Statistics for genomic data science in health

Cyril Dalmasso
cyril.dalmasso@univ-evry.fr

Université d'Evry Val d'Essonne

2019-2020

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

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

Statistics and genetics/genomics

Historical perspective

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

Genomics

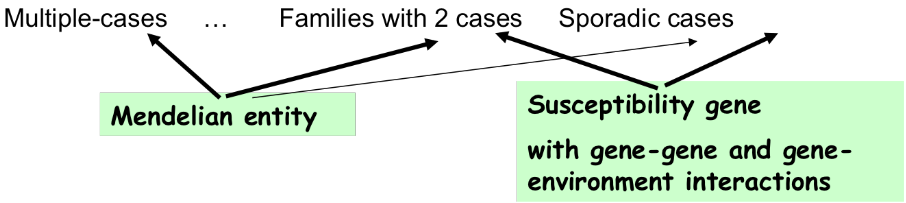
Genomics is the study of genomes

Genetic factors in a medical context

Monogenic diseases

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

Multifactorial diseases



Multiple-cases … Families with 2 cases Sporadic cases

Mendelian entity

Susceptibility gene

with gene-gene and gene-environment interactions

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 ?
  → **linkage analyses** (family based)
  powerful in gene identification of mendelian diseases
- Which are the susceptibility genes?
  → **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

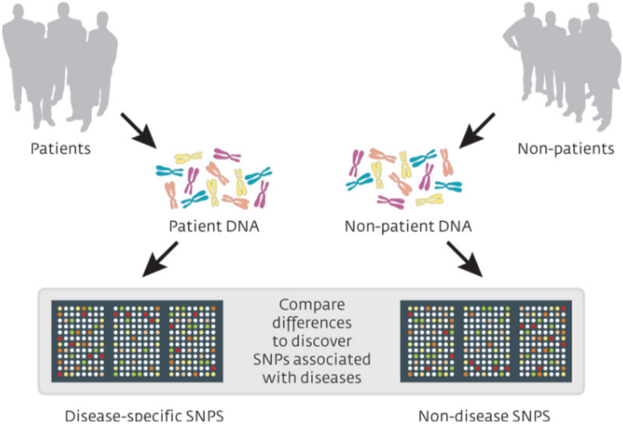
Remark : Association with polymorphisms that are not themselves causal risk factors can be used to localize the trait gene

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, ...)

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

1. Introduction to GWAS/WGS/WES
2. Data structure
   - Single Nucleotide Polymorphism
   - Technologies
   - Preprocessing
3. Single-marker analyses
4. Multi-marker analyses

Genetic markers

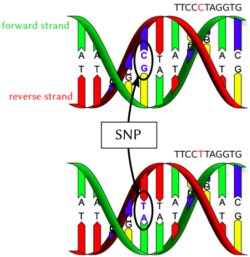
Definition

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

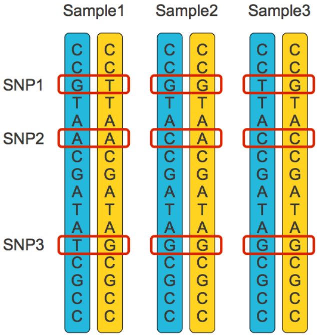
Single nucleotide polymorphism (SNP)

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)


Sample1
Sample2
Sample3
SNP1
SNP2
SNP3
C C G T A A C G A T A T C G C C
C C T T A A C G A T A G C G C C
C C G T A C C G A T A G C G C C
C C G T A A C G A T A G C G C C
C C T T A C C G A T A G C G C C
C C G T A C C G A T A G C G C C

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

$_{MarkerA}\cdot^{MarkerB}$	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)

