

Genome Wide Association Studies

Part I: Theoretical issues

Ahmed Rebai, Phd

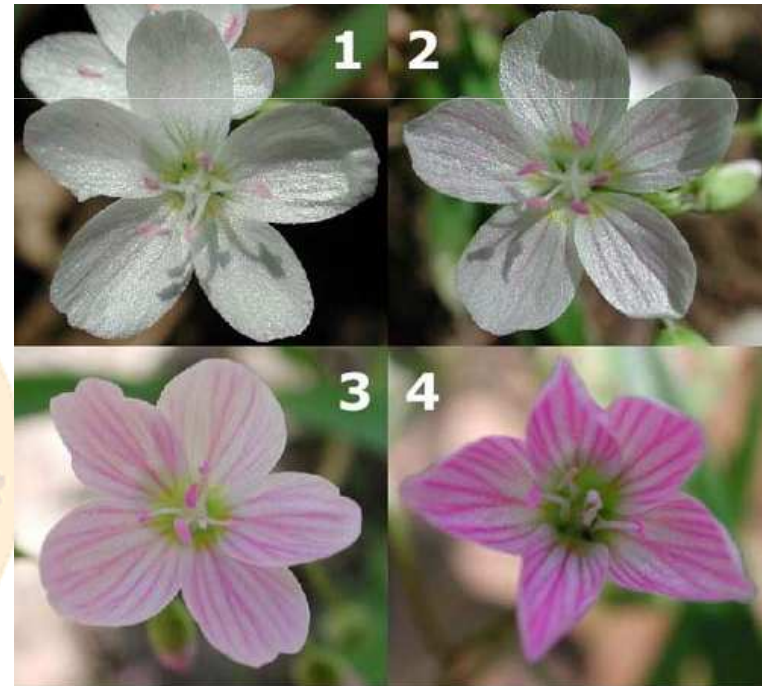
ahmed.rebai@cbs.rnrt.tn

Screening the genome

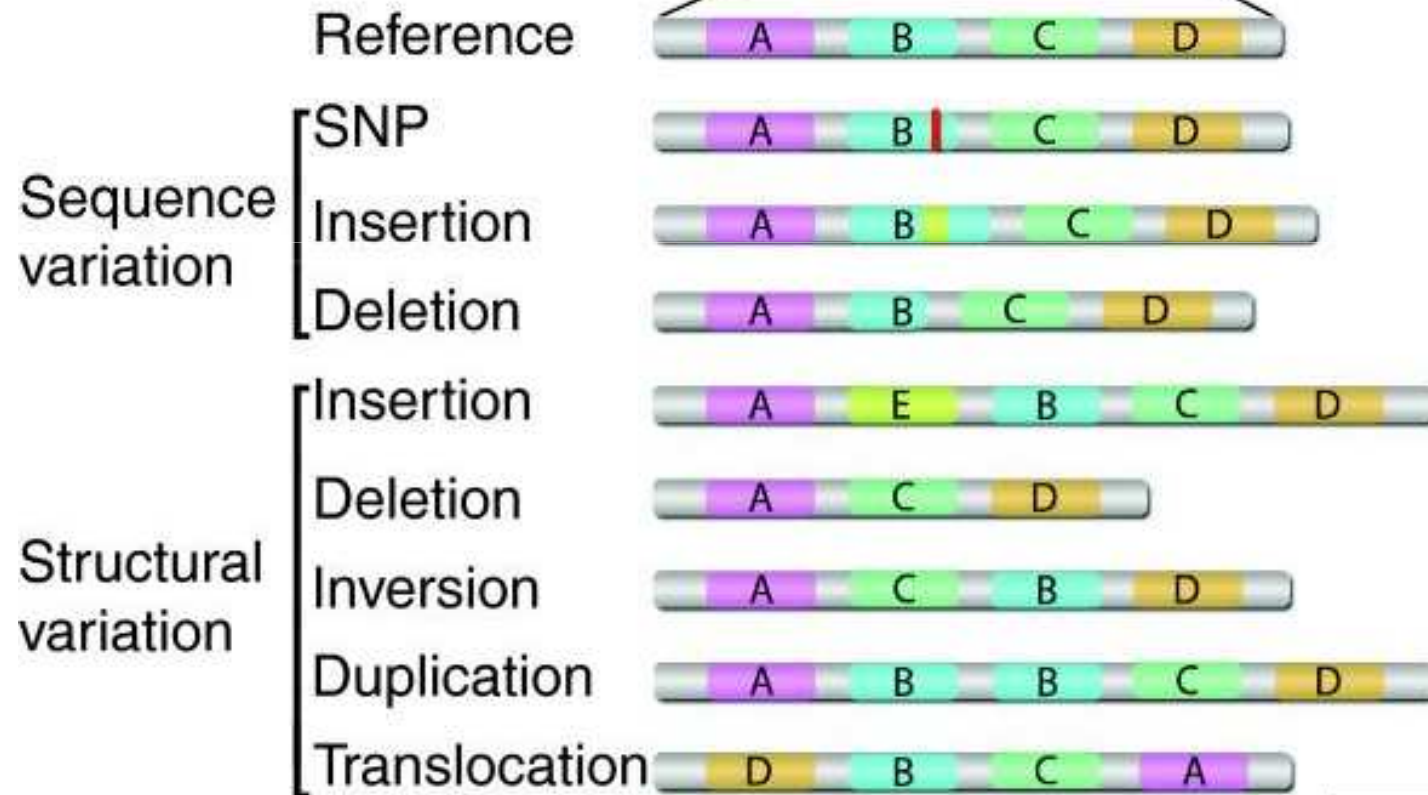
- Human inherited diseases have a genetic basis that needs to be unraveled
- Diseases range from Mendelian (single gene!) to complex (multiple genes, pathways, environment,..)
- Look for DNA sequence changes (single base changes, duplication, deletions,..) that might explain the phenotype spectrum

What is polymorphism?

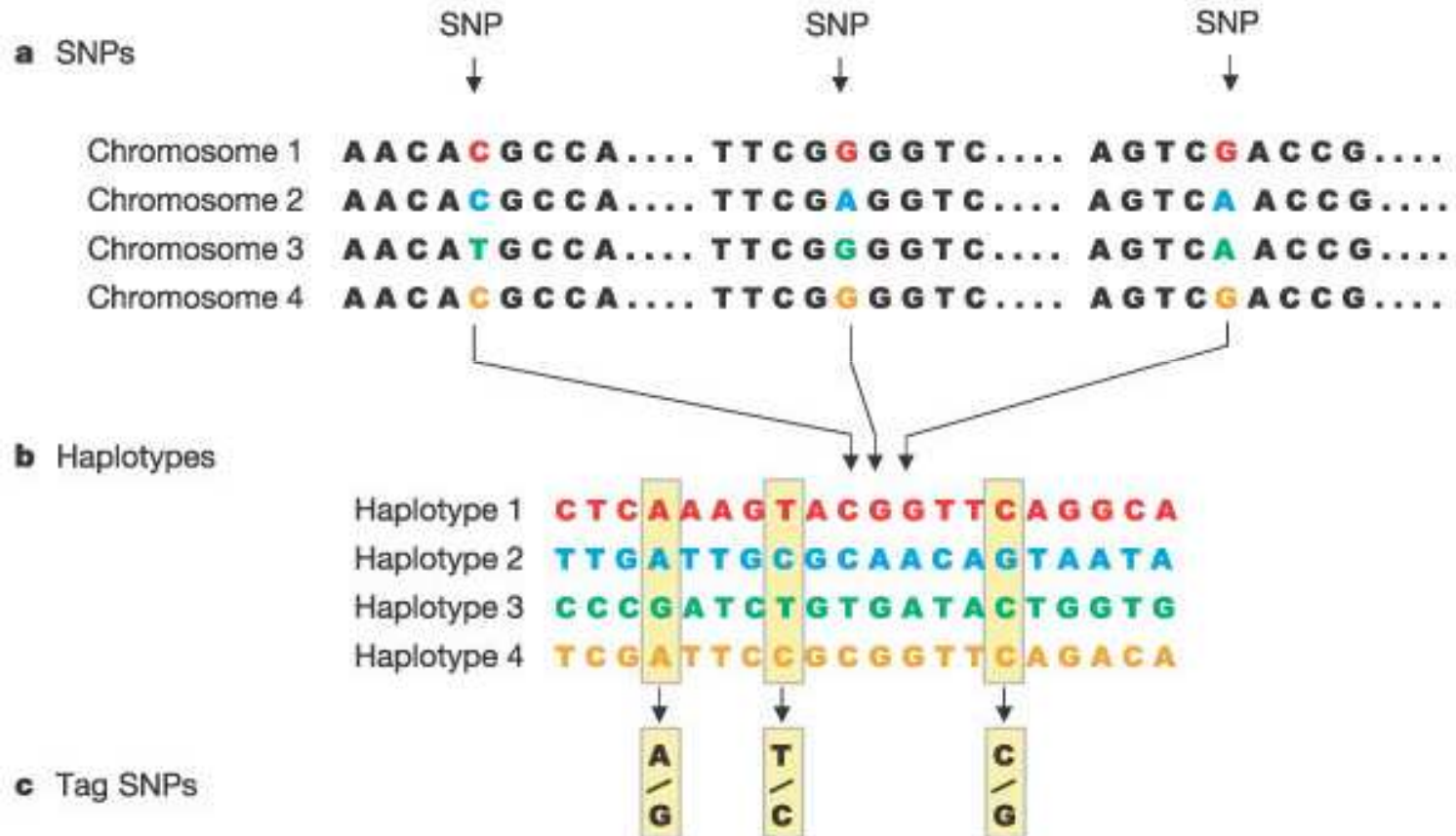
- Anything that differ between individuals, species,..



