 \Rightarrow \Gamma_{i,\alpha_1}(T_i) \subset \Gamma_{i,\alpha_2}(T_i)$$

where $\Gamma_{i,\alpha_j}(T_i)$ is the rejection region for an α_j level.

Remark

In the following, we will consider rejection regions such that $\Gamma_{i,\alpha_1}(T_i) = [\gamma_i, +\infty[$

Type I error criteria

General decision pattern

Reality \ *Decision*	H_0 not rejected	H_0 rejected	Total
H_0 true	U	V	m_0
H_0 false	T	S	m_1
Total	W	R	m

Let $I_0 = \{i \in I : \theta \in \Theta_{0i}\}$ and $I_1 = I \backslash I_0$

- $R = |\{i \in I : \varphi_i = 1\}| = V + S$
- $S = |\{i \in I_1 : \varphi_i = 1\}|$
- $T = m_1 - S$
- $V = |\{i \in I_0 : \varphi_i = 1\}|$
- $U = m_0 - V$

Type I error criteria

Remarks

- U, V, T, S, m_0 and m_1 depend on the unknown value of θ and are not observable
- Only r, m and w can be observed
- R, S, T, U and V are random variables.
- $(m, m_0, m_1) \in \mathbb{N}^3$

Type I error criteria

Definitions

Let $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta \in \Theta}, \mathcal{H}_0)$ a multiple testing problem and $\varphi = \{\varphi_i : i \in I\}$ a *MTP*

- Family-wise error rate

$$FWER_\theta(\varphi) = \mathbb{P}_\theta(V > 0)$$

- Generalized family-wise error rate

$$k - FWER_\theta(\varphi) = \mathbb{P}_\theta(V > k)$$

Type I error criteria

- False discovery proportion

$$FDP_\theta(\varphi) = \frac{V}{R \vee 1}$$

Remark: $FDP_\theta(\varphi)$ cannot be controled.

- False discovery rate

$$FDR_\theta(\varphi) = \mathbb{E}_\theta(FDP_\theta(\varphi))$$

- Positive false discovery rate

$$pFDR_\theta(\varphi) = \mathbb{E}_\theta\left(\frac{V}{R}|R > 0\right)$$

Remark: $pFDR_\theta(\varphi)$ cannot be controled.

Type I error criteria

- Per-family error rate

$$PFER_\theta(\varphi) = \mathbb{E}_\theta(V)$$

- Per-comparison error rate

$$PCER_\theta(\varphi) = \mathbb{E}_\theta\left(\frac{V}{m}\right) = \frac{PFER}{m}$$

- False discovery exceedance rate

$$FDX_\theta(\varphi) = \mathbb{P}_\theta(FDP_\theta(\varphi) > c); c \in (0,1)$$

Strong control / Weak control

Definitions

Let $\mathcal{E}_\theta(\varphi)$ a type I error criterion.

- The *MTP* φ is said to control $\mathcal{E}_\theta(\varphi)$ in the strong sense at level $\alpha \in [0,1]$ if

$$\sup_{\theta \in \Theta} \mathcal{E}_\theta(\varphi) \leq \alpha$$

- The *MTP* φ is said to control $\mathcal{E}_\theta(\varphi)$ in the weak sense at level $\alpha \in [0,1]$ if

$$\forall \theta \in \Theta_0; \mathcal{E}_\theta(\varphi) \leq \alpha$$

Remark

Strong control implies weak control

Error criteria comparison

Lemma 1

1. $FDR_\theta(\varphi) = pFDR_\theta(\varphi) \times \mathbb{P}_\theta(R > 0)$
2. If $m_0 = m$ then $FDR_\theta(\varphi) = FWER(\varphi)$
3. $\forall \theta \in \Theta; FDR_\theta(\varphi) \leq FWER_\theta(\varphi)$

Proof

Exercice

Power

Definitions

- Any power

$$anyPwr_\theta(\varphi) = \mathbb{P}(S > 0)$$

- All power

$$allPwr_\theta(\varphi) = \mathbb{P}(S = m_1)$$

- Average power

$$avgPwr_\theta(\varphi) = \mathbb{E}\left(\frac{S}{m_1}\right)$$

- True discovery rate

$$TDR_\theta(\varphi) = \mathbb{E}\left(\frac{S}{R}\right)$$

Exercice

Let consider three independent tests. Two null hypotheses are true, one is false. The following table give the joint distribution of (V, R).

r	0	1	1	2	2	2	3	3	3	3
v	0	0	1	0	1	2	0	1	2	3
$\mathbb{P}(R = r, V = v)$	0.2	0.4	0.02	–	0.2	0.1	–	–	–	–

1. Complete the table.
2. Calculate the *FDR* value for this scenario.

Unadjusted p-values

Definition

Let $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta \in \Theta})$ a statistical model and let φ a one-dimentional test for the single pair of hypotheses:

- $H_0 : \theta \in \Theta_0$
- $H_1 : \theta \in \Theta \backslash \Theta_0$

Assume that φ is based on a real valued test statistic $T : \mathcal{X} \to \mathbb{R}$ with a rejection region $\Gamma_\alpha \subset \mathbb{R}$ such that

$$\forall x \in \mathcal{X}, \varphi(x) = 1 \Leftrightarrow T(x) \in \Gamma_\alpha$$

The p-value of an observation $x \in \mathcal{X}$ with respect to φ is defined by

$$p_\phi(x) = \inf_{\alpha : T(x) \in \Gamma_\alpha} \left(\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(x) \in \Gamma_\alpha) \right)$$