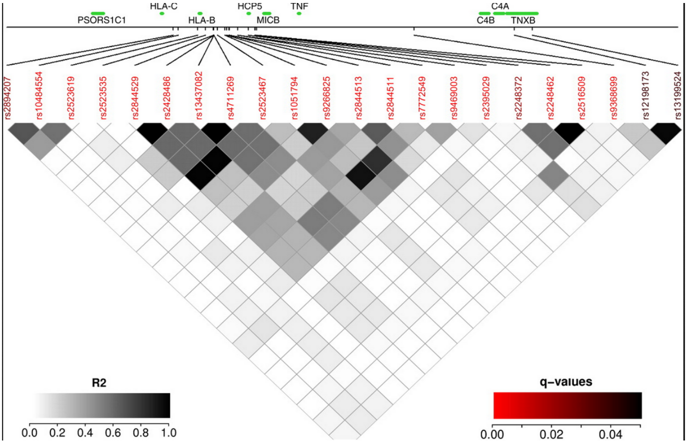
Linkage disequilibrium

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

1. Introduction to GWAS/WGS/WES

2. Data structure
   - Single Nucleotide Polymorphism
   - Technologies
   - Preprocessing

3. Single-marker analyses

4. Multi-marker analyses

Genomic technologies

Technologies

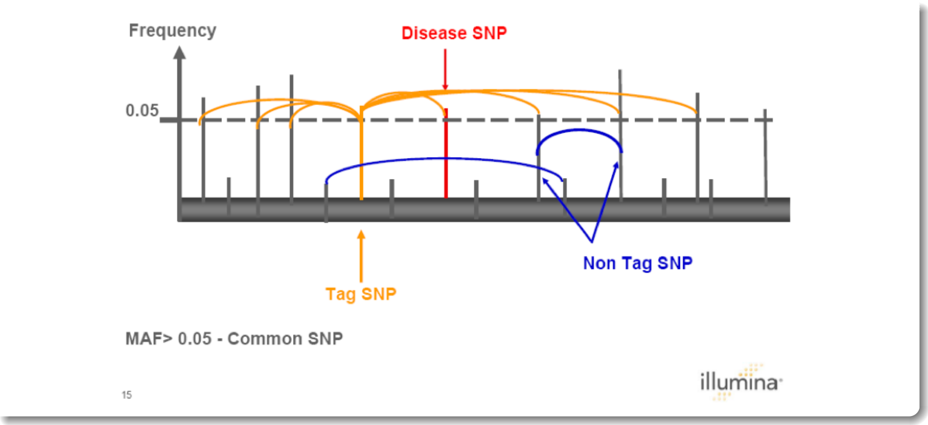
- Whole genome sequencing (WGS)
- Whole exome sequencing (WES)
- SNP genotyping (microarrays)

Microarrays

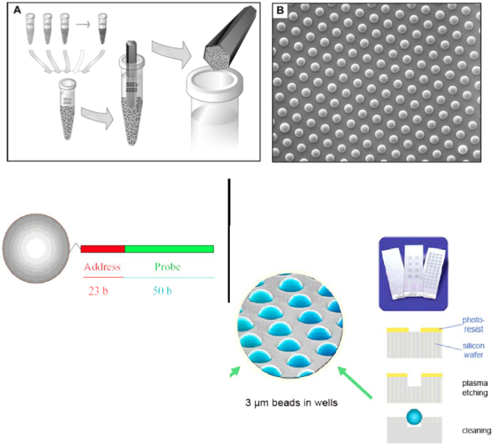
Key concept for GWAS

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


Frequency
Disease SNP
0.05
Non Tag SNP
Tag SNP
MAF> 0.05 - Common SNP
illumina
15

Illumina SNP arrays



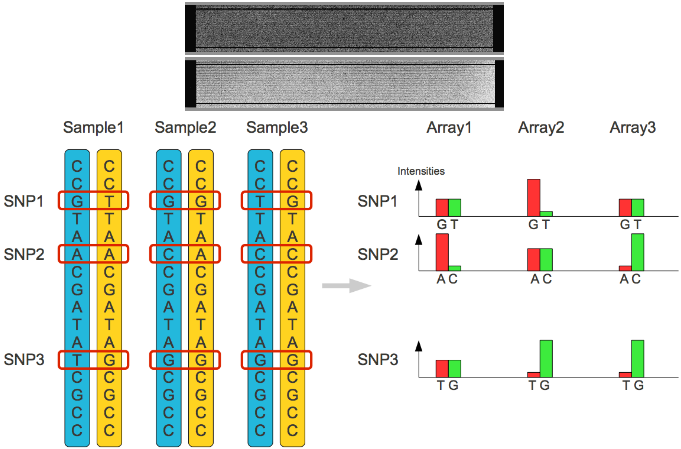
Oliphant et al. Biotechniques. 2002.