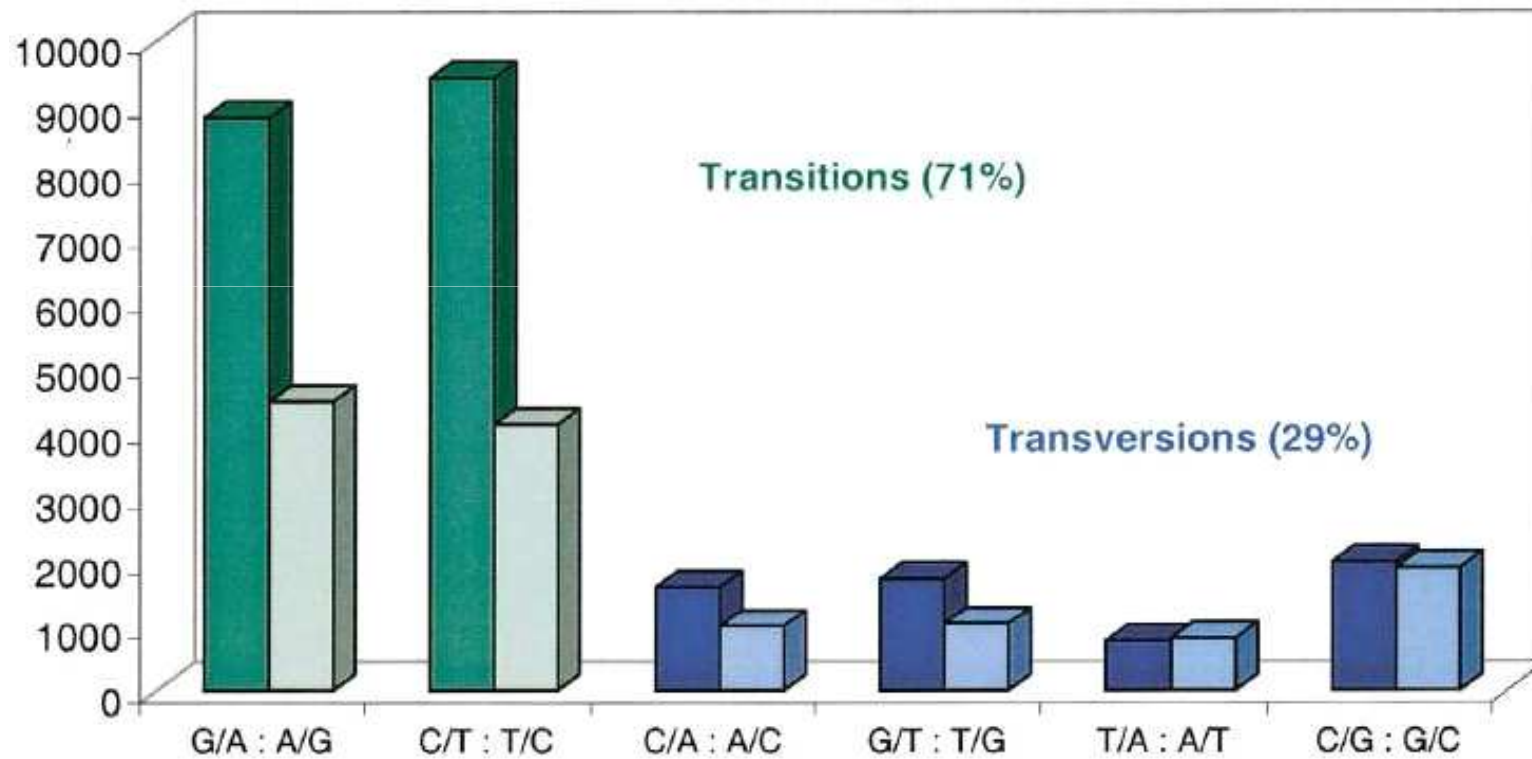
Polymorphic Sequences



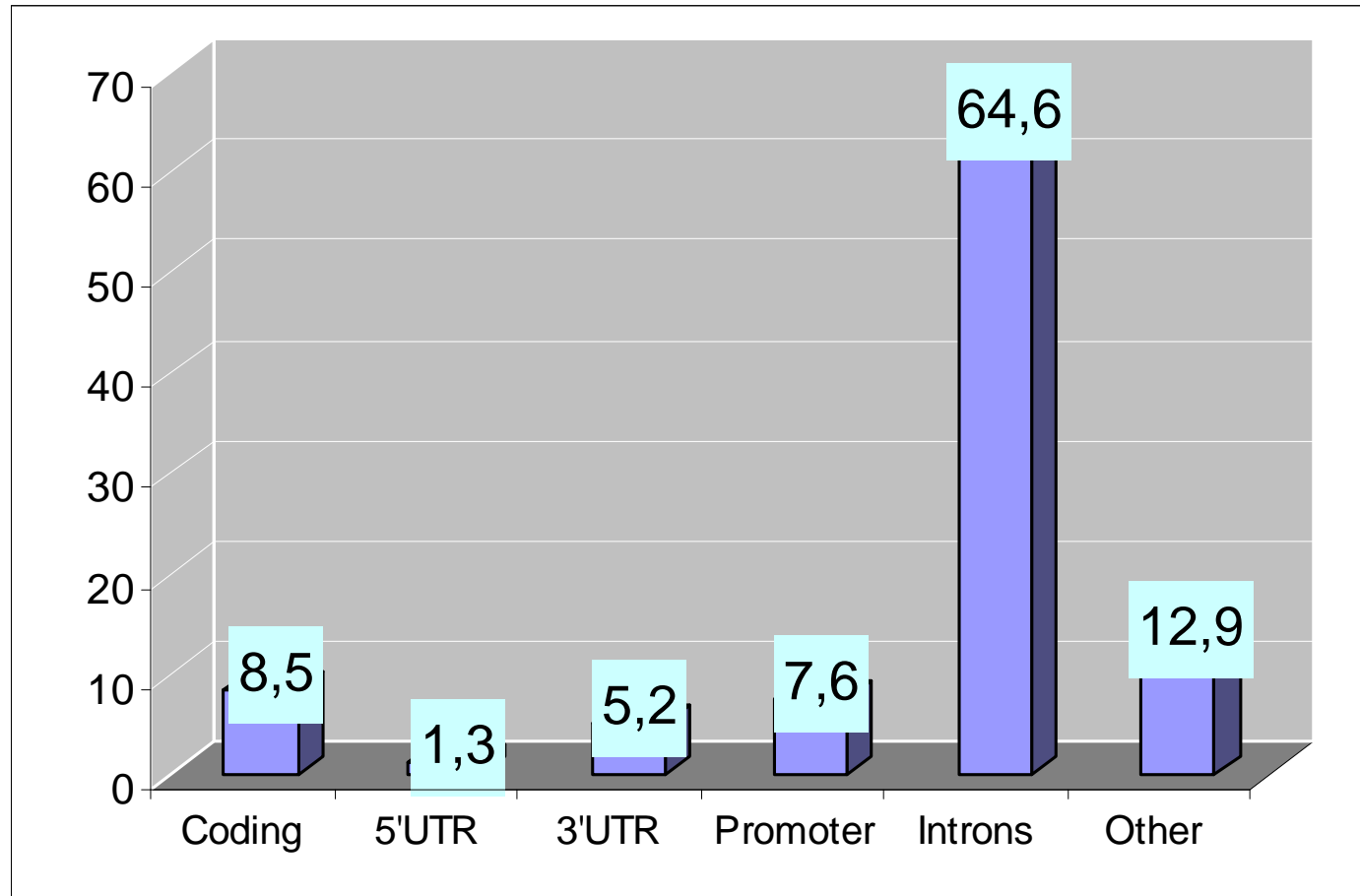
SNPs



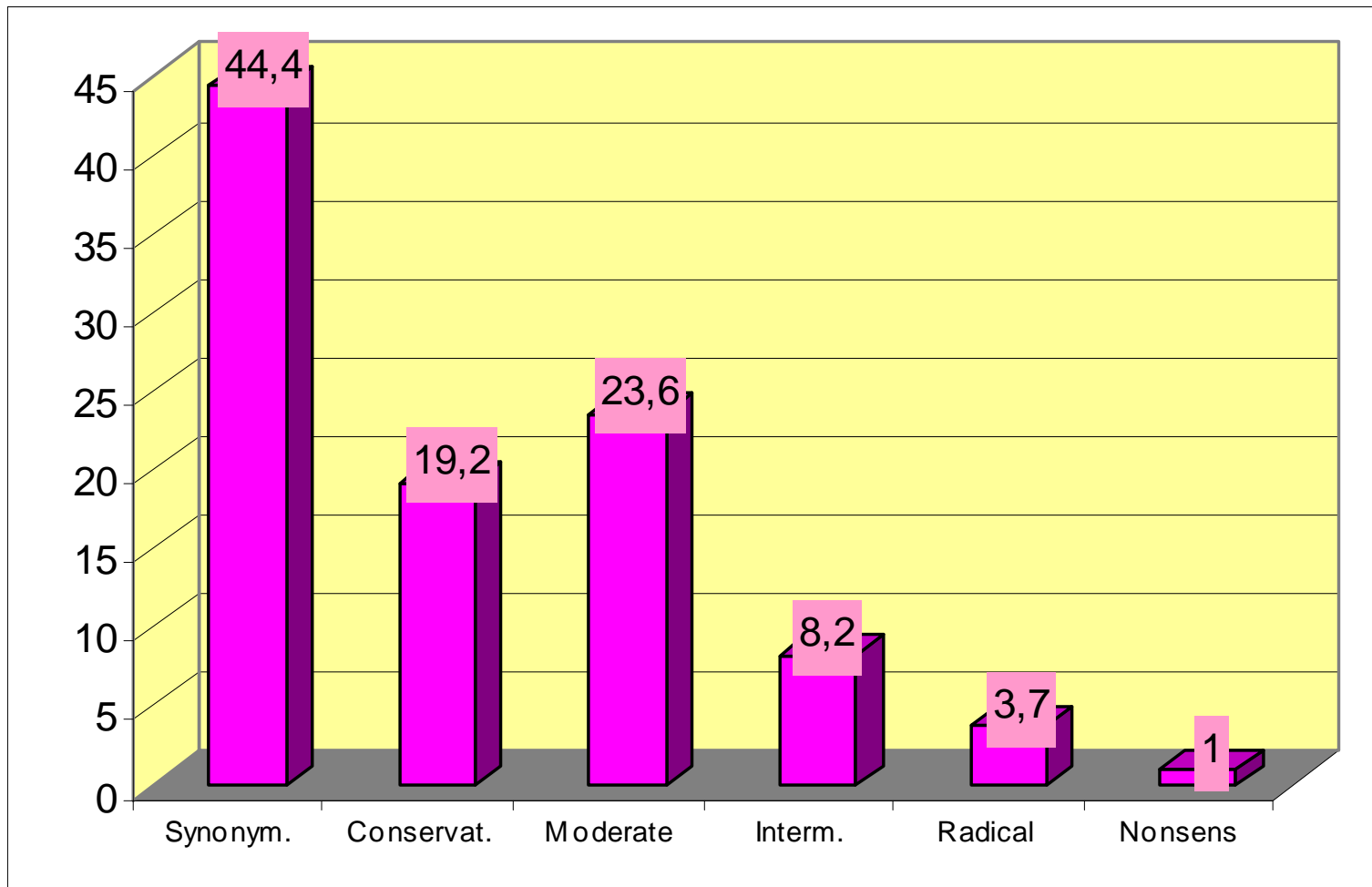
Classes of SNP



Location of SNP in gene regions



cSNP effects

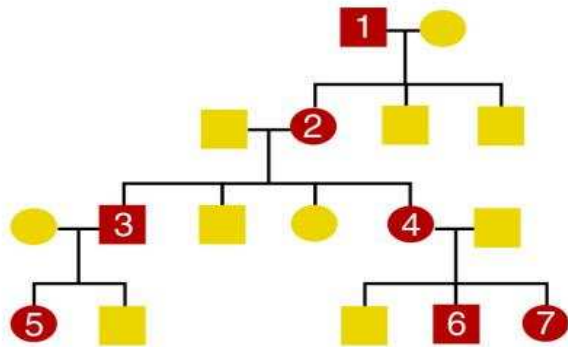


Design of association studies

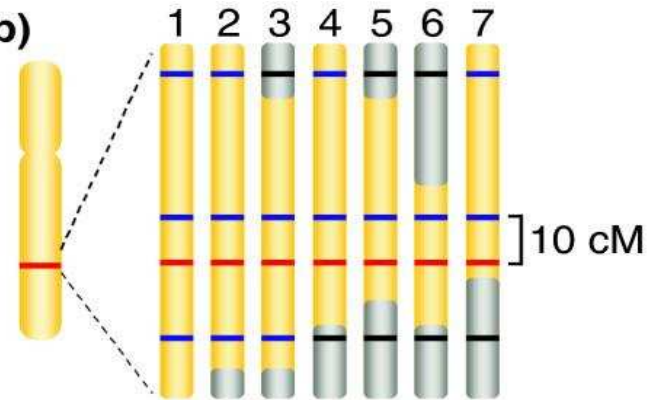
- **Family-based**: data consists in families (trios, nuclear, pedigrees,..) segregating for the phenotype
- **Population-based**: two samples one of cases (one class of phenotype) the other of (matched) controls genotyped for SNPs

Designs and methods

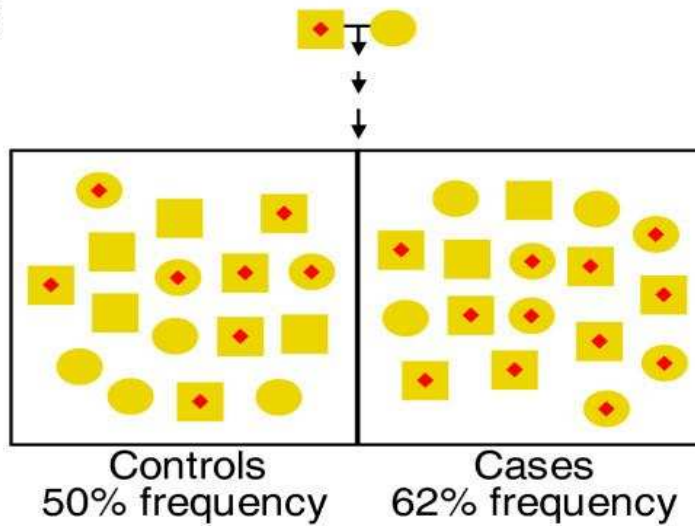
(a)



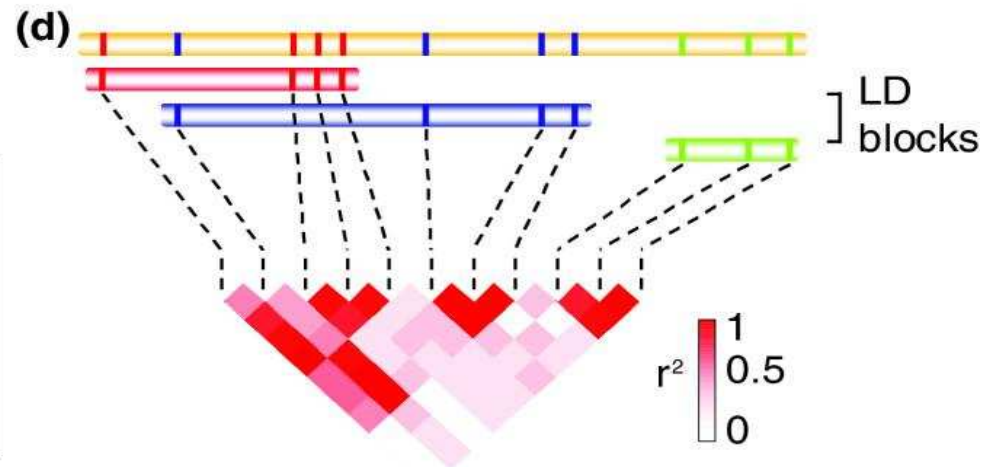
(b)



(c)



(d)