Unadjusted p-values

Remarks

- Rejection regions are such that $\alpha_1 \leq \alpha_2 \Rightarrow \Gamma_{\alpha_1} \subseteq \Gamma_{\alpha_2}$
- If Θ_0 contains only one single element θ_0 and if P_{θ_0} is continuous, then
$$p_\varphi(x) = \inf(\alpha \in [0,1] : T(x) \in \Gamma_\alpha)$$
- Let $\Omega_{\mathcal{X}}$ denote the domain of $\mathcal{X}$.
$$\begin{array}{rcl} P_\varphi & : & \Omega_{\mathcal{X}} \to [0,1] \\ & & \omega \mapsto p_\varphi(x(\omega)) \end{array}$$
can be regarded as a random variable. It is just a test statistic.
- Let $\alpha \in (0,1)$ a fixed given significance level and assume that P_{θ_0} is continuous, then we have the duality $\varphi(x) = 1 \Leftrightarrow p_\varphi(x) \leq \alpha$

Unadjusted p-values

Theorem

Consider the m-vector of test statistics $T = (T_1, ..., T_m)$ and define a collection of m rejection regions $\{\Gamma_{\alpha_i} : i = 1, ..., m\}$ based solely on the marginal null distribution and such that

- i) $\sup_{\theta \in \Theta_0} \mathbb{P}(T_i \in \Gamma_{\alpha_i}) \leq \alpha$
- ii) Nested assumption $\alpha_1 \leq \alpha_2 \Rightarrow \Gamma_{\alpha_1} \subseteq \Gamma_{\alpha_2}$

1. The unadjusted p-values are stochastically larger than a uniform ditribution on the interval $[0, 1]$, that is:

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(p_i \leq z) \leq z$$

2. For a simple null hypothesis ($\Theta_0 = \{\theta_0\}$) and a continuousnull distribution $\mathbb{P}_{\theta_0}$

$$p_\varphi \sim \mathcal{U}([0, 1])$$

Adjusted p-values

Definition

Consider any multiple testing procedure (MTP) with rejection regions Γ_α. Then, we can define an m-vector of adjusted p-values $P^* = (P_1^*, ..., P_m^*)$ such that

$$P_i^* = \inf\{\alpha \in (0,1) : H_{0i} \text{ is rejecte at nominal multiple testing level } \alpha\}$$

Remarks

- The adjusted and unadjusted p-values can be calculated without prior specification of significance level α.
- Adjusted and unadjusted p-values reflect the strength of the evidence against each null hypothesis.

Bootstrap estimation of the test statistic/p-value null distribution

Procedure

- Generate B bootstrap samples

$$X^b = \{X_j^b : j = 1, ..., n\}, b = 1, ..., B$$

For the b^{th} sample, the $X_j^b, j = 1, ..., n$ are n iid copies of a random variable $X^* \sim \mathbb{P}_\theta$

- For each bootstrap sample X^b compute an m-vector of test statistics (or p-values) $T^b = (T_i^b : i = 1, ..., m)$ that can be arranged in an mxB matrix T^B with rows correponding to the m null hypotheses and columns to the B bootstrap samples.

Bootstrap estimation of the test statistic/p-value null distribution

Procedure

- Define m marginal cumulativedistributions functions p_i^B as the empirical CDF of the rows of the matrix T^B that is:

$$\hat{F}_i^B(z) = \frac{1}{B} \sum_{b=1}^{B} 1_{\{T_i^b \leq z\}}$$

Classes of MTP procedures

Margin-based procedures

For a margin-based MTP, each marginal test φ_i can be calibrated to keep a local significance level α_{loc}. The multiple test $\varphi = [\varphi_i : 1 \leq i \leq m]$ is then build up from these marginal tests by adjusting α_{loc} for the multiplicity of the problem.

Margin-based procedures

Single step procedure

Single step procedures carry out each individual test $\varphi_i; 1 \leq i \leq m$ at local significance level α_{loc} where α_{loc} is the result of a multiplicity correction of α

Example

The Bonferroni procedure consists in choosing $\alpha_{loc} = \frac{\alpha}{m}$

Margin-based procedures

Step-down procedures

Step-down procedures rely on an ordering of the hypothesis $H_{0(1)} \leq ... \leq H_{0(m)}$ which is induced by the order of the marginal p-values $p_{(1)} \leq \cdots \leq p_{(m)}$

Principle

1. Order the hypotheses
2. Test $H_{0(i)}$ at level $\alpha_{(i)}$
3. - If $H_{0(i)}$ is rejected, repeat step 2 for $H_{0(i+1)}$ (tested at level $\alpha_{(i+1)}$)
   - If $H_{0(i)}$ is not rejected, STOP.

Margin-based procedures

Step-up procedures

Step-up procedures also rely on an ordering of the hypothesis $H_{0(1)} \leq ... \leq H_{0(m)}$.

Principle

1. Order the hypotheses
2. Test $H_{0(m)}$ at level $\alpha_{(m)}$
3.
   - If $H_{0(m)}$ is is rejected, reject $H_{0(j)}$ for all $j \leq m$
   - If $H_{0(m)}$ is not rejected, repeat step 2 for $H_{0(m-1)}$.