Affymetrix SNP arrays 6.0

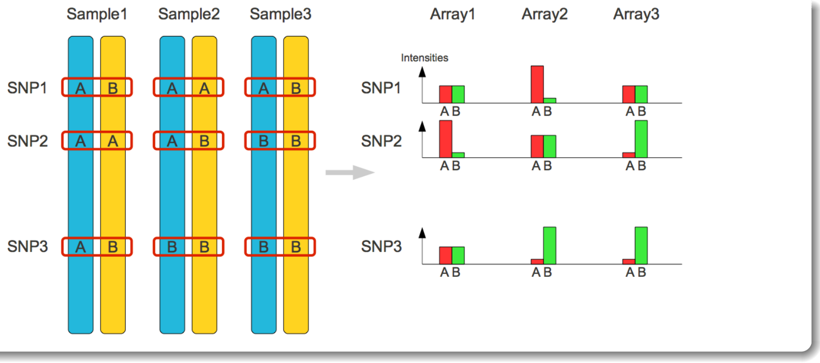
Affymetrix SNP Array 6.0

- 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

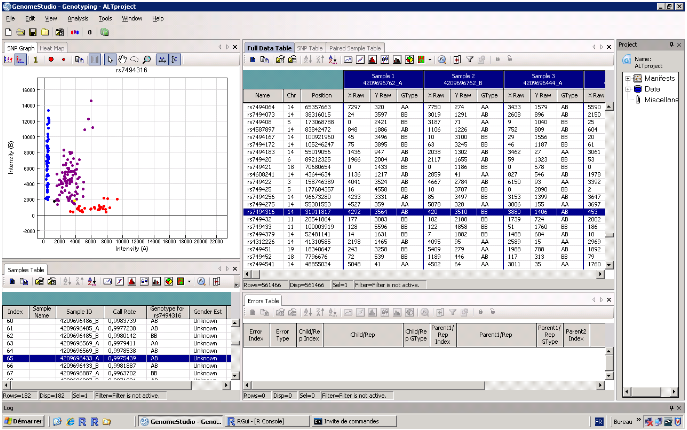
Intensity values for both alleles



Intensity values for both alleles



Genome Studio (Illumina)



			Sample 1 4209696762_A			Sample 2 4209696762_B			Sample 3 4209696444_A			
Name	Chr	Position	X Raw	Y Raw	GType	X Raw	Y Raw	GType	X Raw	Y Raw	GType	X Raw
rs7494064	14	65357663	7297	320	AA	7750	274	AA	3433	1579	AB	5590
rs7494073	14	38316015	24	3597	BB	3019	1291	AB	2608	896	AB	2150
rs749408	5	173068788	0	2421	BB	3187	71	AA	9	1040	BB	25
rs4587897	14	83842472	848	1886	AB	1106	1226	AB	752	809	AB	604
rs7494167	14	100921960	45	3496	BB	10	3100	BB	29	1556	BB	20
rs7494172	14	105246247	75	3895	BB	63	3245	BB	46	1187	BB	61
rs7494183	14	55019056	1436	947	AB	2038	1302	AB	3462	27	AA	3061
rs749420	6	89212325	1966	2004	AB	2117	1655	AB	59	1323	BB	53
rs749421	18	70680654	0	1433	BB	0	1186	BB	0	578	BB	0
rs4608241	14	43644634	1136	1217	AB	2859	41	AA	827	546	AB	1978
rs749422	3	158746389	4041	3524	AB	4667	2784	AB	6150	93	AA	3392
rs749425	5	177684357	16	4558	BB	10	3707	BB	0	2090	BB	2
rs7494256	14	96673280	4233	3331	AB	85	3497	BB	3153	1399	AB	3647
rs7494275	14	55301553	4527	359	AA	5078	328	AA	3006	155	AA	3697
rs7494316	14	31911817	4292	3564	AB	420	3510	BB	3880	1406	AB	453
rs749432	11	20541864	177	3083	BB	102	2188	BB	1739	724	AB	2002
rs749433	11	100003919	128	5596	BB	122	4858	BB	51	1760	BB	186
rs7494379	14	52481141	14	1631	BB	7	1882	BB	1488	604	AB	10
rs4312226	14	41310585	2198	1465	AB	4095	95	AA	2589	15	AA	2969
rs749451	19	18340647	243	3258	BB	5409	279	AA	1988	788	AB	1892
rs749452	18	7796676	72	539	BB	1189	446	AB	117	313	BB	79
rs7494541	14	48855034	5048	41	AA	4502	64	AA	3011	35	AA	1760