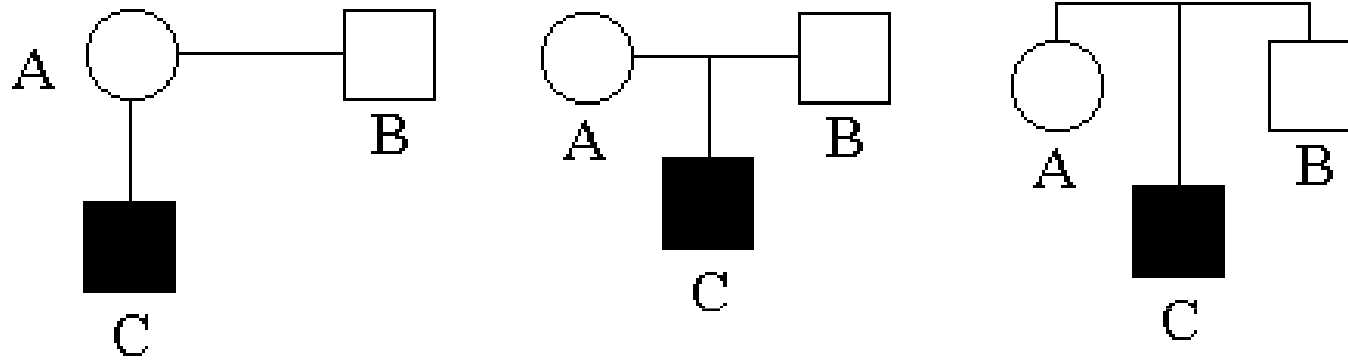
	Case-Control	Cohort	Trio
Assumptions	<p>Case and control participants are drawn from the same population</p> <p>Case participants are representative of all cases of the disease, or limitations on diagnostic specificity and representativeness are clearly specified</p> <p>Genomic and epidemiologic data are collected similarly in cases and controls</p> <p>Differences in allele frequencies relate to the outcome of interest rather than differences in background population between cases and controls</p>	<p>Participants under study are more representative of the population from which they are drawn</p> <p>Diseases and traits are ascertained similarly in individuals with and without the gene variant</p>	<p>Disease-related alleles are transmitted in excess of 50% to affected offspring from heterozygous parents</p>
Advantages	<p>Short time frame</p> <p>Large numbers of case and control participants can be assembled</p> <p>Optimal epidemiologic design for studying rare diseases</p>	<p>Cases are incident (developing during observation) and free of survival bias</p> <p>Direct measure of risk</p> <p>Fewer biases than case-control studies</p> <p>Continuum of health-related measures available in population samples not selected for presence of disease</p>	<p>Controls for population structure; immune to population stratification</p> <p>Allows checks for Mendelian inheritance patterns in genotyping quality control</p> <p>Logistically simpler for studies of children's conditions</p> <p>Does not require phenotyping of parents</p>
Disadvantages	<p>Prone to a number of biases including population stratification</p> <p>Cases are usually prevalent cases, may exclude fatal or short episodes, or mild or silent cases</p> <p>Overestimate relative risk for common diseases</p>	<p>Large sample size needed for genotyping if incidence is low</p> <p>Expensive and lengthy follow-up</p> <p>Existing consent may be insufficient for GWA genotyping or data sharing</p> <p>Requires variation in trait being studied</p> <p>Poorly suited for studying rare diseases</p>	<p>May be difficult to assemble both parents and offspring, especially in disorders with older ages of onset</p> <p>Highly sensitive to genotyping error</p>



Copyright restrictions may apply.

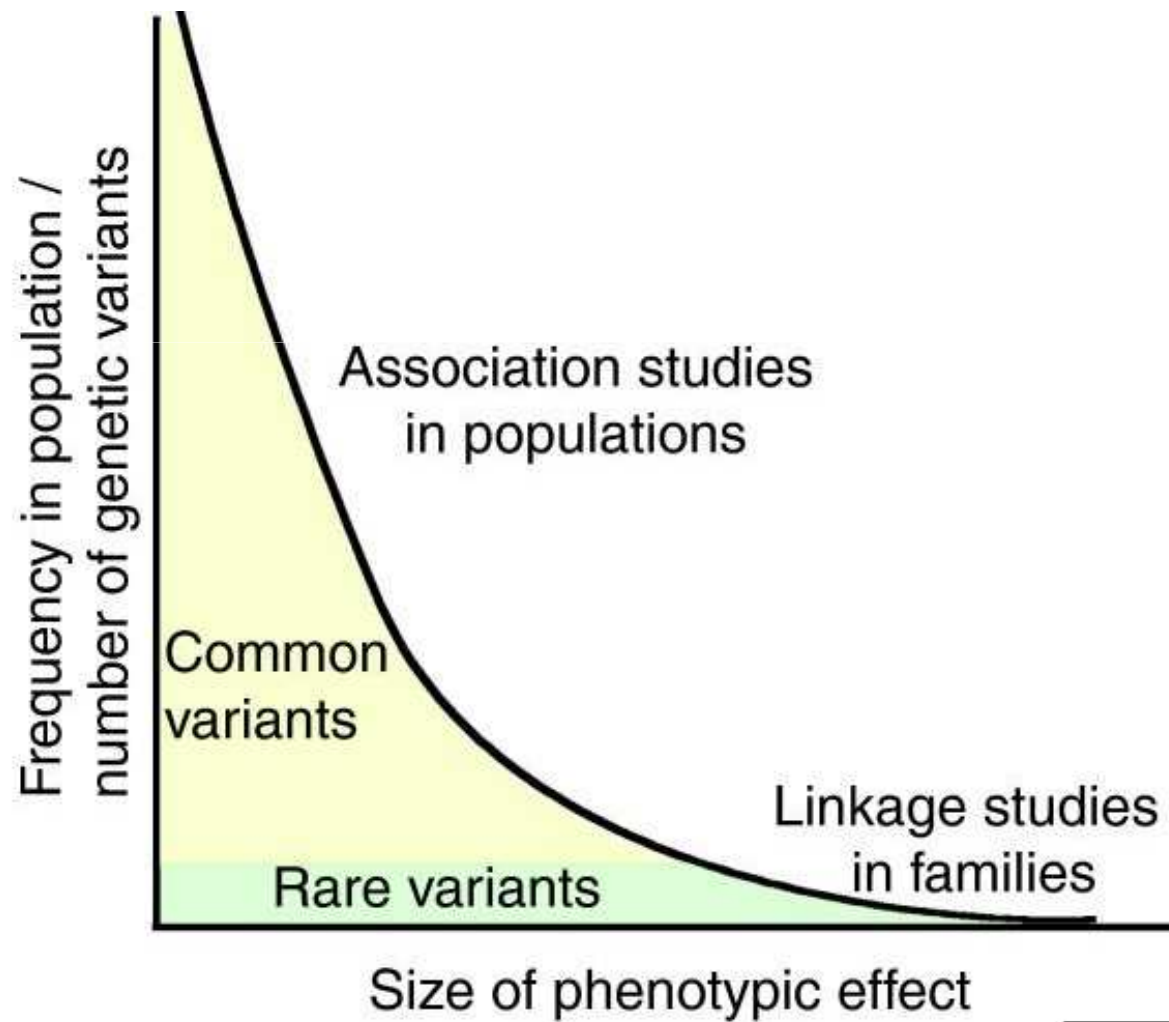
Pearson, T. A. et al. JAMA 2008;299:1335-1344

Trios vs populations



- Not easy to have the parents for late-onset diseases
- More individuals need to be genotyped

Family vs population



Types of population association studies

- **The candidate polymorphism approach:** a SNP 'suspected' of being involved in the disease causation
- **Candidate gene approach:** typing 5-50 SNP within a gene which is either a
 - Positional candidate from a prior linkage study
 - Functional candidate based on homology with a gene of known function in a model species
- **Fine mapping:** hundreds of SNP in a candidate region (1-10 Mb), containing 5-50 genes identified by a linkage genome scan.
- **The genomewide scan approach:** >300,000 SNP distributed throughout the genome

Candidate SNP



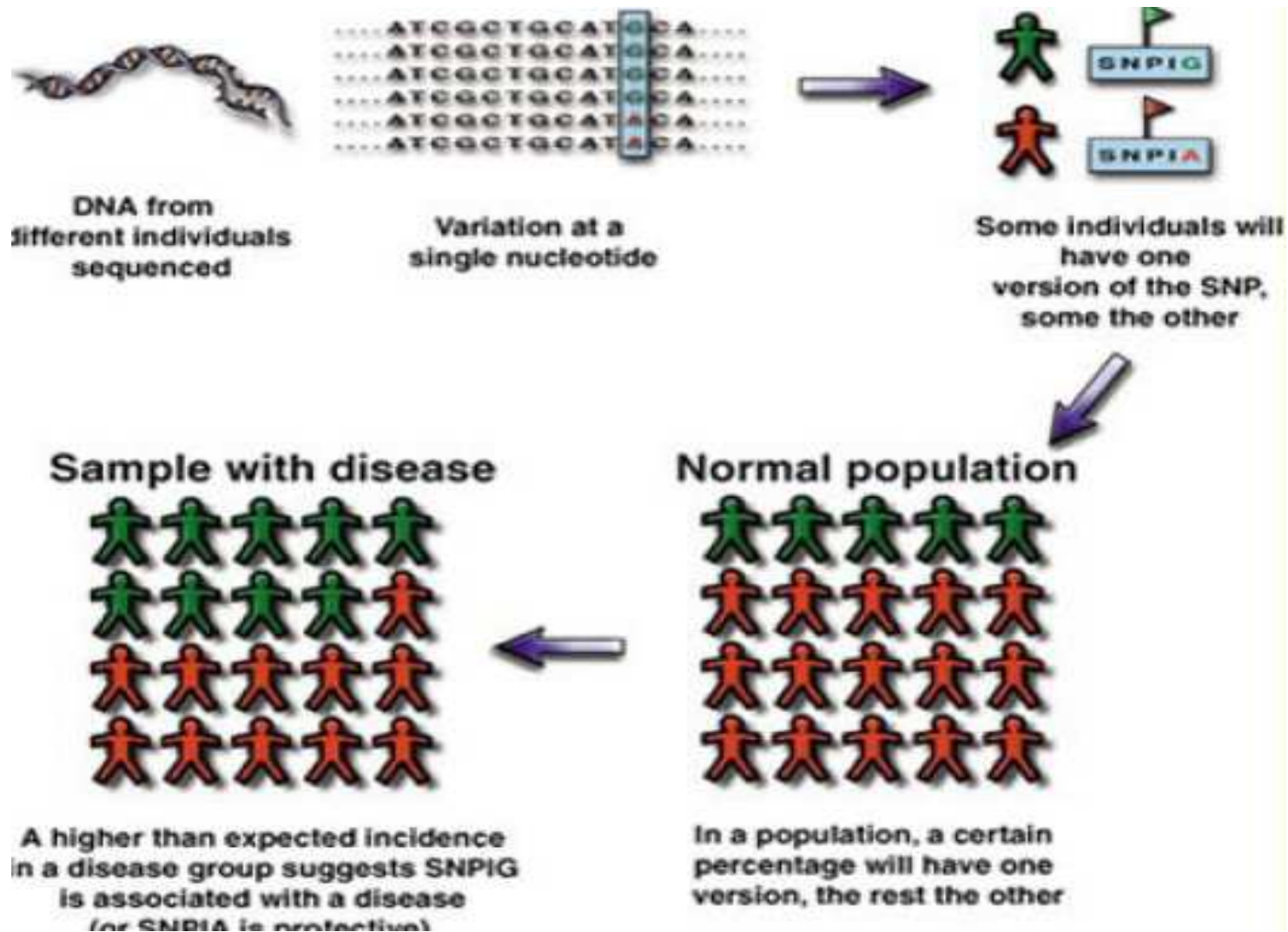
Genome-wide SNPs



GWAS

- Searching for associated SNP in a candidate gene is like looking for a lost key in a dark street
- Typing 10 million SNPs is too costly and laborious (billions of genotypes)
- Searching for an optimal set of 300 to 500 thousands SNP for use in GWAS

Basic principle of AS



GWAS data: so simple!

0	1	1	1	1	0	1	0	2	1	2	2	0	1	0	0	0	1	1	Control	
2	0	1	1	1	2	0	0	0	1	0	1	1	0	1	1	0	1	0	0	Control
2	0	1	2	2	0	1	2	1	0	0	1	1	0	1	0	0	1	1	1	Control
1	2	1	1	2	1	1	1	1	0	1	1	1	0	0	2	2	2	0	2	Control
1	1	2	1	0	1	2	1	1	1	1	2	1	2	1	2	1	2	1	1	Case
2	2	1	2	0	1	0	0	0	1	2	2	1	2	1	2	1	0	2	1	Case
0	1	1	0	0	2	1	0	0	2	1	1	1	2	1	1	2	0	1	0	Case
0	1	1	0	0	1	0	2	2	1	1	1	1	2	0	1	2	1	1	2	Case

Preliminary analyses

Checking data:
Testing before testing!

Hardy-Weinberg Equilibrium

- If the population is:
 - Panmictic: random matings and of large size
 - There is no migration
- And the locus:
 - Is not subject to selection
- Then genotype frequencies can be deduced from allele frequencies (p frequency of A):

$$AA: p^2 \quad Aa: 2p(1-p) \quad aa: (1-p)^2$$

HWE

- **Deviation from HWE** can be due to inbreeding, population stratification, selection
- Test HWE in the control sample as data quality check: discard SNP that significantly departure from HWE at $\alpha=10^{-4}$
- Ignore the case where departure can be due to tendancy to miscall heterozygotes as homozygotes in deletion polymorphisms that could be important in disease causation

Tests of HWE

- Compare observed to expected genotype counts using Pearson chi-square test of goodness of fit: with 3 genotypes and 1 parameter estimated (p) this is a test with 1 df
- Inappropriate for rare variants (low genotype counts): use Fisher Exact Test (FET)
- Other Exact tests are available in the *R language* (e.g. *Genetics package*, ...)