Closed test procedures

Definition 1

The system $\mathcal{H}_m = \{H_{0i}\}, i \in I = \{1, ..., m\}$ is closed under intersection $(\cap - closed)$ if

$$\forall \emptyset \neq J \subseteq I, H_J = \cap_{j \in J} H_j = \emptyset \text{ or } H_J = \cap_{j \in J} H_j \in \mathcal{H}_m$$

Definition 2

The test φ is coherent if $\forall (i,j) \in I^2, H_i \subseteq H_j \Rightarrow \{\varphi_j = 1\} \subseteq \{\varphi_i = 1\}$

Closed test procedures

Theorem 1

Let $\mathcal{H} = \{H_i : i \in I\}$ a $\cap-$ *closed* system of hypotheses and let $\phi = \{\varphi_i, i \in I\}$ a coherent multiple testing procedure at local level α, then ϕ is a strongly FWER controlling multiple test at level α for $(\mathcal{X}, \mathcal{F}, \mathcal{P}, \mathcal{H})$

Closed test procedures

Theorem 2

Let $\mathcal{H} = \{H_i : i \in I\}$ a $\cap -$ *closed* system of hypotheses and let $\phi = \{\phi_i, i \in I\}$ a coherent multiple testing procedure at local level α, then we define the closed multiple test procedure (closed test) $\bar{\phi} = (\bar{\phi}_i : i \in I)$ based on ϕ by

$$\forall i \in I : \bar{\phi}_i(x) = min_{\{j:H_j \subseteq H_i\}} \phi_j(x)$$

It holds

1. The closed test $\bar{\phi}$ strongly controls the *FWER* at level α
2. For all $\emptyset \neq I' \subset I$, the restricted closed test $\bar{\phi}' = (\bar{\phi} : i \in I')$ is a strongly (at level α) FWER controlling multiple test for $\mathcal{H}' = \{H_i : i \in I'\}$
3. Both tests $\bar{\phi}$ and $\bar{\phi}'$ are coherent

Multivariate procedures

Goal

To incorporate the dependency structure of the data explicitely into the mutiple testing procedure in order to optimize the power

Classes of multivariate procedures

- Resampling-based procedures
- Methods based on the Central Lmit Theorem
- ...

Adaptive procedures

Goal

To improve the power of a type I error rate controlling procedure by incorporating in it a part $\mathcal{V}$ of the underlying distribution P_θ

Remark

Adaptation can relate to

- Dependence structure ($\mathcal{V} = \Sigma$)
- π_0 the proportion of true null hypotheses
- Alternative hypotheses structure

FWER control

Bonferroni procedure

The Bonferroni procedure is defined by

$$\hat{\varphi}_j^{Bonf} = 1_{\{P_j \leq \frac{\alpha}{m}\}}; 1 \leq j \leq m$$

Proposition

The Bonferroni procedure provides strong control of the *FWER* at level $\frac{m_0}{m}\alpha \leq \alpha$

Adjusted p-values

$$P_j^* = \alpha P_j$$

Remark

For large m values, the power of the Bonferroni procedure is very low.

FWER control

Sidak procedure

The Sidak procedures is defined by

$$\hat{\varphi}_j^{Sid} = 1_{\{P_j \leq 1-(1-\alpha)^{\frac{1}{m}}\}}; 1 \leq j \leq m$$

Proposition

The Sidak procedure provides strong control of the *FWER* at level α if $(P_1, ..., P_m)$ are jointly stochastically independent

Adjusted p-values

$$P_j^* = 1 - (1 - P_j)^m$$

FWER control

Holm procedure

The Holm procedure (step-down Bonferroni) is defined by

$$\hat{\varphi}_j^{Sid} = 1_{\{P_j \leq \frac{\alpha}{m-i+1}\}}; 1 \leq j \leq m$$

Proposition

The Holm procedure provides strong control of the *FWER* at level α.

Adjusted p-values

$$P_j^* = max((m-i+1)P_(i), P_{(i-1)}^*)$$

FWER control

Adaptation to the dependence structure

Let consider a Gaussian model

$$X \sim N(H\mu, \Sigma)$$

where $H \in \{0,1\}^m$, $\mu \in \mathbb{R}^{*m}$, $p_j(X) = \Phi(x_j)$ Σ known.

Theorem

If $t_\alpha(\Sigma) = inf\{t \in [0,1] : F_\Sigma(t) \geq \alpha\}$ with $F_\Sigma(t) = \mathbb{P}_{X \sim N(0,\Sigma)}(min_j(p_j(x)) \leq t)$, then

- $\forall \theta \in \Theta, FWER_\theta(t_\alpha(\Sigma)) \leq \alpha$
- for $\theta = (H, \mu)$ with $H_j = 0 \forall j$, $FWER_\theta(t_\alpha(\Sigma)) = \alpha$

FWER control

Westfall and Young procedure

The Westfall and Young's procedure is a multivariate resampling based method for estimating the adjusted p-values from a set of raw p-values:

$$p_i^* = \mathbb{P}_\theta(min_{j \in I_0} P_j \leq p_i) \leq \mathbb{P}_\theta(min_{j \in I} P_j \leq p_i)$$

FWER control

Westfall and Young procedure

Principle:

1. Calculate $p_i (i \in I)$
2. Calculate $p_i^b, i \in I, b \in \{1, ..., B\}$ (pour B bootstrap)
3. Estimate the adjusted p-values:

$$\hat{p}_i^* = \frac{\sum_{b=1}^{B} 1_{\{min_{j \in I}(p_j^b) \leq p_i\}}}{B}$$

FWER control

Exercice

1. Implement and apply the following MTP:
   - Bonferroni
   - Sidak
   - Holm
   - Westfall and Young
2. Check the results using the `p.adjust` function and the `multtest` package.