Rows=561466 Disp=561466 Sel=1 Filter=Filter is not active.

Index	Sample Name	Sample ID	Call Rate	Genotype for rs7494316	Gender Est
60		4209696486_B	0,9983739	AB	Unknown
61		4209696485_A	0,9977238	AB	Unknown
62		4209696485_B	0,9980142	BB	Unknown
63		4209696569_A	0,9979411	AA	Unknown
64		4209696569_B	0,9978538	AB	Unknown
65		4209696433_A	0,9975439	AB	Unknown
66		4209696433_B	0,9981887	AB	Unknown
67		4209696887_A	0,9963702	BB	Unknown

Rows=182 Disp=182 Sel=1 Filter=Filter is not active.

Error Index	Error Type	Child/Rep Index	Child/Rep	Child/Rep GType	Parent1/Rep Index	Parent1/Rep	Parent1/Rep GType	Parent2 Index

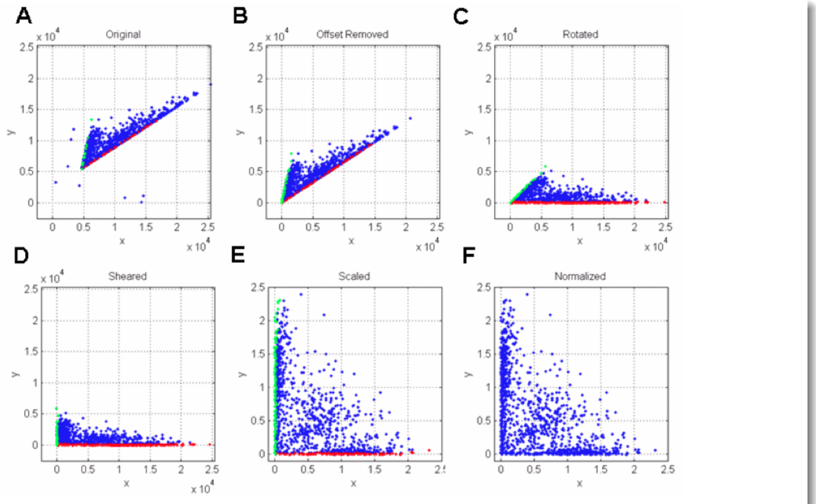
Rows=0 Disp=0 Sel=0 Filter=Filter is not active.