Pearson chi-square

- Let $D_A = P_{AA} - p^2$
- Testing HWE is testing $D_A = 0$

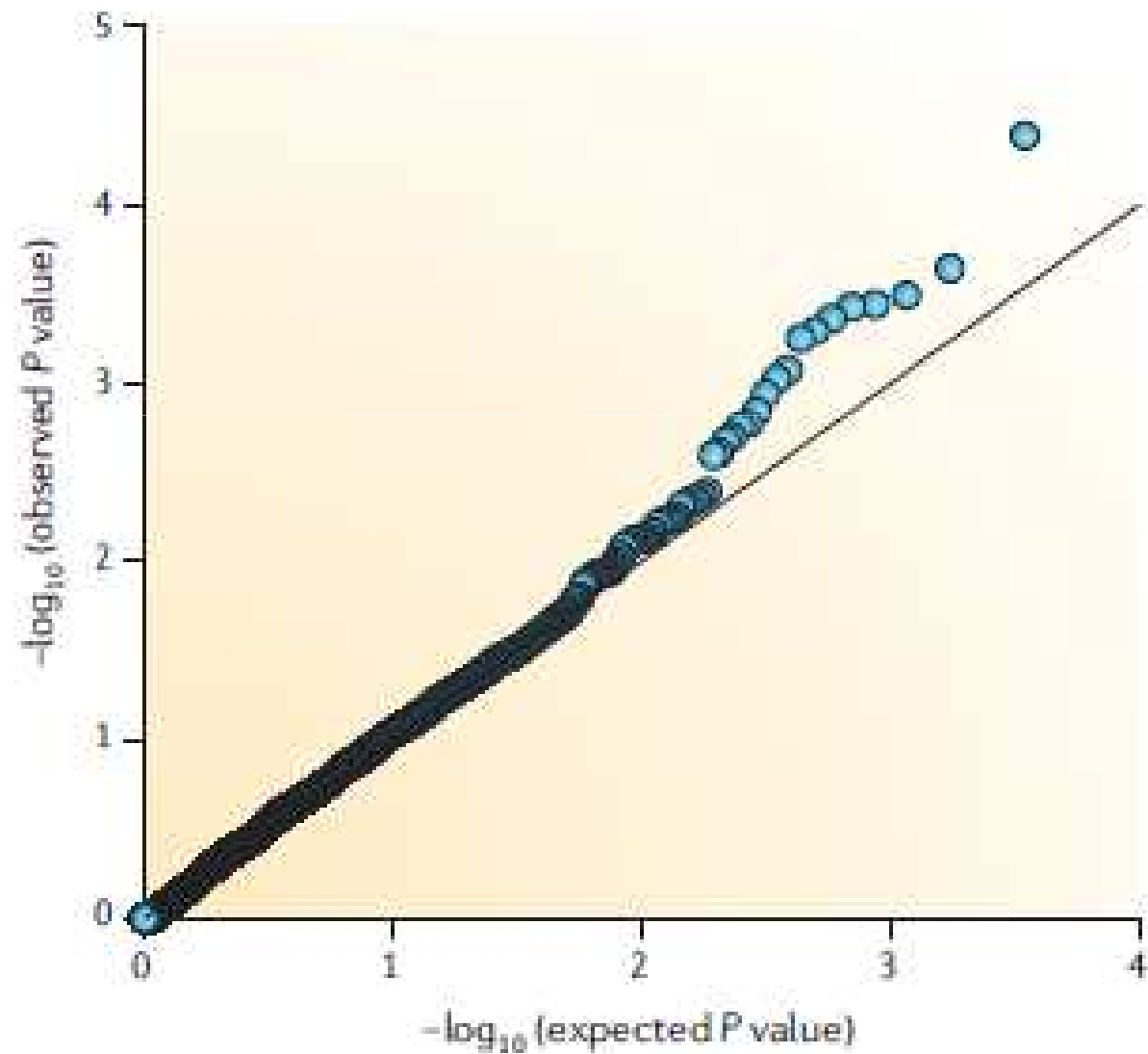
- $$\chi^2 = \frac{N D_A^2}{(p(1-p))^2}$$

- Compute *p-value* = $Pr(\chi^2_{1df} > \chi^2_{obs})$
- If *p-value* < 0,05 (or 0,0001) then Deviation from HWE

Graphical test for many SNP

- A Quantile-Quantile plot or QQ-plot of *p-values* for L SNPs:
 - sort *p-values* by decreasing order
 - plot the $-\log(i^{\text{th}} \text{ } p\text{-value})$ against $-\log(i/(L+1))$
- SNP that deviate from the diagonal line are not in HWE
- This replace correcting for multiple testing

QQ-plot for HWE



Missing genotype data

- A problem for multipoint SNP analyses
- **Data imputation**: replace missing genotypes with predicted ones
- **Predicted genotypes**: that best fits with genotypes at neighbouring SNP using:
 - **Best prediction** based on some statistical criteria (e.g. maximum likelihood)
 - **Randomly selected** from a probability distribution (resampling methods)
 - **Hot-deck**: replace with that of an individual whose genotype matches at neighboring SNP
 - **Regression models** using genotypes of all individuals

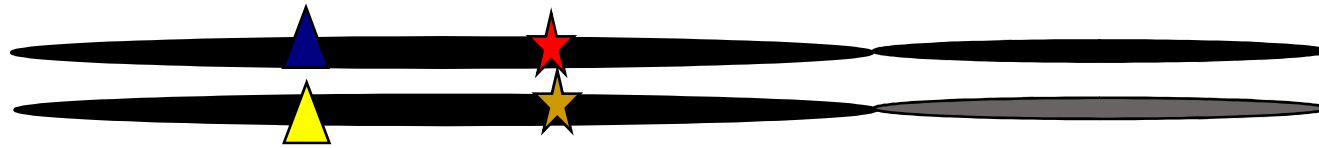
Missing genotypes

- All these approaches assume that data are missing at random (independently from the genotype) which is often doubtful due to:
 - Bad matching of cases and controls
 - Heterozygotes are genotyped as homozygotes
- Differential rate of missingness can be checked by testing association between missing status and disease status (code 0 for missing and 1 for non-missing)

Haplotypes from genotypes

- If interested in many tightly linked SNPs it is very useful to use haplotypes
- A haplotype is a set for alleles carried by one chromosome (phased)
- Haplotype of an individual can be:
 - Determined by Laboratory-based methods
 - Inferred from family members
 - Estimated using statistical methods (need genotypes of unrelated individuals)
- True haplotypes are more informative than genotypes but inferred are less (unless LD is high)

Haplotypes



a SNPs

	SNP	SNP	SNP
	↓	↓	↓
Chromosome 1	AAC A CGCCA....	TTCG G GGTC....	AGTC G ACCG....
Chromosome 2	AAC A CGCCA....	TTCG A GGTC....	AGTC A ACCG....
Chromosome 3	AAC A TGCCA....	TTCG G GGTC....	AGTC A ACCG....
Chromosome 4	AAC A CGCCA....	TTCG G GGTC....	AGTC G ACCG....

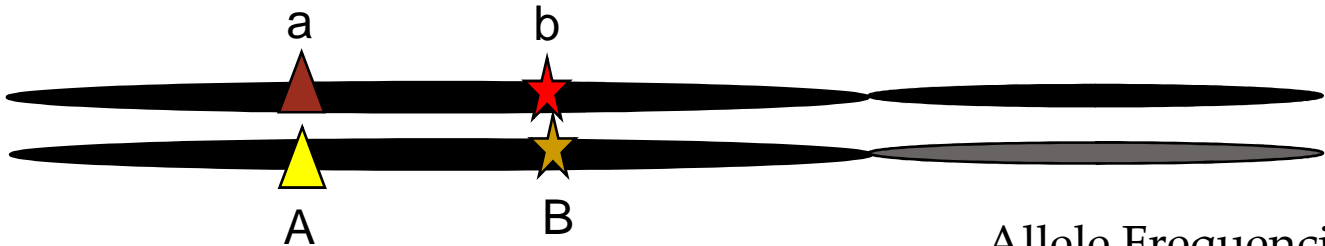
b Haplotypes

Haplotype 1	C	T	C	A	A	A	G	T	A	C	G	G	T	C	A	G	G	C	A	
Haplotype 2	T	T	G	A	T	T	G	C	G	C	A	A	C	A	G	T	A	A	T	A
Haplotype 3	C	C	C	G	A	T	C	T	G	T	G	A	T	A	C	T	G	G	T	G
Haplotype 4	T	C	G	A	T	T	C	C	G	C	G	G	T	T	C	A	G	A	C	A

↓	↓	↓
A /	T /	C /
G	C	G

c Tag SNPs

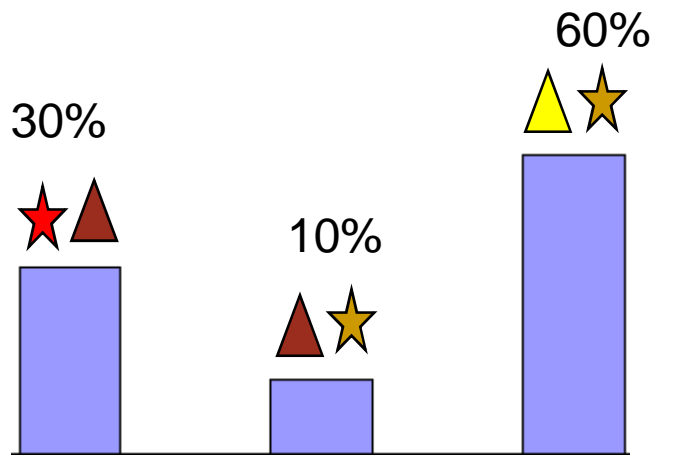
Linkage Disequilibrium



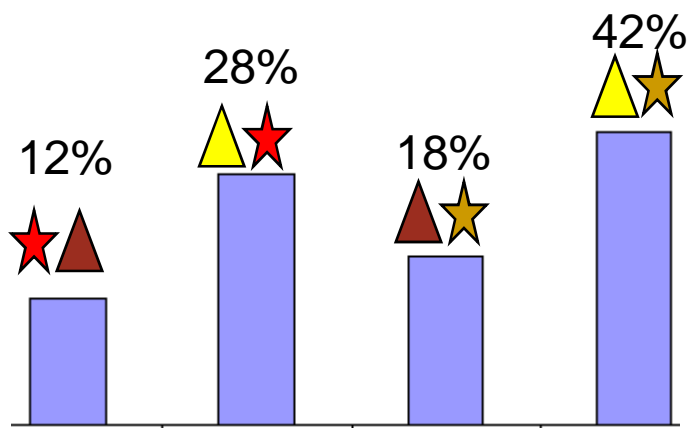
SNP1 SNP2

Allele Frequencies

- ★ 40%
- ☆ 60%
- ▲ 30%
- △ 70%



Linkage Disequilibrium (LD)



No LD

LD measures

$$D = P_{AB} - P_A P_B$$

- Choose allele A and B such that $D > 0$ and $P_A > P_B$, then

$$D' = \frac{D}{D_{\max}} = \frac{D}{(1 - P_A)P_B}$$

- $D' = 1$ denotes complete LD

Correlation: more practical

- This is correlation from the 2x2 contingency table of haplotype counts

- Or
$$r^2 = \frac{D^2}{P_A(1-P_A)P_B(1-P_B)}$$

$$r^2 = (D')^2 \frac{P_B(1-P_A)}{P_A(1-P_B)}$$

LD Measures

- D' can be large (indicate high LD) even when one allele is very rare, which is of little practical interest
- Nr^2 is the chi-square test in 2x2 table of haplotype counts
- r^2 is directly related to statistical power: if disease risk is multiplicative and HWE holds then r^2 between a SNP and a causal variant is the sample size required to detect association by directly typing the causal variant, relative to that required to achieve the same power when typing the SNP.

In other words

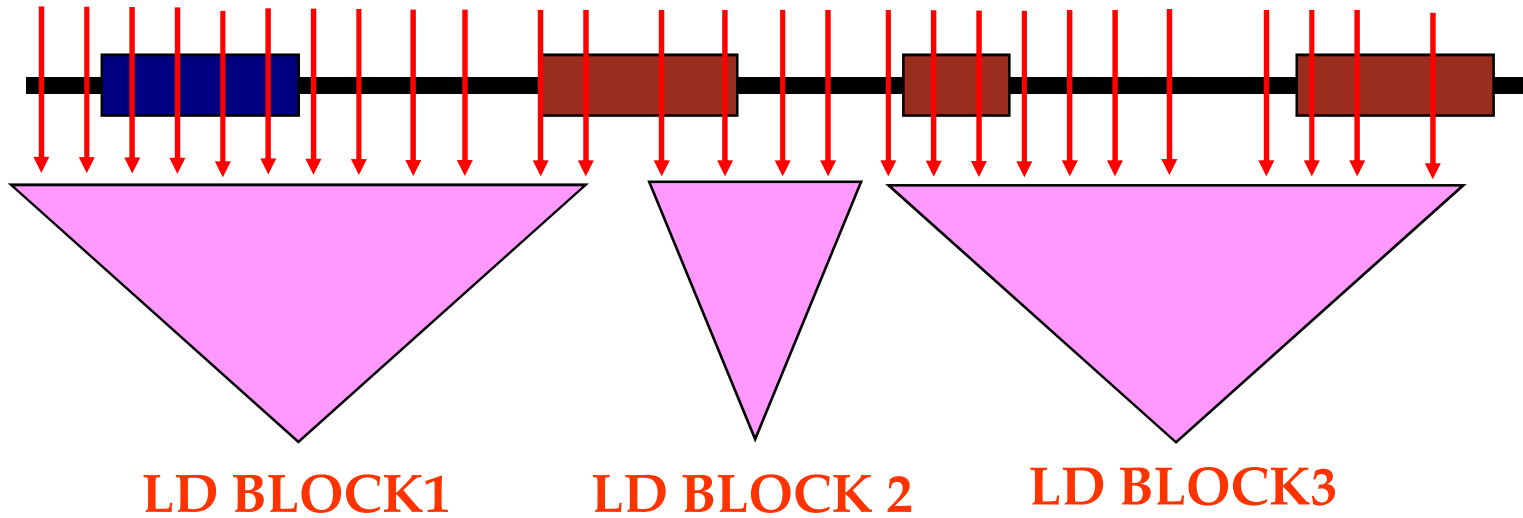
- If you have a SNP having an $r^2=0.10$ with a causal variant and if
- you need a sample of **100 individuals** to detect association with the causal variant with **80% power**
- Then you need **$100/0.1=1000$ individuals** to detect association (with 80% power) with the SNP.