FDR control

Benjamini and Hochberg

Denote $p_{(1)} \leq p_{(m)}$ the ordered p-values for a multiple testing problem $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta \in \Theta}, \mathcal{H}_m)$ and $H_{0(1)} \leq ... \leq H_{0(m)}$ the ordered null hypotheses. The linear step up test φ^{LSU} (often called Benjamini and Hochberg procedure vf^{BH}) rejects exactly the hypotheses $H_{(1)}, ..., H_{(k)}$ where

$$k = max(i \in I : p_{(i)} \leq \frac{i\alpha}{m})$$

If the maximum does not exists, then no hypothesis is rejected

FDR control

Theorem

Consider the following assumption:

D1 $\forall \theta \in \Theta : \forall j \in I : \forall i \in I_0 : \mathbb{P}_\theta(R \geq j | p_i \leq t)$ is non-increasing in $t \in]0; \alpha]$

D2
- p-values are stochastically larger than $\mathcal{U}([0,1])$
- $\forall \theta \in \Theta$, the p-values $(p_i : i \in I_0)$ are iid
- $\forall \theta \in \Theta$, the random vectors $(p_i; i \in I_0)$ and $(p_i; i \in I_1)$ are stochastically independent

Then,

- under D1: $FDR_\theta(\varphi^{BH}) \leq \frac{m_0}{m}\alpha$
- under D2: $FDR_\theta(\varphi^{BH}) = \frac{m_0}{m}\alpha$

FDR control

Remark

The adaptive procedure (adaptive to the proportion of true null hypotheses) is given by replacing α by $\frac{\alpha}{\hat{\pi}_0}$ where $\hat{\pi}_0$ is an estimation of the proportion of true null hypotheses $\frac{m_0}{m}$

FDR control

Benjamini and Yekutieli procedure

For any dependency structure among $P_1, ..., P_m$, it holds:

$$\forall \theta \in \Theta : FDR_\theta(\varphi^{BH}) \leq \frac{m_0}{m}\alpha \sum_{j=1}^{m} \frac{1}{j}$$

The BH procedure with α replaced by $\frac{\alpha}{\sum_{j=1}^{m}}$ controls the FDR under arbitrary dependency among $P_1, ..., P_m$. This procedure is denoted φ^{BY}

Empirical Bayes procedures

Two-components mixture model

$$F(t) = \mathbb{P}(H = 0)F_0(t) + (1 - \mathbb{P}(H = 0))F_1(t)$$

Empirical Bayes procedures

pFDR as a posterior probability

Under the two-components mixture model:

$$pFDR = \mathbb{P}(H = 0 | T \in \Gamma)$$

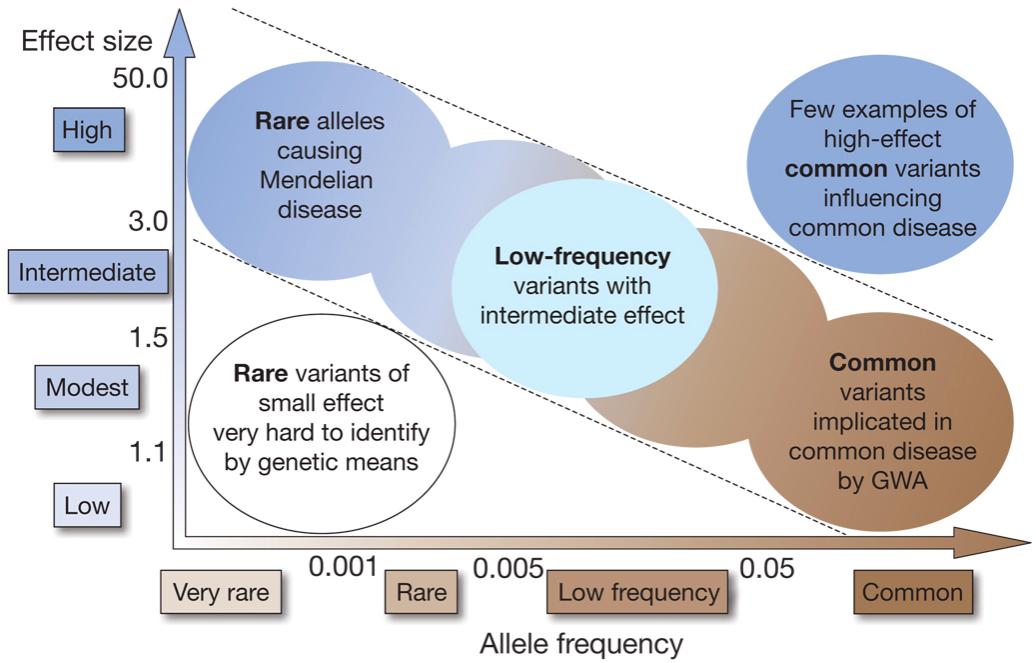
Empirical Bayes procedures

local false discovery rate

The local false discovery rate is a local version of the *pFDR*:

$$lfdr(t) = \mathbb{P}(H = 0 | T = t)$$

Power of association studies



Manolio et al. Finding the missing heritability of complex diseases. Nature. 2009.

Heritability

Quantitative trait

Quantitative genetic model from Ronald Fisher (1918):

$$P = \mu + G + E$$

where

- G is the total genome effect
- E is the environment effect

Heritability

Quantitative trait

If G and E are independent:

$$\sigma^2_P = \sigma^2_G + \sigma^2_E$$

where

- Heritability definition: Proportion of trait variance which is due to all genetic effects

$$H^2 = \frac{\sigma^2_G}{\sigma^2_P}$$

Missing heritability

Quantitative trait

NEWS FEATURE PERSONAL GENOMES NATURE|Vol 456|6 November 2008

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Missing heritability

Missing heritability

Significant GWAS SNPs explain a small proportion of disease heritability

Possible reasons

- GxG and GxE interactions
- A large number of causal variants, each with a small effect
- Epigenetics
- Rare variants

Association studies

GWAS

Captures nearly all common variants

Sequencing (NGS)

Captures all common and rare variants

Genome sequencing

- Whole Genome Sequencing (WGS) -> sequencing of the entire genome
- Whole exome sequencing (WES) -> Sequencing only the coding regions of the genome (1% of tge genome contain 85% of variability)

Genome sequencing allows to capture rare and common variations

SNP arrays vs. Sequencing

SNP arrays

- Data easily stored and analyzed
- Allele calling is standardized
- Experiment well understood
- Number of statistical tests known and carefully considered
- SNP interrogated directly and indirectly