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)



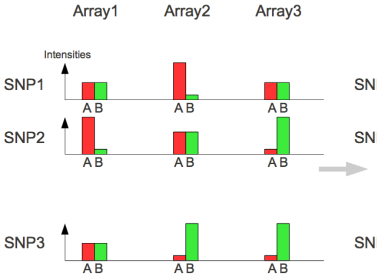
www.illumina.com

Normalization - Example (Illumina)

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

Genotyping

Objective

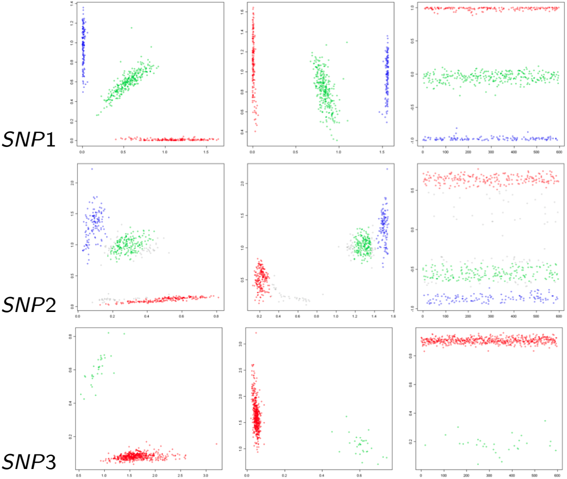


	Array1	Array2	Array3
SNP1	2	1	2
SNP2	1	2	3
	...	...	...
SNP3	2	3	3

1 = AA (homozygous)
2 = AB (heterozygous)
3 = BB (homozygous)

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, ...
- ...

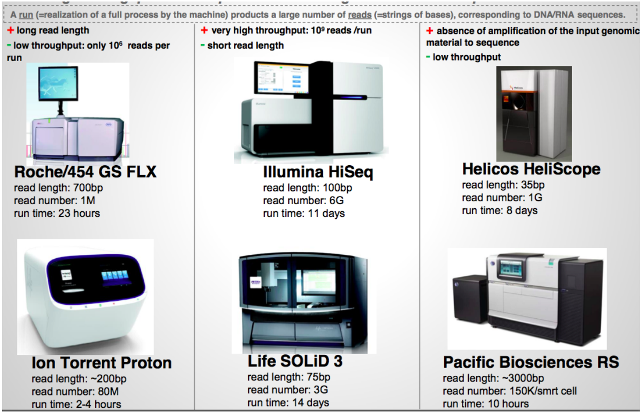
Comparison of genotying algorithms for Illumina's SNP arrays

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 understtod
- 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

1 Introduction to GWAS/WGS/WES

2 Data structure
- Single Nucleotide Polymorphism
- Technologies
- Preprocessing

3 Single-marker analyses

4 Multi-marker analyses

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

- Cretate 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.

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

1 Introduction to GWAS/WGS/WES

2 Data structure

3 Single-marker analyses
- Statistical tests
- Multiple testing
- Population stratification

4 Multi-marker analyses

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:

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

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

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:

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

- 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 :

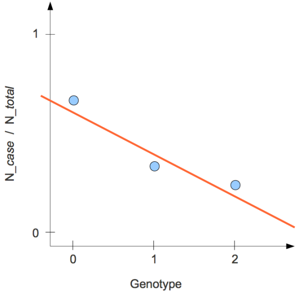
$$H_0 : \pi_0 = \pi_1 = \pi_2 \text{ vs } H_1 : \pi_0 < \pi_1 < \pi_2$$

- Test statistic :

$$\frac{\hat{\beta}}{Var(\hat{\beta})} \xrightarrow[H_0]{} \chi^2_1$$

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:

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

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

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

- 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

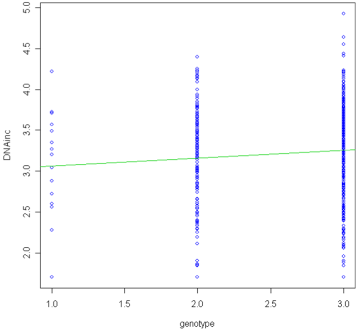
Quantitative Trait Loci (QTL) mapping aim at identifying genetic loci that influence the phenotypic variation of a quantitative trait