SNP tagging

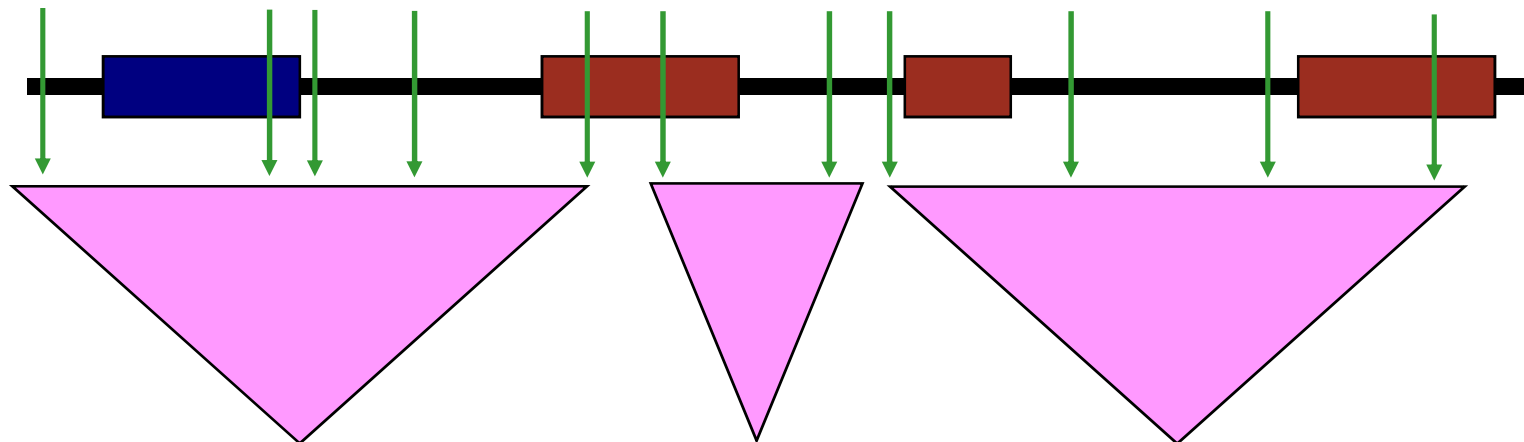
- Select a minimal numbers of SNP that retain as much as possible of the genetic variation of the full SNP set

LD blocks and TagSNP

→ SNP



→ tSNP



Methods for SNP Tagging

- **Simple**: for each pair of neighbor SNP discard the one (having the most missing data) if $r^2 > 0,9$
- **Sophisticated**: find the smallest number of SNPs that need to be genotyped to cover the other SNPs at an $r^2 \geq 0.8$
- Regression methods
- Linear Dynamic programming

Usefulness of tagging

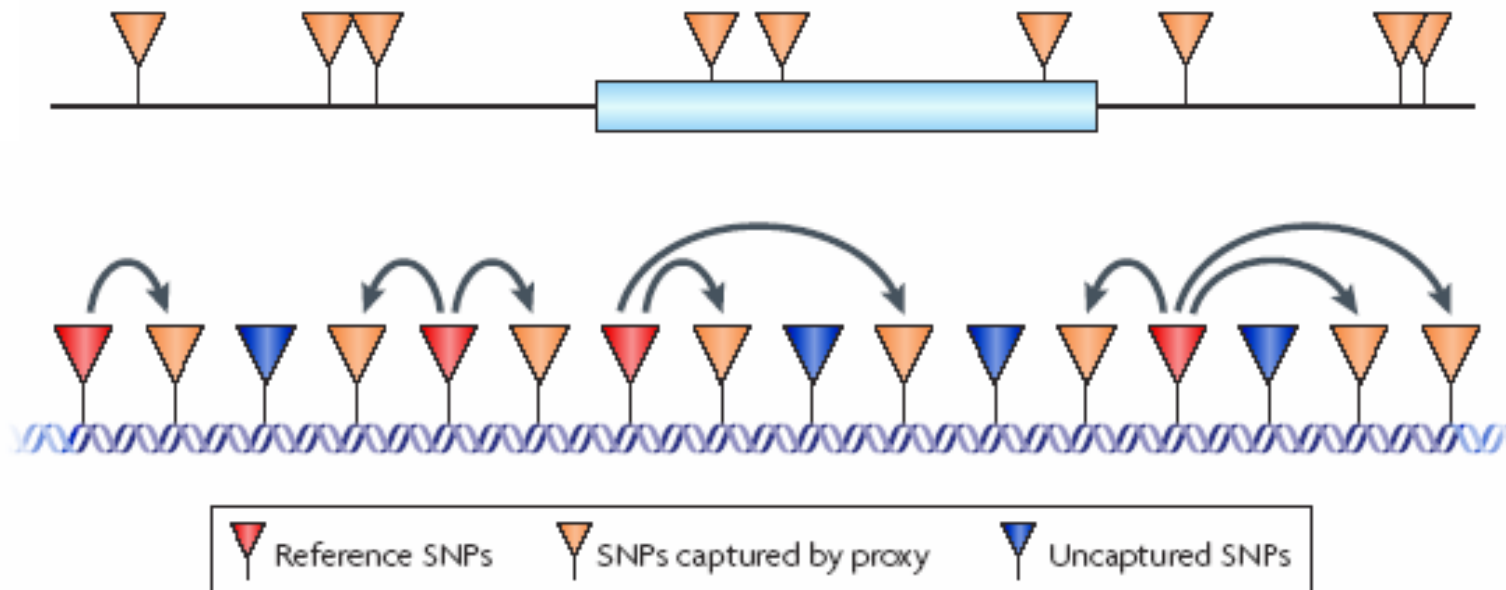
- The HapMap project
- **Transferability**: a tag SNP selected in one population might not perform well in another but in general it is good
- **Use only tagSNP** for analysis even if all have been genotyped.
- Some SNPs are not captured !

Missed SNPs

a Direct:
catalogue and test all functional variants for association



b Indirect:
use a dense SNP map and test for linkage disequilibrium



HapMap Project

- The goal was to determine the common patterns of DNA sequence variation in the human genome (a Haplotype Map) by characterizing :
 - Sequence variants
 - Their frequencies
 - Correlation between them
- From population with african, asian and european ancestry

Hapmap phases

- **The phase I** was to genotype one SNP every 5 kb in 270 individuals from 4 geographic regions :
 - 30 individuals from the Yuruba (Nigeria)
 - 30 from the CEPH project in Utah
 - 45 Han chinese
 - 45 Japenene from Tokyo
- **Phase II**: typing 4 million SNPs in the same samples (completed in 2005)
- **Phase III**: other population samples (open)

Visit:
www.hapmap.org

Tests of association

A single SNP

Pearson chi-square Test

- If we construct the table of genotype counts in cases and controls

Genotype	<i>AA</i>	<i>Aa</i>	<i>aa</i>
Cases	P_1	Q_1	R_1
Controls	P_0	Q_0	R_0

- Use the chi-square test (2 df) or FET
- In complex traits the contribution to disease risk of SNPs is often **roughly additive** (risk of *Aa* is intermediate between that of *AA* and *aa*). The chi-square is not the best test in this case. To improve power we can use an **allelic test**.

Risk models

- There are four possible risk models for any given SNP depending on relative risk;
- Take genotype *aa* as a reference genotype with risk equal to 1 then:

Genotype	<i>AA</i>	<i>Aa</i>	<i>aa</i>
Additive	2γ	γ	1
Dominant	γ	γ	1
Recessive	γ	1	1
Multiplicative	γ^2	γ	1

Allelic test

- Define the allele count table from genotypes

Allele	<i>A</i>	<i>a</i>
Cases	$2P_1 + Q_1$	$2R_1 + Q_1$
Controls	$2P_0 + Q_0$	$2R_0 + Q_0$

- Chi-square test with 1 df
- Not recommended because it requires HWE in cases and controls combined and risk estimates are not interpretable

Allelic test

		Allele		
		A	a	
Disease Status	+	a	b	a+b
	-	c	d	c+d
	a+c	b+d	2N	

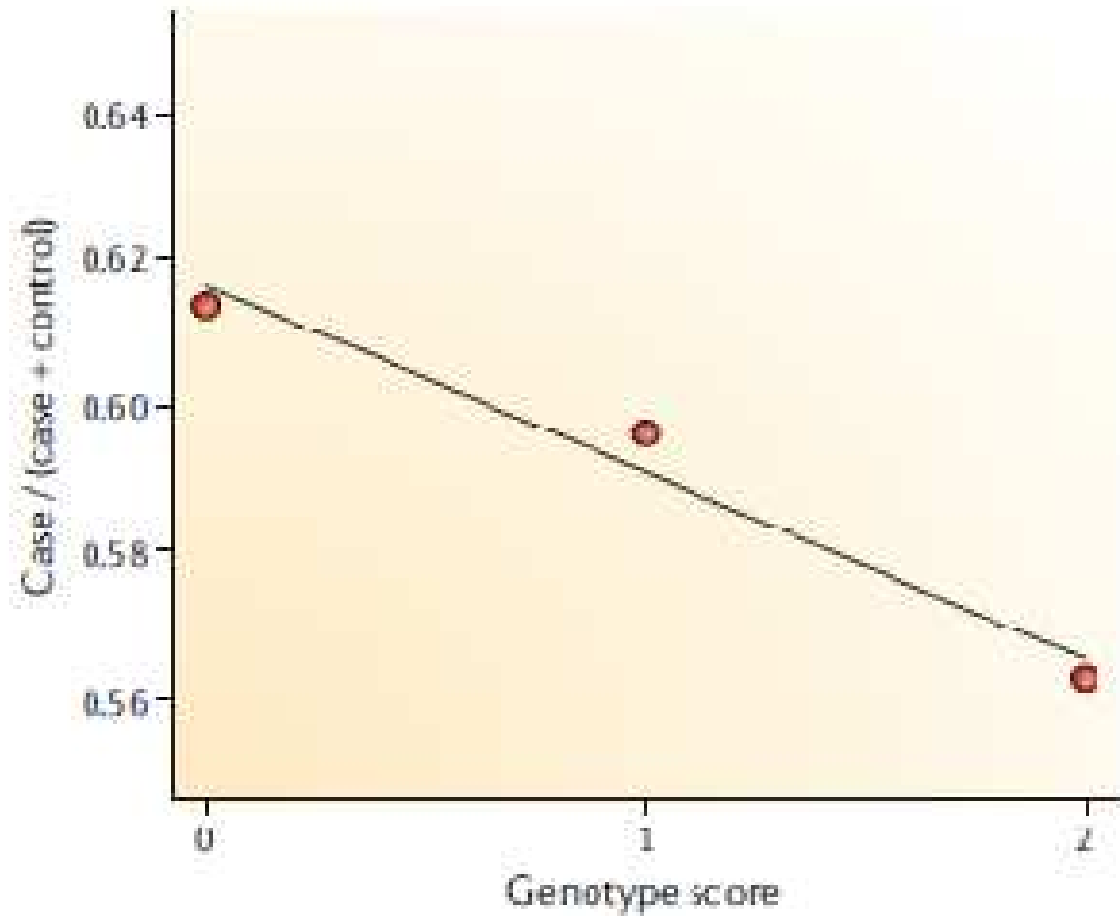
$$\chi^2 = N \frac{(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

$$p\text{-value} = \text{Prob}(\chi^2_{1df} > \chi^2_{obs})$$

Improved allelic tests

- Nuel et al (2006) proposed an **exact allelic test** that is not biased by departure from HWE (implemented in R).
- Song and Elston (2006) proposed a **correction for allelic trend test** when HWE does not hold.
- **The Cochran-Armitage test** is a conservative allelic test not relying on HWE: fit a horizontal line to proportion of cases in the three genotypic classes

Armitage test of association



Logistic regression

- Let us denote by π_i the disease risk for individual i ($\pi_i = \text{Prob}(y_i=1)$), the model consists in stating that
- $\text{Logit}(\pi) = \log(\pi/(1-\pi)) =$
 - β_0 for aa
 - β_1 for Aa
 - β_2 for AA
- To test association we test: $\beta_0 = \beta_1 = \beta_2$
- If we set :
 - $\beta_1 = (\beta_0 + \beta_2)/2$ we get an additive model
 - $\beta_1 = \beta_0$ we get recessive model
 - $\beta_1 = \beta_2$ we get dominant model

Logistic regression

- The advantages are that:
 - Many SNPs can be included in the same model, allowing test for epistasis and gene by environment interaction
 - SNP effect can be tested while adjusting for covariates such as age of onset, gender, ...

Which test to use?