Sequencing

- Requires massive storage capacity
- Allele and Structural Variation calling still in flux
- Experiment not clearly defined
- SNPs interogated at different depths
- Different error rates for different NGS platforms

Gene and pathway level analysis

Limitations of SNP level analyses

- Lack of power (multiple testing problem)
- Causal SNP in LD with multiple types SNPs
- Most common diseases are multifactorial
- Lack of reproducibility
- Biological interpretation

Gene and pathway level analysis

Multi-SNP analyses

- Idea: group SNPs to form SNP sets and test them as a unit
- SNP sets :
  - Genes
  - Pathways
  - Evolutionary conserved regions
  - Moving windows
  - Any group based on an outcome variable
- Databases : Ingenuity, MetaCore, Kegg, Gene ontology (GO), ...
- Use information on network structures

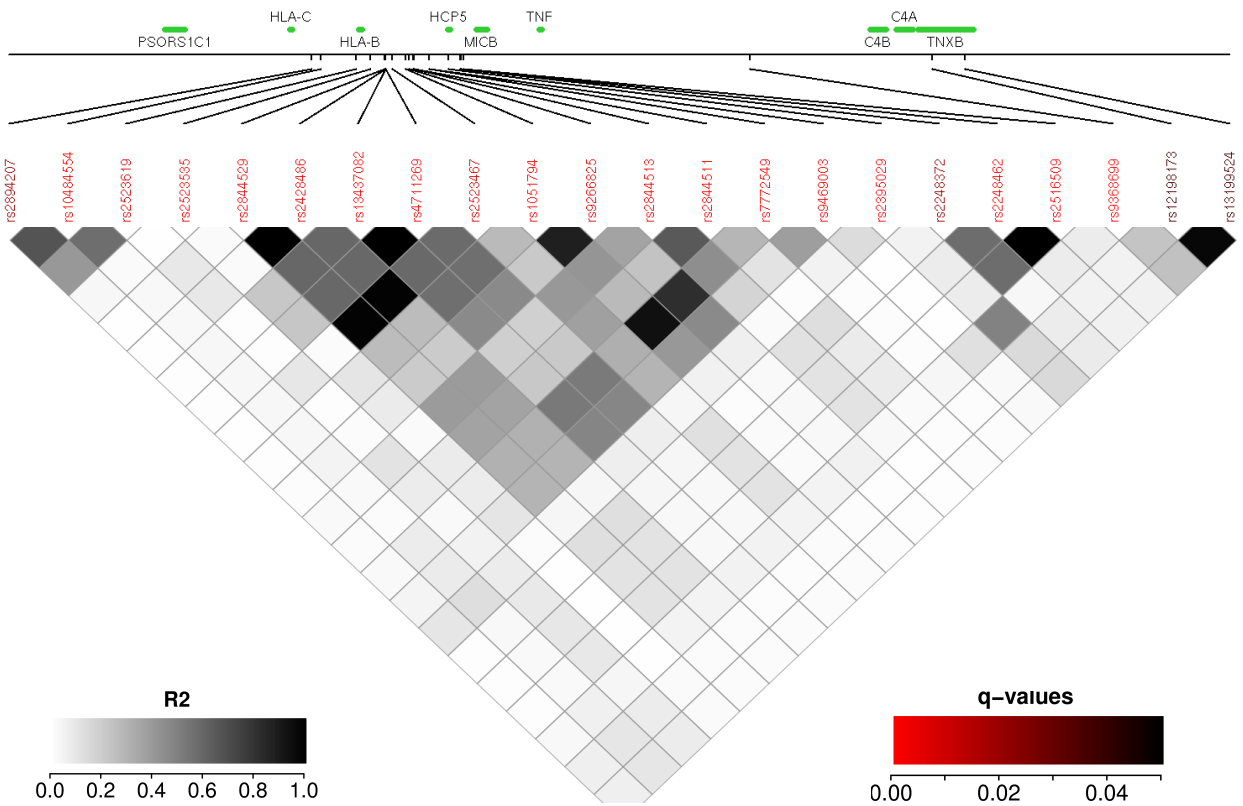
Gene and pathway level analysis

Advantages of multi-SNP analyses

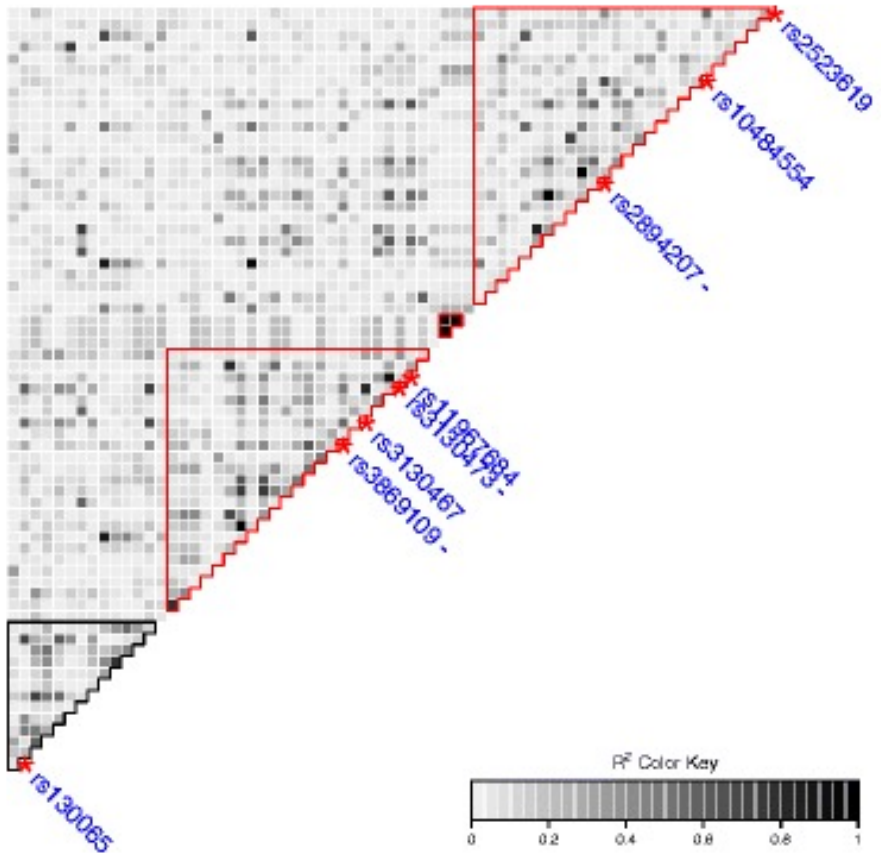
- Dimentionality reduction
- Capture multi-SNP effects
- Biologically meaningful unit