Genetic models

- Dominant
- Recessive
- Additive
- Multiplicative

Quantitative trait

Linear regression model

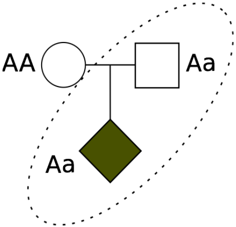


$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

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

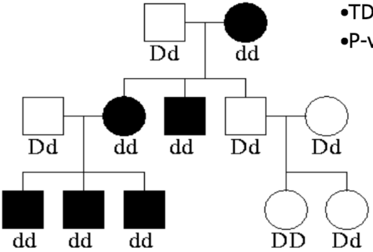
Test statistic:

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

Family based association tests

Transmission Disequilibrium Test (TDT)

Example



- $n_{d|D}=5$
- $n_{D|d}=0$
- TDT-chisq=5
- P-value=.025

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

1 Introduction to GWAS/WGS/WES

2 Data structure

3 Single-marker analyses
- Statistical tests
- Multiple testing
- Population stratification

4 Multi-marker analyses

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 ⇒ laboratory cost
- False negatives ⇒ discovery/publication cost

Multiple testing

Strategy

Reality \ *Decision*	H_0 not rejected	H_0 rejected	Total
H_0 true	TN	FP	m_0
H_0 false	FN	TP	m_1
Total	N	P	m

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

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

Multiple testing

Multiple testing error criteria

Decision *Reality*	H_0 not rejected	H_0 rejected	Total
H_0 true	TN	FP	m_0
H_0 false	FN	TP	m_1
Total	N	P	m

Family-wise error rate: $FWER = \Pr(FP > 1)$

False discovery rate: $FDR = \mathbb{E}(Q)$ with $Q = \begin{cases} \frac{FP}{P} & \text{if } P \neq 0 \\ 0 & \text{if } P = 0 \end{cases}$

Multiple testing

Adjusted p-values

Adjusted p-values extend the p-value concept to the multiple testing framework:

$$p_j^* = inf\{\alpha \in [0,1] | H_{0j} \text{ rejected at threshold } \alpha\}$$

Use: H_0 rejected if $p^* < \alpha$

FWER procedures

Bonferroni

Procedure:	m tests at level $\alpha^* = \alpha/m$
Adjusted p-values:	$p_i^* = m \times p_i$

FWER control

The Bonferroni procedure controls the FWER (strong sense) without any assumption on dependences.

Let $I = \{i|H_i = 0\}$

$$FWER = \Pr(V > 0) = Pr(min_{i\in I}P_i \leq \alpha^*) = Pr(\cup_{i\in I}\{P_i \leq \alpha^*\})$$
$$\leq Pr(\cup_{i=1}^{m}\{P_i \leq \alpha^*\}) \leq \sum_{i=1}^{m} Pr(P_i \leq \alpha^*) = \sum_{i=1}^{m}\alpha^* = \alpha$$

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

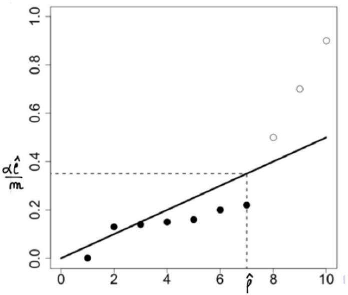
To be used with caution!

FDR procedures

Benjamini Hochberg (BH)

BH is a step-up procedure with $\alpha^*_{(i)} = \frac{\alpha i}{m}$: rejection of the $\hat{k}$ hypotheses with the smallest p-values where

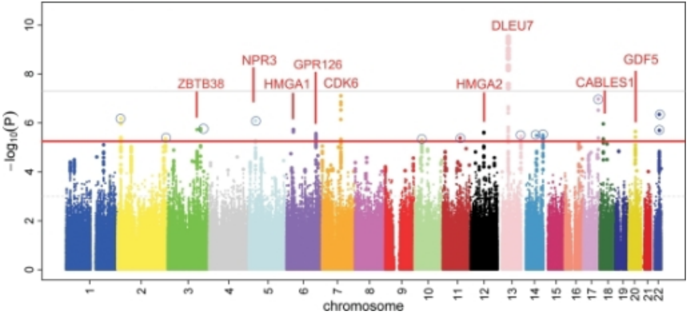
$$\hat{k} = max\{0 \leq k \leq m : p_{(k)} \leq \frac{\alpha k}{m}\}$$



E. Roquain

Results presentation

Manhattan plot



Estrada et al, Hum Mol Genet, 2009 .