- There is no generally accepted answer!
- FET spread over the range of risk models but less powerful to detect near-additive risks.
- Armitage: good for additive models, weak power for other models
- The problem is that the model is unknown
- Take the Max of test statistics over models
- Armitage for rare variants, FET elsewhere
- Bayesian Testing

Bayesian testing: a different way of thinking

- Instead of computing a p-value (probability of having the test value by chance) we compute a Posterior Probability of Association (PPA):
 - Choose a value of the prior probability of association π (10^{-4} to 10^{-6})
 - Compute the Bayes Factor for each SNP
BF=Pr(Data/ Association)/Pr(Data/no association)
 - Calculate the Posterior Odd and then PPA

$$PO = BF \frac{\pi}{1 - \pi} \quad \text{and} \quad PPA = \frac{PO}{1 + PO}$$

Example

Trait	SNP*	p-values [†]		$\log_{10}(\text{BF})^{\S}$	PPA	
		Trend test	General test		$\pi = 10^{-4}$	$\pi = 10^{-5}$
BD	rs420259	2.2×10^{-4}	6.3×10^{-8}	4.1	0.56	0.11
CD	rs9858542	7.7×10^{-7}	3.6×10^{-8}	4.7	0.83	0.33
T2D	rs9939609	5.2×10^{-8}	1.9×10^{-7}	5.3	0.95	0.67
CD	rs17221417	9.4×10^{-12}	4.0×10^{-11}	8.9	0.99999	0.99987
T1D	rs17606736	2.2×10^{-15}	1.5×10^{-14}	12.5	1.00000	1.00000

Advantages of Bayesian

- **Allows averaging over genetic models** by computing a combined BF between models
- **Allows Averaging over effect sizes:** SNP with higher to low risk
- **Allows incorporating external biological information:** SNP near genes, with known biological function, with low frequency, conserved among species, .. can be given higher π

Measuring Risk

- A measure of risk is the odds ratio:

$$OR = \frac{a/(a+c)}{b/(b+d)} \approx \frac{a/c}{b/d} = \frac{ad}{bc}$$

- If $OR=1$, no association

$$CI_{95\%} = \frac{ad}{bc} \exp \left[\pm 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right]$$

- If CI contains 1, no significant association (at 5%)

	A	a	
Disease +	a	b	a+b
-	c	d	c+d
	a+c	b+d	2N

Population attributable risk

- Represents the excess risk of disease in those having the risk allele with those not having it

$$PAR = \frac{K(OR - 1)}{K(OR - 1) + 1}$$

K is the prevalence of carriers in the population

- Can be approximated, for a rare dominant risk allele by

$$PAR \approx P_{Aa} \frac{1 - 2p(1 - P_{aa})}{P_{aa}(1 - 2p)}$$

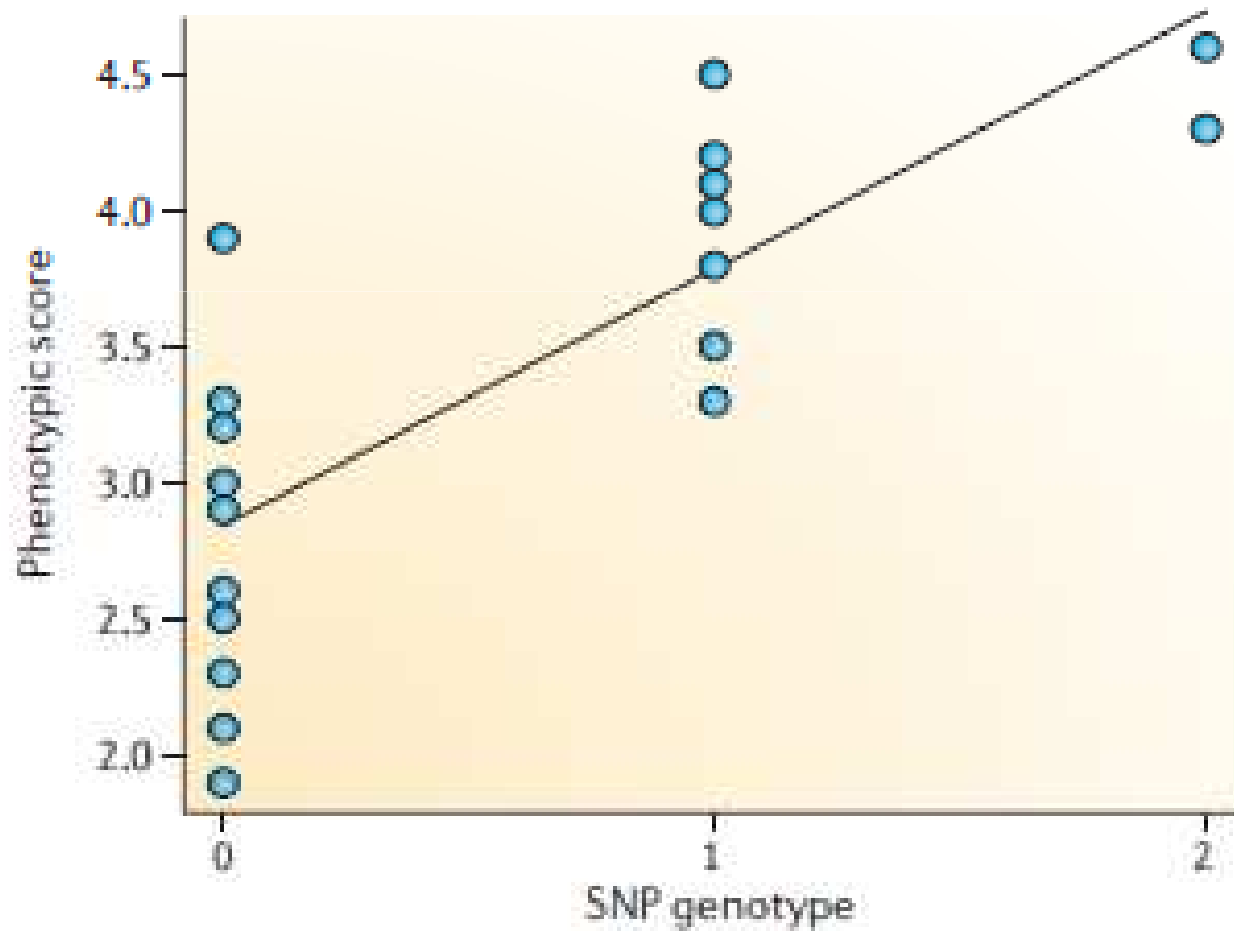
Categorical Phenotypes

- Categorical traits can be:
- **Unordered**: disease subtypes and association can be tested by multinomial regression
- **Ordered** such as disease severity (mild, moderate, severe) and we need a method that gives more weight to the most severely affected cases (diagnosis is more certain, causal genes contribute more)
- If we assume that the risk for category k relative to category $(k-1)$ is the same for all k , then we can build a score test (generalization of Armitage test)

Continuous phenotype

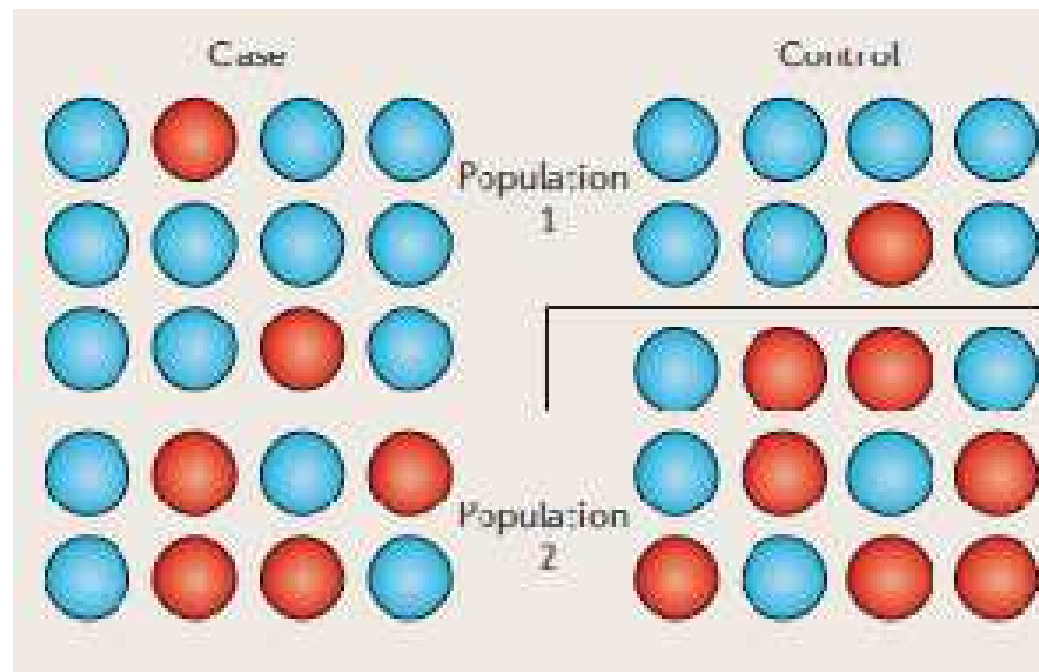
- We use mean comparison (analysis of variance) or linear regression between the three genotypes
- Both require the trait to be **Normally** distributed for each genotype class and have the same variance;
- If not a transformation of the trait might be necessary (log, inverse, square root, box-cox)

Linear regression



Complicating factors!

- Population stratification can generate spurious genotype-phenotype association



Genomic Controls

- We consider a set of about 100 « null » SNPs (that are mostly not related to the disease)
- The Armitage test is computed for each null SNP
- Compute $\hat{\lambda}$, the median of test values divided by its expectation
- If $\hat{\lambda} > 1$ (which is indicative of stratification), then divide test value by $\hat{\lambda}$
- **Caveats:** Limited in applicability, conservative, problem in choosing null SNPs

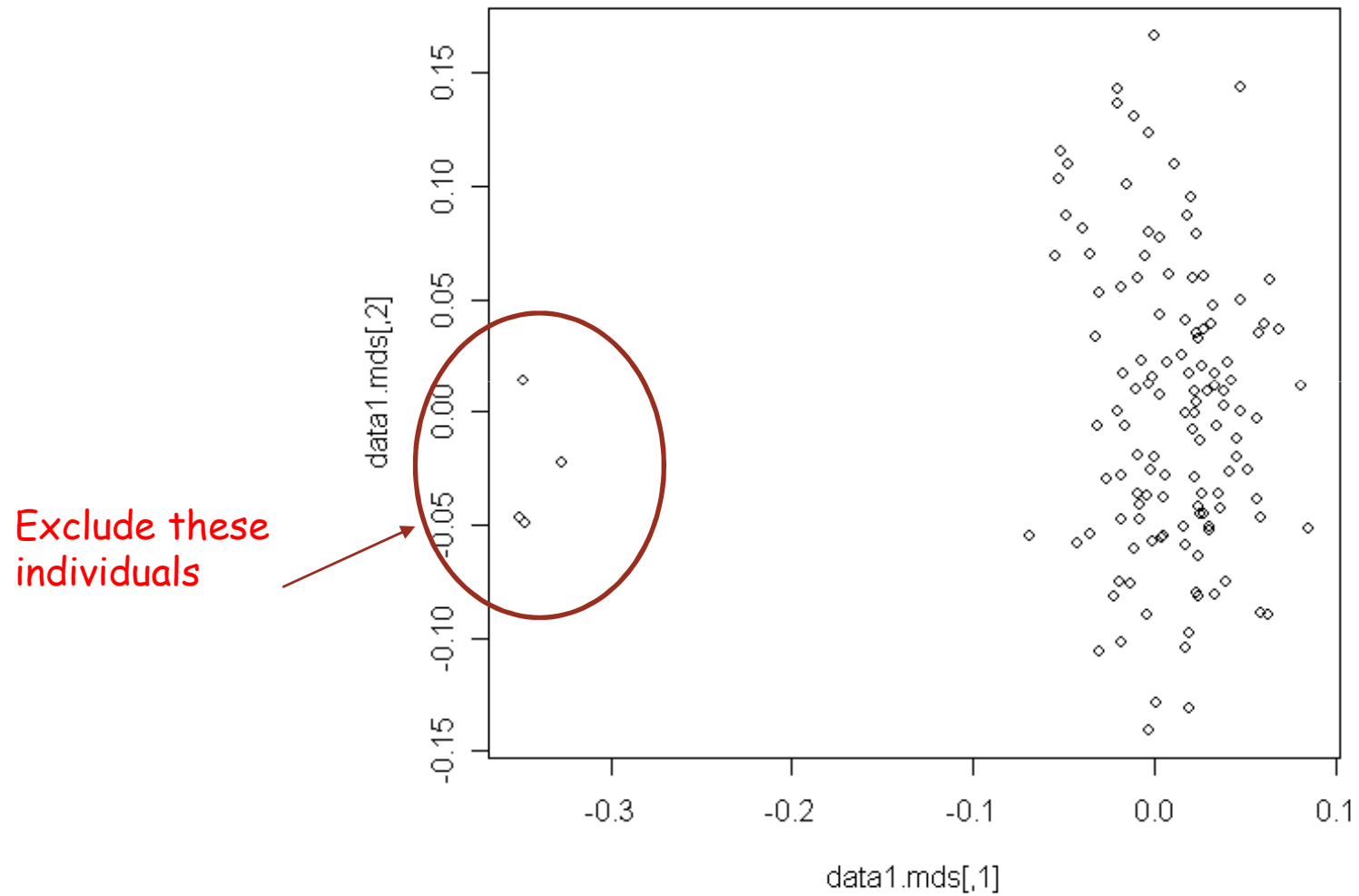
Structured association methods

- Searches for the best sub-population structuring by optimizing some criteria
- Allocate individuals to hypothetical sub-populations
- Test for association conditional on this allocation
- **Caveats:**
 - Computationally demanding,
 - Subpopulations are theoretical constructs and have no direct interpretation

Other methods

- **Include null SNP as covariates in regression analyses:** computationally fast, more flexible than GC but it is recommended to assess type-I error by simulation.
- **Use Principal Component Analysis** to diagnose population structure using null SNPs
- **Mixed-model approaches** that estimates kinship (relatedness between individuals)

Kinship between individuals



Power and sample size

- In statistical testing we consider:
 - a null hypothesis H_0 : « no association »

versus

- an alternative hypothesis H_1 : « association »
- This results in two types of error
 - The first (type-I, α) is fixed (chosen) and
 - The second (type-II, β) can be calculated for given values of disease variant parameters (risk and allele frequency), a given risk model and a given sample size.