Gene and pathway level analysis

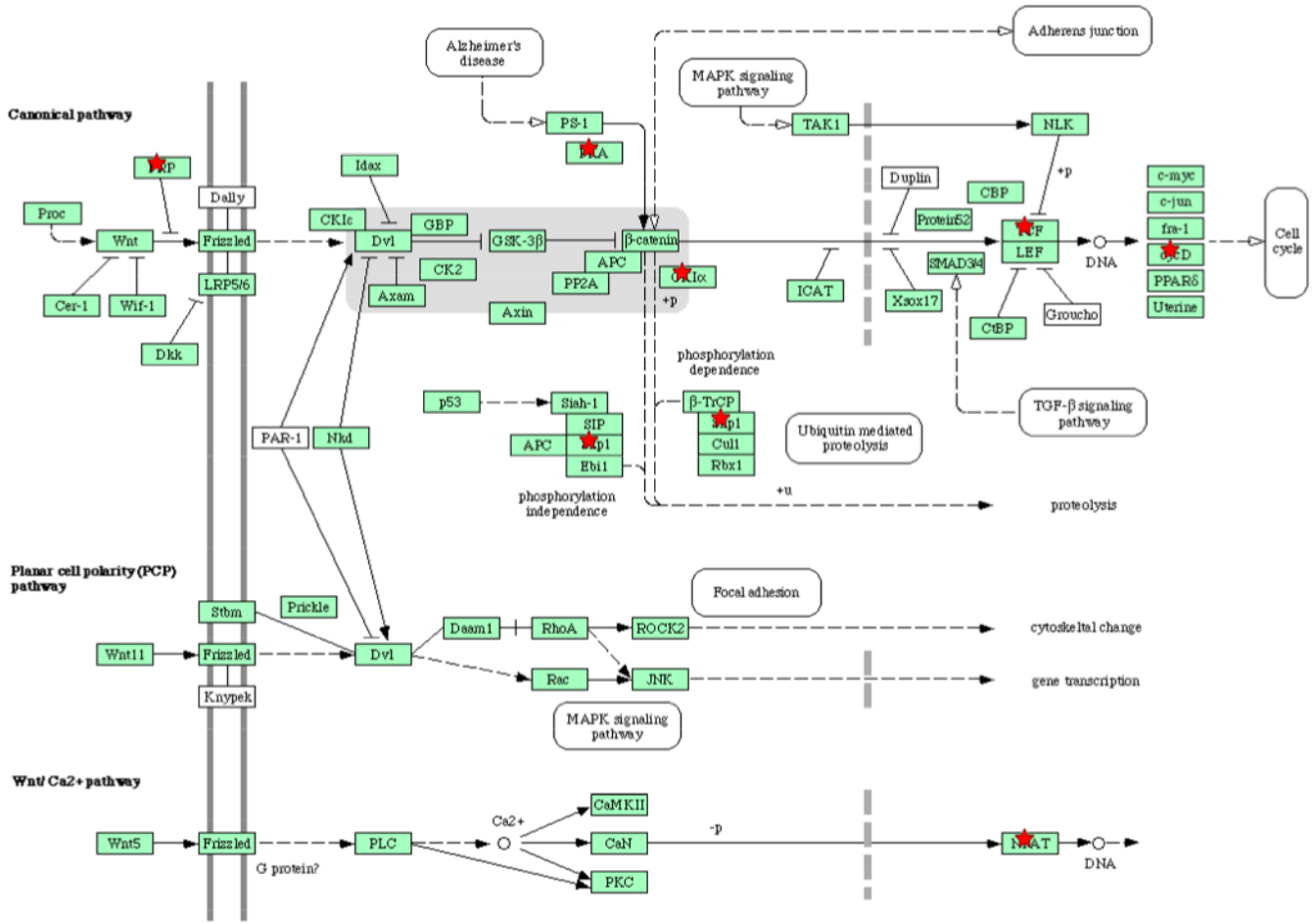
Linkage Disequilibrium (LD) - correlation structure



Example - LD block



Example - pathway



Gene and pathway level analysis

Question

How to test if the gene/pathway is associated with the phenotype?

Gene and pathway level analysis

Statistical methods

- Gene level analysis
  - Minimum p-value tests (minP)
  - Combined p-value approaches
  - Average/collapsing tests
  - Variance component tests
- Pahway level analysis
  - Over-representation analysis (ORA)
  - Gene set enrichment analysis (GSEA)
  - minP, collapsing, combined p-value, VC tests
  - Graphical methods

$\Rightarrow$ See rare variants analysis

Gene and pathway level analysis

Minimum p-value

- Idea: the smallest individual SNP p-value represents the entire group
- Advantage: easy to run
- Problem: How taking into account for having taken the smallest p-value? (Bonferroni, estimation of the effective number of tests,permutations,...)