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

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

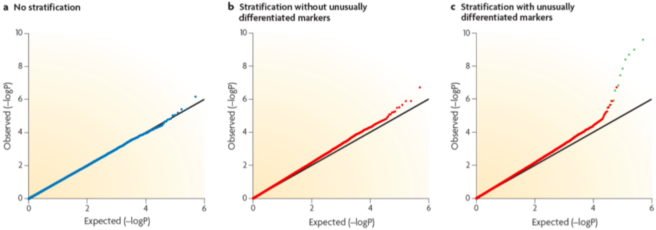


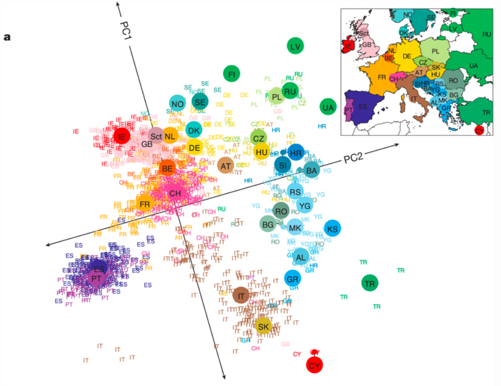
Figure 1 | **P–P plots for the visualization of stratification or other confounders.** The figure shows simulated P–P plots under three scenarios for genome-wide scans with no causal markers. **a** | No stratification: *p*-values fit the expected distribution. **b** | Stratification without unusually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

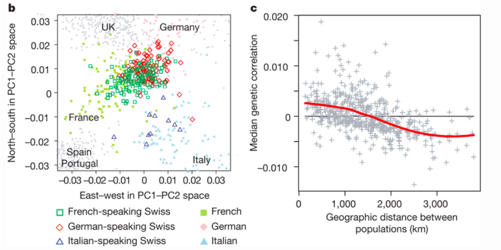


Novembre J et al. Genes mirror geography within Europe. Nature. 2008

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

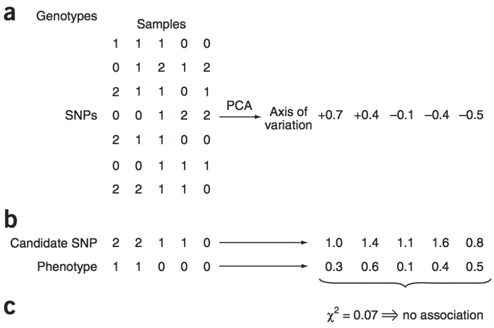
Genomic control

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

Devlin et al., Theor Popul Biol. 2001

Population stratification

Eigenstrat - PCA



a Genotypes

Samples

	1	1	1	0	0							
	0	1	2	1	2							
	2	1	1	0	1							
SNPs	0	0	1	2	2	$\xrightarrow{\text{PCA}}$	Axis of variation	+0.7	+0.4	–0.1	–0.4	–0.5
	2	1	1	0	0							
	0	0	1	1	1							
	2	2	1	1	0							

b

Candidate SNP	2	2	1	1	0	$\longrightarrow$	1.0	1.4	1.1	1.6	0.8
Phenotype	1	1	0	0	0	$\longrightarrow$	0.3	0.6	0.1	0.4	0.5

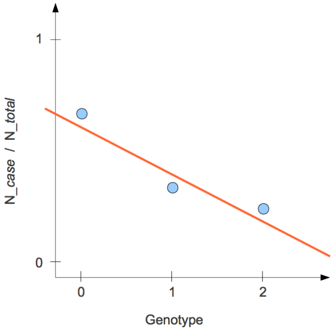
c

$\chi^2 = 0.07 \Longrightarrow$ no association

Price et al. Nature Genetics. 2006.