Errors in statistical testing

Truth: unknown

Decision

	H ₀ True no association	H ₀ false association
Accept H ₀ Declare absence of association	1- α Confidance level	β (type II error)
Reject H ₀ Declare association	α (type I error): 5%	1-β Power

How to compute power?

- Power = Pr(Declaring association / there is actually association)
- If we have the theoretical distribution of the test statistic then

$$Power = \Pr(\chi_{1df}^2 > \chi_{1df, \alpha}^2 / p, \gamma, n, model)$$

- Theoretical power can be computed by analytical approximate formula

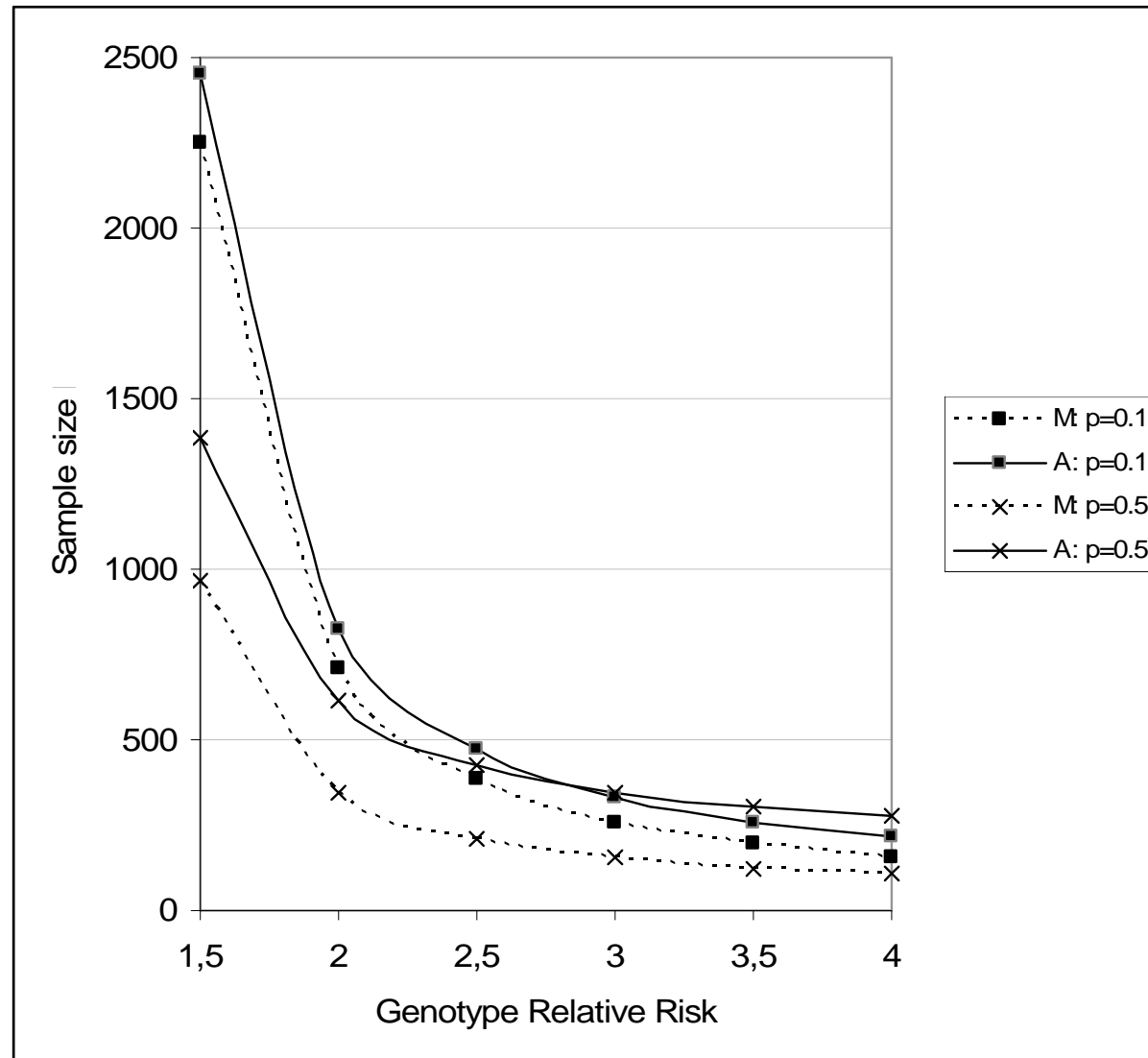
Empirical power

- Power of the sample under study per se can be computed using resampling technique such as bootstrapping or Monte Carlo methods:
- **Bootstrap**: create M new samples by
- **Allocating for each individual a genotype** by random selection from the original genotypes array (with replacement)
- **Compute test statistic** for each sample
- **Estimate power** as the proportion of samples in which association is declared (test value is greater than the predefined threshold at a given α)

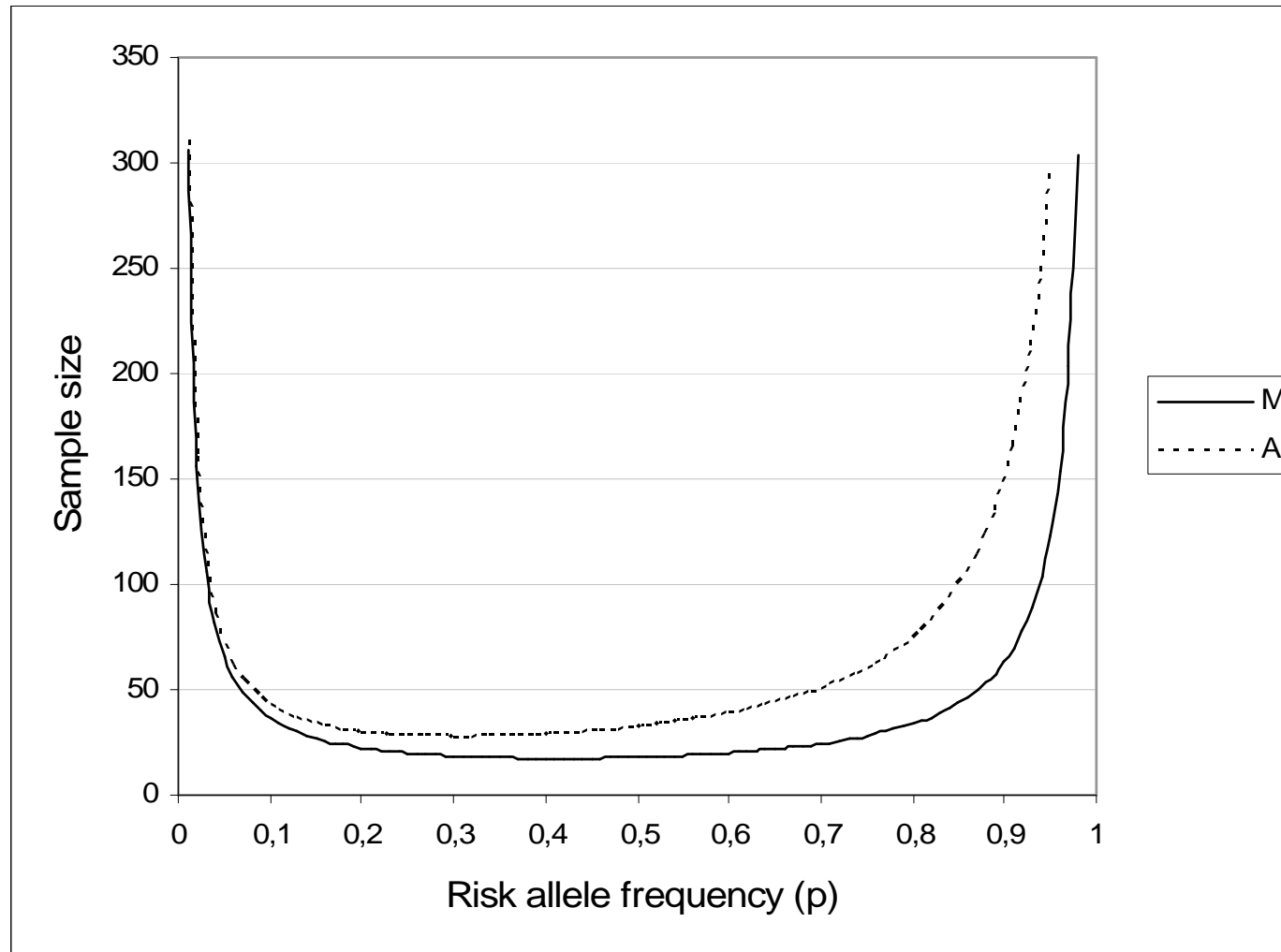
Bootstrap

0 1 1 1 1 1 0 1 0 2 1 2 2 0 1 0 0 0 1 1	Control	2	2
2 0 1 1 1 2 0 0 0 1 0 1 1 0 1 1 0 1 0 0	Control	1	0
2 0 1 2 2 0 1 2 1 0 0 1 1 0 1 0 0 1 1 1	Control	1	1
1 2 1 1 2 1 1 1 1 0 1 1 1 0 0 2 2 2 0 2	Control	2	2
1 1 2 1 0 1 2 1 1 1 1 2 1 2 1 2 1 2 1 1	Case	2	2
2 2 1 2 0 1 0 0 0 1 2 2 1 2 1 2 1 0 2 1	Case	2	1
0 1 1 0 0 2 1 0 0 2 1 1 1 2 1 1 2 0 1 0	Case	1	0
0 1 1 0 0 1 0 2 2 1 1 1 1 2 0 1 2 1 1 2	Case	0	0

Power and gene risk



Power and allele frequency



Testing association

The multiple SNP scenario

Unphased genotypes: Logistic regression

- A model including all SNPs as well as covariates, interaction effects,...
- A score test with $2L$ df (L df if we assume additivity)
- Use only tagging SNP to eliminate redundancy and increase power
- Use stepwise selection procedure to avoid highly correlated SNPs
- Assessing significance is problematic!

Combining single locus tests

- Use cumulative sums of single locus tests and identify those that are of particular interest
- Detecting local high-scoring segments, groups of neighbor SNPs that have small association p-values by methods and algorithms similar to those used in finding sequence patterns.

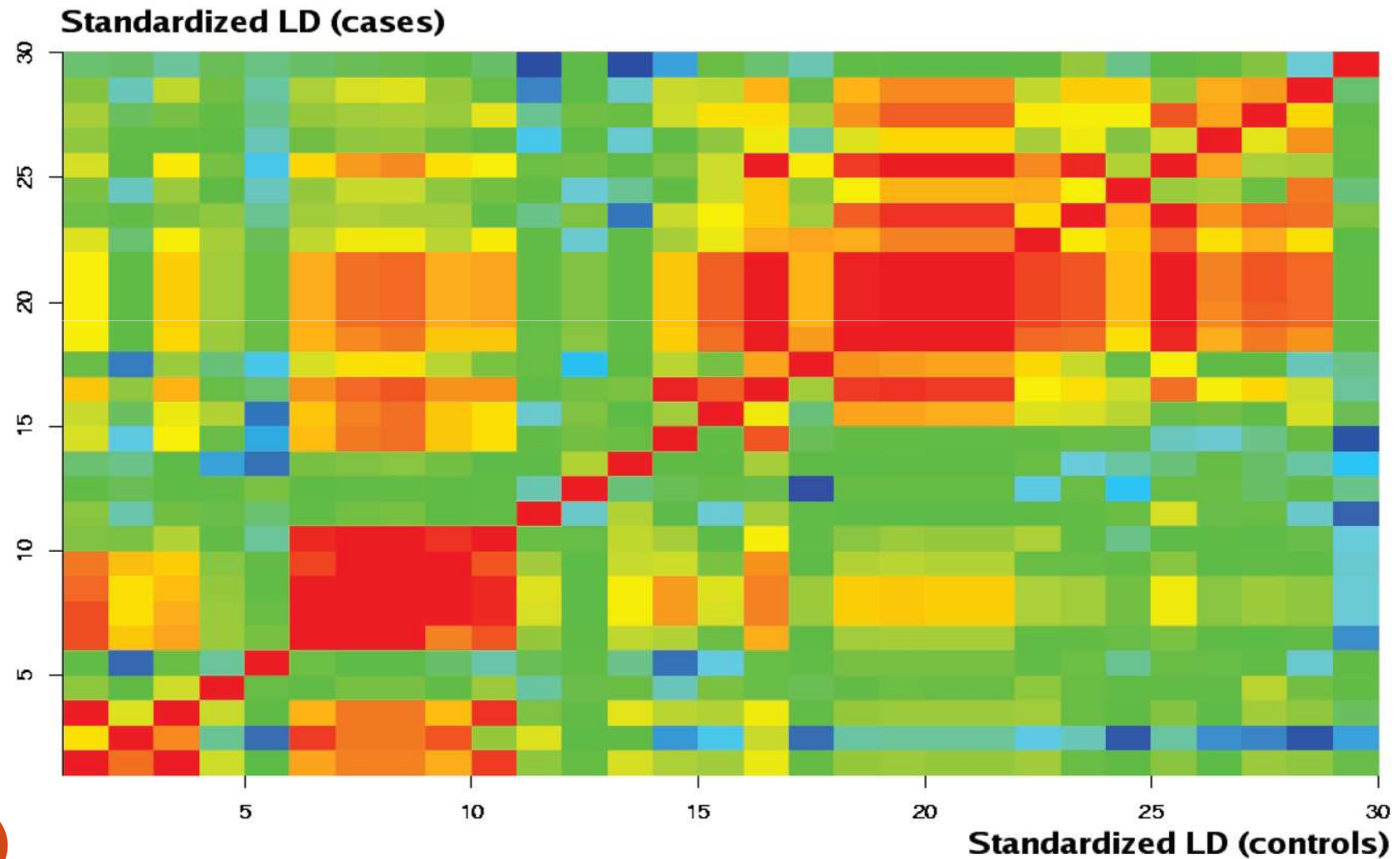
Haplotype-based methods

- Reduce the number of df in models
- Capture correlation structure of SNP in LD blocks
- Capture combined effect of highly linked cis-acting causal variants
- **Caveats:** haplotypes are not observed but inferred and it is hard to account for the uncertainty of their inference

Haplotype tests

- Use a $2 \times k$ contingency table (problem of zero cells for rare haplotypes) or Compare frequencies of haplotypes (rely on HWE and near-additive risk)
- Haplotypes are treated as categorical variables in regression analyses
- Compare patterns of LD between cases and controls (Zaykin et al, 2006)

Contrasting LD patterns



Problems

- **Rare haplotypes**: including them results in loss of power if haplotypes are similar but correspond to distinct causal variants
- **Solution**: Combine rare haplotypes in controls into a single category
- LD block vary with sample size, SNP density and block definition
- Use clustering to identify sets of haplotypes sharing common ancestry

Three major complicating factors

- Missing data
- Epistasis
- Gene-environment interaction

Missing Genotype imputation

- Seen before!