Gene and pathway level analysis

Combined p-value approaches

- Idea: combine the p-values across the SNPs in the group
- Example: Fisher's method ($X^2_{2k} = -2\sum_{i=1}^{k} ln(p_i)$)
- Problem: p-values are supposed independent for most combination approaches

Gene and pathway level analysis

Averaging/Collapsing

- Idea: build a meta-SNP $C_i = \sum_{i=1}^{k} \omega_j x_{ij}$ and test association between C_i and the outcome
- Common approaches:
  - Simple average
  - Inverse of MAF
  - p-values from previous studies
  - PCA
  - Supervised approaches

Gene and pathway level analysis

Variance component tests

- Regression model:

$$\mathbb{E}(g(y_i)) = \alpha Z_i + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

- Null hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_p$
- Mixed model: if $\mathbb{E}(\beta) = 0$ and $V(\beta) = \tau^2$, then

$$H_0 : \tau^2 = 0$$

Gene and pathway level analysis

Over-representation analysis (ORA)

- Idea: From a list of significant SNPs, look for an over-representation of the SNPs in the group
- Common approaches:
  - Fisher's exact test / Hypergeometric test

	Significant	Not significant	
In group	N_{11}	N_{12}	N_{1+}
Not in group	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	N

  - χ^2 independance test
  - Binomial test

Gene and pathway level analysis

Gene Set Enrichment Analysis (GSEA)

1. Rank all SNPs based on their p-values
2. Calculate an enrichment score for the group G:

$$ES(G) = max_{1 \leq j \leq N} \sum_{i=1}^{j} X_i$$

where $X_i = \begin{cases} \sqrt{\frac{x_{1+}}{x_{s+}}} & \text{if } SNP_i \in G \\ -\sqrt{\frac{x_{s+}}{x_{1+}}} & \text{if } SNP_i \notin G \end{cases}$

3. Evaluate significance based on permutations

Rare variants

- No consensual threshold
- Most of human variants are rare
- Functional variants tend to be rare

Rare variants

Challenges

- Lots of rare variant $\Rightarrow$ Large multiple testing problem
- Large sample size required to oberve one particular rare variant
- Individual power depends on allele frequency

Current strategy

Region based approach

Test the joint effect of pre-specified group of sequence variants

- Sequencing study unit: region (gene, moving window, exons, ...)
- Types of tests
  - Collapsing/burden tests
  - Variance component based tests
  - Omnibus tests

Collapsing tests

Principle

Aggregate rare variant information in a region into a single summary measure

- CAST
- MZ
- Weighted Sum Tests
- ...

Collapsing tests

Multiple linear regression model

- Regression model:

$$\mathbb{E}(g(y_i)) = \alpha Z_i + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

- Null hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_p$

Collapsing tests

Model

Assume: $\beta_1 = \beta_2 = ... = \beta_p = \beta$

$$\mathbb{E}(g(y_i)) = \alpha Z_i + \beta C_i$$

where $C_i = \sum X_{ij}$

Collapsing tests

Other possibilities

- CAST: $C_i = 1_{(\sum x_{ij} > 0)}$
- MZ: $C_i = \sum 1_{(x_{ij} > 0)}$ (dominant model)
- Weigted burden test: $C_i = \sum \omega_j X_{ij}$
  - Unsupervised approaches
  - Supervised approaches (require permutation or bootstraping for significance)

Warning

Loss of power if:

- both protective and deleterious effects
- only a few variants have an effect

Sequence Kernel Association Test (SKAT)

Principle

- Compare pair-wise similarity in phenotype between subjects to pair-wise similarity in genotypes at the rare variants
- Similarity in genotypes is measured with a kernel $K(G_i, G_{i'})$

Sequence Kernel Association Test (SKAT)

Variance component tests

- Regression model:

$$\mathbb{E}(g(y_i)) = \alpha Z_i + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

- Null hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_p$
- Mixed model: if $\mathbb{E}(\beta) = 0$ and $V(\beta) = \tau^2$, then

$$H_0 : \tau^2 = 0$$

Sequence Kernel Association Test (SKAT)

Variance component tests

- Regression model:

$$\mathbb{E}(g(y_i)) = \alpha Z_i + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

- Null hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_p$
- Mixed model: if $\mathbb{E}(\beta) = 0$ and $V(\beta) = \omega_j \tau^2$, then

$$H_0 : \tau^2 = 0$$

- Score test statistic: $Q_{skat} = (y - \mu_0)' K (y - \mu_0)$ where

$$K = GWWG'$$

with $W = diag(\omega_j)$

SKAT-O

Optimal unified strategy

$$Q_{optimal} = \rho Q_{collapse} + (1 - \rho) Q_{SKAT}$$

Principle

Use data to adaptively estimate ρ in order to maximize power

Additional concerns

- Quality controls
- Population stratification
- Accomodating common variants

Epistasis

Definitions

- Biological definition
- Statistical definition

Epistasis

Biological definition

- Originally in mendelian genetics
- One locus masks the effect of another locus
- No environmental contribution

Epistasis

Table 1. Example of phenotypes (e.g. hair colour) obtained from different genotypes at two loci interacting epistatically, under Bateson's (1909) definition of epistasis

	Genotype at locus G		
Genotype at locus B	*g/g*	*g/G*	*G/G*
b/b	White	Grey	Grey
b/B	Black	Grey	Grey
B/B	Black	Grey	Grey

Cordell H, *Human Molecular Genetics*, 2002

Epistasis

Biological definition

"situation in which the qualitative nature of the mechanism of action of a factor is affected by the presence or absence of the other" (Siemiatycki, 1981)

Table 2. Example of penetrance table for two loci interacting epistatically in a general sense

	Genotype at locus B		
Genotype at locus A	*b/b*	*b/B*	*B/B*
a/a	0	0	0
a/A	0	1	1
A/A	0	1	1