Population stratification

Eigenstrat - Cochran Armitage trend test



Generalization

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

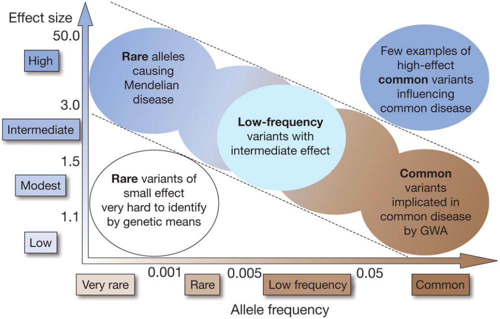
Population stratification

PCA

Warning: not adapted to familial data

1. Introduction to GWAS/WGS/WES
2. Data structure
3. Single-marker analyses
4. Multi-marker analyses
    - Gene and pathway level analysis
    - Methods for combining information from single-marker coefficients

Power of association studies


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Low-frequency variants with intermediate effect
Few examples of high-effect **common** variants influencing common disease
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

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:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

where

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

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2}$$

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 understtod
- 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

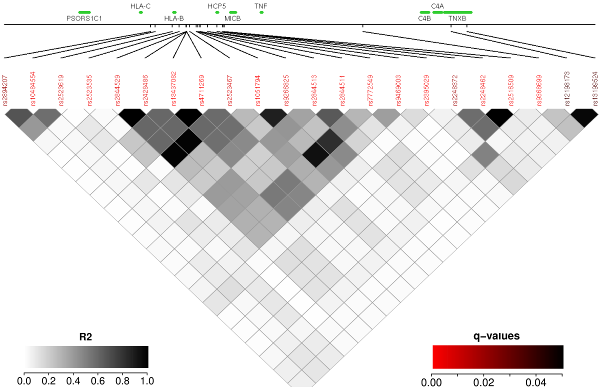
Gene and pathway level analysis

Advantages of multi-SNP analyses

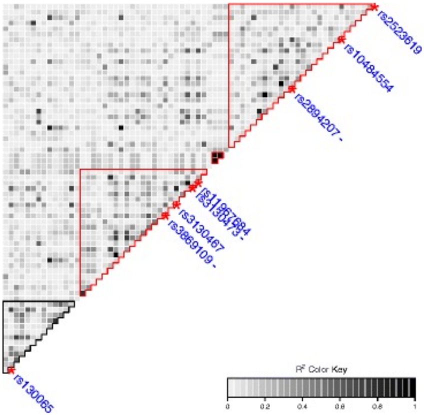
- Dimentionality reduction
- Capture multi-SNP effects
- Biologically meaningful unit

Gene and pathway level analysis

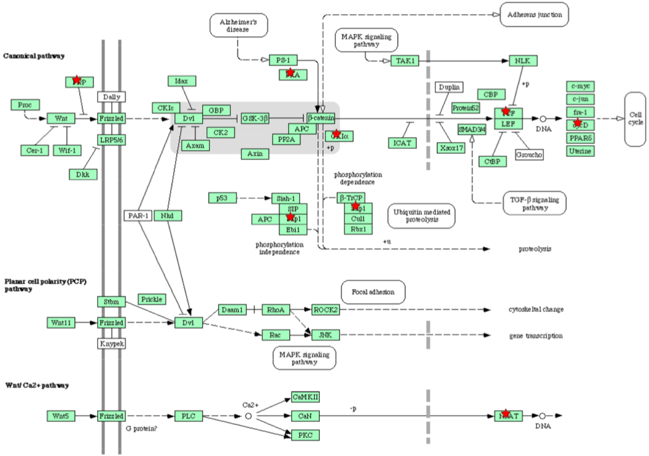
Linkage Disequilibrium (LD) - correlation structure



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

⇒ 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 $(X_{2k}^2 = -2 \sum_{i=1}^{k} ln(p_i))$
- 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 ⇒ 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:

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

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

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