Cordell H, *Human Molecular Genetics*, 2002

Epistasis

Statistical definition

$$g(\mathbb{E}(y)) = \sum_i \beta_i SNP_i + \sum_{i \neq j} \gamma_{ij} SNP_i x SNP_j$$

Detection

- Exhaustive combination
- Main effects filter
- Biological based filter

Epistasis

Exhaustive combination

- Pro: Allows to investigate all possible interactions
- Cons: Huge number of combinations (practically unfeasible)

Main effects filter

- Pro: Easy to interpretand computationally feasible
- Cons: Evaluates only genes with large main effects

Biological based filter

- Pro: Fewer statistical tests
- Cons: Limited by current state of knowledge

Epistasis

Modelling strategies

- Standard regression analyses

$$g(\mathbb{E}(y)) = \sum_i \beta_i SNP_i + \sum_{i \neq j} \gamma_{ij} SNP_i x SNP_j$$

- -epistasis option in PLINK

Epistasis

Gene level analyses

Advantages:

- Results biologically interpretable
- Genetic effects more detectable
- Reduction of the variables number

Epistasis

Gene level analyses

Existing gene scale methods :

- For two or few genes
  - → PCA + logistic regression *(He et al. 2011, Li et al. 2009, Zhang et al. 2008)*
  - → PLS + logistic regression *(Wang T et al. 2009)*
- For a larger number of genes
  - → PCA + LASSO *(D'Angelo et al. 2009)*
  - → PCA + pathway-guided penalized regression *(Wang X et al. 2014)*

From Stanislas V, 2016

Epistasis

Group modelling approach

	$SNP_{1,1}$	..	SNP_{1,p_1}	..	$SNP_{G,1}$	..	SNP_{G,p_G}	Pheno
Ind_1	1		0		0		1	y_1
Ind_2	0		0		2		1	y_2
.	2		1		1		2	.
.	0		1		0		0	.
Ind_i	0		2		1		0	y_i
	$\underbrace{\quad}_{gene_1}$				$\underbrace{\quad}_{gene_G}$			

We note
$SNP_{1,1} = X_{1,1}$

r, s two genes

model :

$$g(E[\boldsymbol{y}|\boldsymbol{X}]) = \underbrace{\sum_{g}\sum_{p_g}\beta_{g,p_g}\boldsymbol{X}_{g,p_g}}_{\text{Main effects}} + \underbrace{\sum_{r,s}\gamma_{r,s}\boldsymbol{Z}_{r,s}}_{\text{Interaction effects}}$$

$$\boldsymbol{\beta} = \left(\underbrace{\beta_{1,1},\beta_{1,2},\cdots,\beta_{1,p_1}}_{gene_1},\cdots,\underbrace{\beta_{G,1},\cdots,\beta_{G,p_G}}_{gene_G}\right)^T \qquad \gamma = \left(\gamma_{12},\cdots \underbrace{\gamma_{1G}}_{\gamma_{1G,1},\cdots,\gamma_{1G,q}} \cdots,\gamma_{(G-1)G}\right)$$

q : # of interaction variables for a couple

From Stanislas V, 2016

Epistasis

Interaction variable construction

We consider $f_u(\boldsymbol{X}^r, \boldsymbol{X}^s)$ to represent the interaction between genes r, s.

$$\hat{\boldsymbol{u}} = arg \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} cov^2(\boldsymbol{y}, f_u(\boldsymbol{X}^r, \boldsymbol{X}^s))$$

We set : $f_u(\boldsymbol{X}^r, \boldsymbol{X}^s) = \boldsymbol{W}^{rs}\boldsymbol{u}$ with $\boldsymbol{W}^{rs} = \{X^r_{ij} X^s_{ik}\}_{i=1\cdots n}^{j=1\cdots,p_r;k=1,\cdots,p_s}$

$$\max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} ||\hat{cov}[\boldsymbol{W}^{rs}\boldsymbol{u}, \boldsymbol{y}]||^2 = \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} \boldsymbol{u}^T \boldsymbol{W}^{rs\,T} \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{W}^{rs}\boldsymbol{u}$$

$\boldsymbol{u}$: eigen vector associated to the largest eigenvalue of $\boldsymbol{W}^{rs\,T}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{W}^{rs}$

$$\boldsymbol{u} = \boldsymbol{W}^{rs\,T}\boldsymbol{y}$$

We then obtain for each couple $(r, s) \rightarrow \boldsymbol{Z}^{rs} = \boldsymbol{W}^{rs}\boldsymbol{u} = \boldsymbol{W}^{rs}\boldsymbol{W}^{rs\,T}\boldsymbol{y}$

From Stanislas V, 2016

Epistasis

Coefficients estimation

Group LASSO regression

$$\hat{\theta} = (\hat{\beta}, \hat{\gamma}) = \underset{\beta,\gamma}{\operatorname{argmin}} \left(\sum_i (y_i - \boldsymbol{X}_i\beta - \boldsymbol{Z}_i\gamma)^2 + \lambda \left[\sum_g \sqrt{p_g} ||\beta^g||_2 + \sum_{rs} \sqrt{p_r p_s} ||\gamma^{rs}||_2 \right] \right)$$

Limits of the groupLASSO regression :

- $P(S^* \subset \hat{S}) \underset{n \to +\infty}{\longrightarrow} 1$ but $|\hat{S}| >> |S^*|$
- Difficult to compute p-value or confidence interval

Adaptive-Ridge Cleaning *Becu JM, 2015*

- Use of a specific penalty for group LASSO
- Permutation test based on Fisher test approach for each group $P_k = \frac{1}{B}\#\{F_k^* \geq F_k\}$

From Stanislas V, 2016

Epistasis

Non prametric approaches

- Decision trees
- Multifactor-dimentionality reduction (MDR)
- Support vector machines