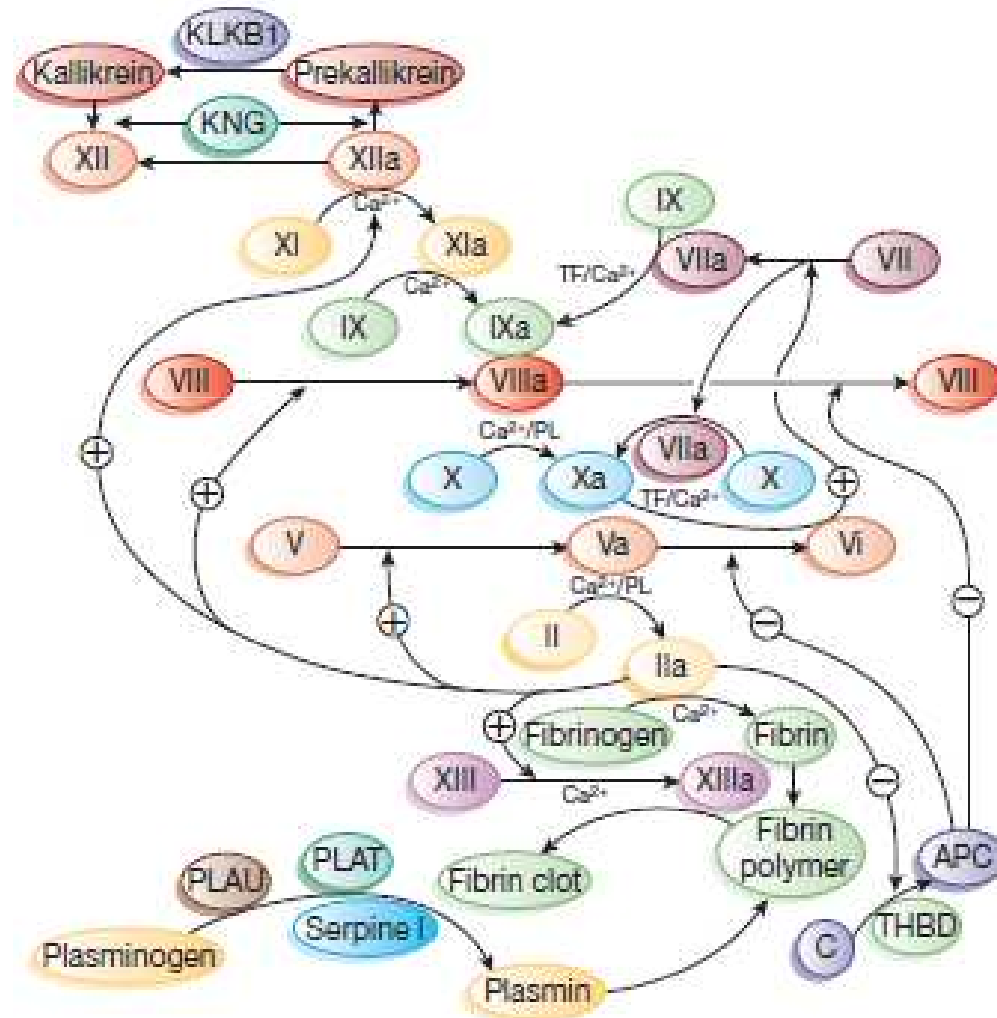
Epistasis

- A variant with a small marginal effect of individual SNPs might turn to have a strong effect in certain genetic background and be of clinical significance
- Is it better to tackle epistasis directly or first focus on marginal effects?
- The inclusion of epistasis is very easy in regression methods but testing all combinations is unwise: should be limited to genes with no marginal effects

Gene-environment interaction

- The risk conferred by alleles or genotypes is not the same across environments
- **Environment** often has a very « loose » definition: nutrition, lifestyle, exposition to 'pollution' (smoking, solvents,..)?
- Test for association in different samples defined according to their environment?

Higher order interactions?



The multiple testing problem

- Particularly acute when testing thousands of SNP but also relevant in single SNP analysis
- From a *frequentist* perspective,
 - If we fix the overall type-I error rate at $\alpha=5\%$.
 - If we want all tests should generate together less than α false positives and
 - If we have L SNP,
 - If SNP are considered independent (not true!)
- we should use a per-SNP significance level of α' such that:

Multiple testing

$$\alpha = 1 - (1 - \alpha')^L \quad \text{so} \quad \alpha' = \frac{\alpha}{L}$$

- Known as the **Bonferroni correction**
- For $L=1 \text{ million}$ we have $\alpha' = 5 \cdot 10^{-8}$
- This is conservative because many SNP are tightly linked (high LD)
- Many other procedures for controlling type-I error exist

Another Bonferroni!

- Use Bonferroni with a corrected n , the number of effective SNPs
 1. Compute correlation matrix for genotype codes (AA = -1, AG = 0, GG = 1) of n SNPs
 2. Compute n eigenvalues, λ_i (principal components) and their variance, $v = \Sigma(\lambda_i - 1)^2 / (n - 1)$.
 3. $n_{\text{eff}} = n[1 - (n - 1)v/n^2]$
- Can be done easily with R language

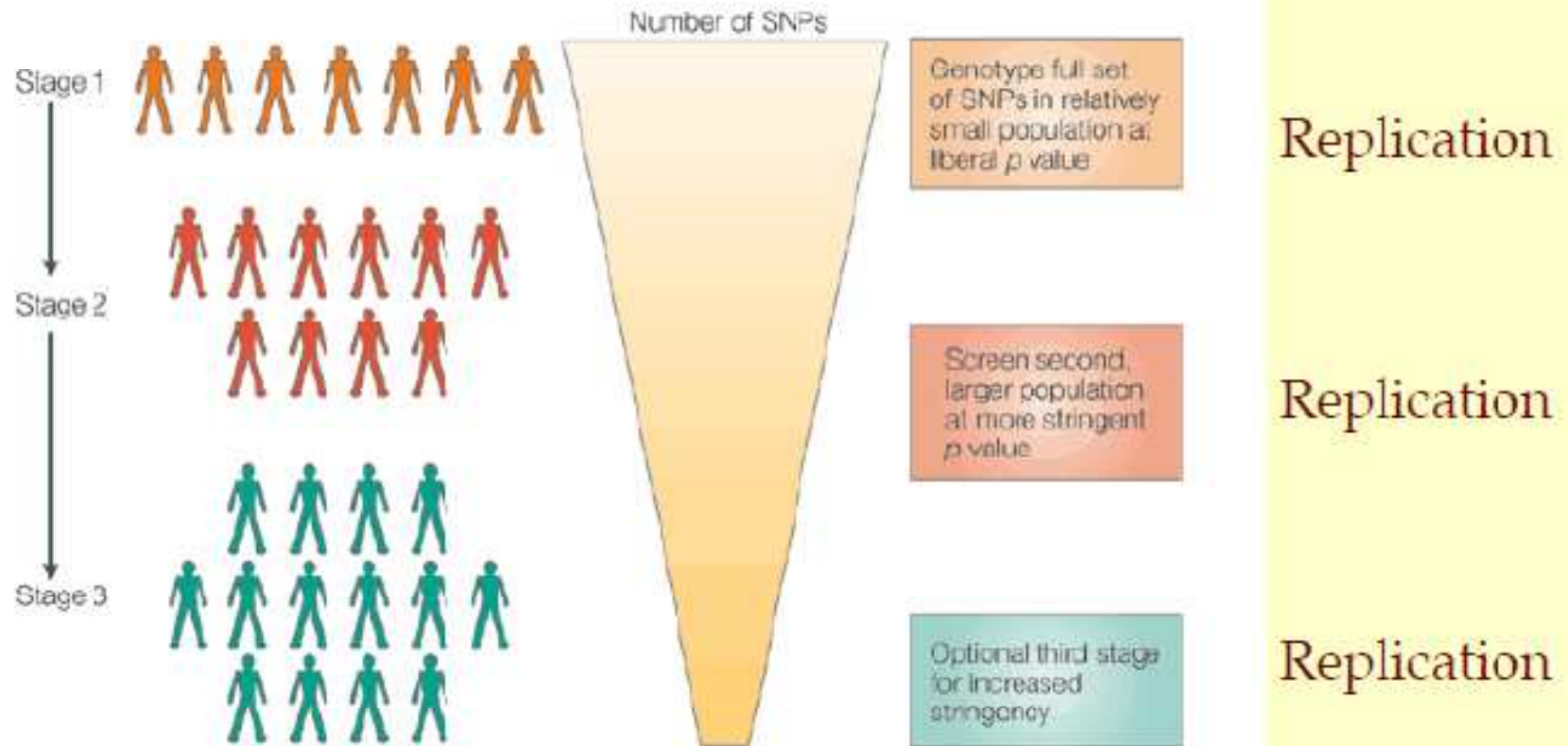
Multiple testing: permutation

- Compute p-values using permutation:
- **Randomize phenotype labels over individuals** while retaining genotypes (the LD structure is conserved but the association with phenotype is broken)
- **Repeat this many times and analyse all the datasets**
- Obtain p-value for each dataset and each SNP
as the proportion of test values that are greater than the observed initial test (with original data)
- **Easily implemented in R language**
- **Computationally demanding (for 1 SNP, a sample of 200:200 and on a PC, 10,000 permutations take 2" so 1 million SNP this gives 4 years!)**

The importance of replication

- Use an **independant sample** (preferably genotyped in a different platform) to confirm an association reported in an initial study
- To not counfound with **cross-validation**:
splitting a sample in two subsets one used to search for association and the other to check the initial findings

Multistage designs



Nature Reviews | **Genetics**

Hirschhorn & Daly Nat. Genet. Rev. 6: 95, 2005
NCI-NHGRI Working Group on Replication Nature 447: 655, 2007

Table 2. Examples of Multistage Designs in Genome-wide Association Studies^a

Stage	3-Stage Study ^b		4-Stage Study ^c	
	Case Participants/ Control Participants	SNPs Analyzed	Case Participants/ Control Participants	SNPs Analyzed
1	400/400	500 000	2000/2000	100 000
2	4000/4000	25 000	2000/2000	1000
3	20 000/20 000	25	2000/2000	20
4			2000/2000	5

Abbreviation: SNP, single-nucleotide polymorphism.

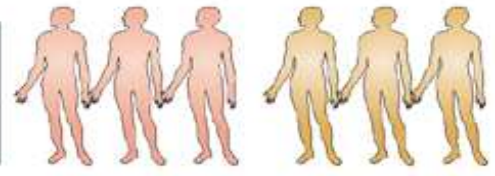
^aBased on hypothetical data.

^bFive SNPs associated with disease.

^cTwo SNPs associated with disease.

Large cohort of cases and controls ($n > 1,000$)

- Matched for confounding variables, such as race, ethnicity and sex
- Stratified in order to maximize signals



Microarray-based SNP genotyping

- ~1 million random marker SNPs or ~25,000 risk-enhancing SNPs (for example, nsSNPs)



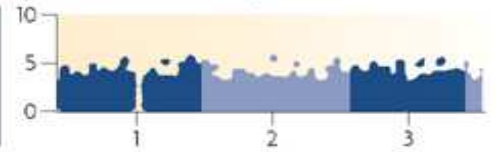
Derivation of haplotypes

- Predicated on International HapMap



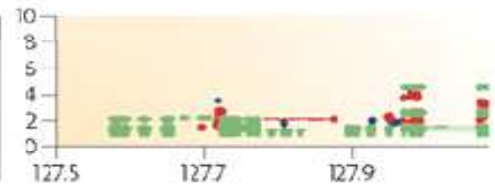
Detection of association signals

- χ^2 or similar test
- Uncorrected $P < 10^{-7}$ or false discovery rate-like correction



Fine mapping of association signal (see FIG. 2)

- Directed genotyping of additional SNPs in region
- Fine mapping of LD in region of association
- Empirical derivation of haplotypes
- Examination of effect of stratification, if available



Replication of association

- Large independent cohort: of cases and controls ($n > 1,000$)
- Genotyping of nominated candidate SNPs (<20)
- χ^2 or similar test; replication of initial signal

Genotypes	CC	AA	CA	Total
Cases observed	59	27	98	184
Controls observed	60	89	36	185
Total	119	116	134	369

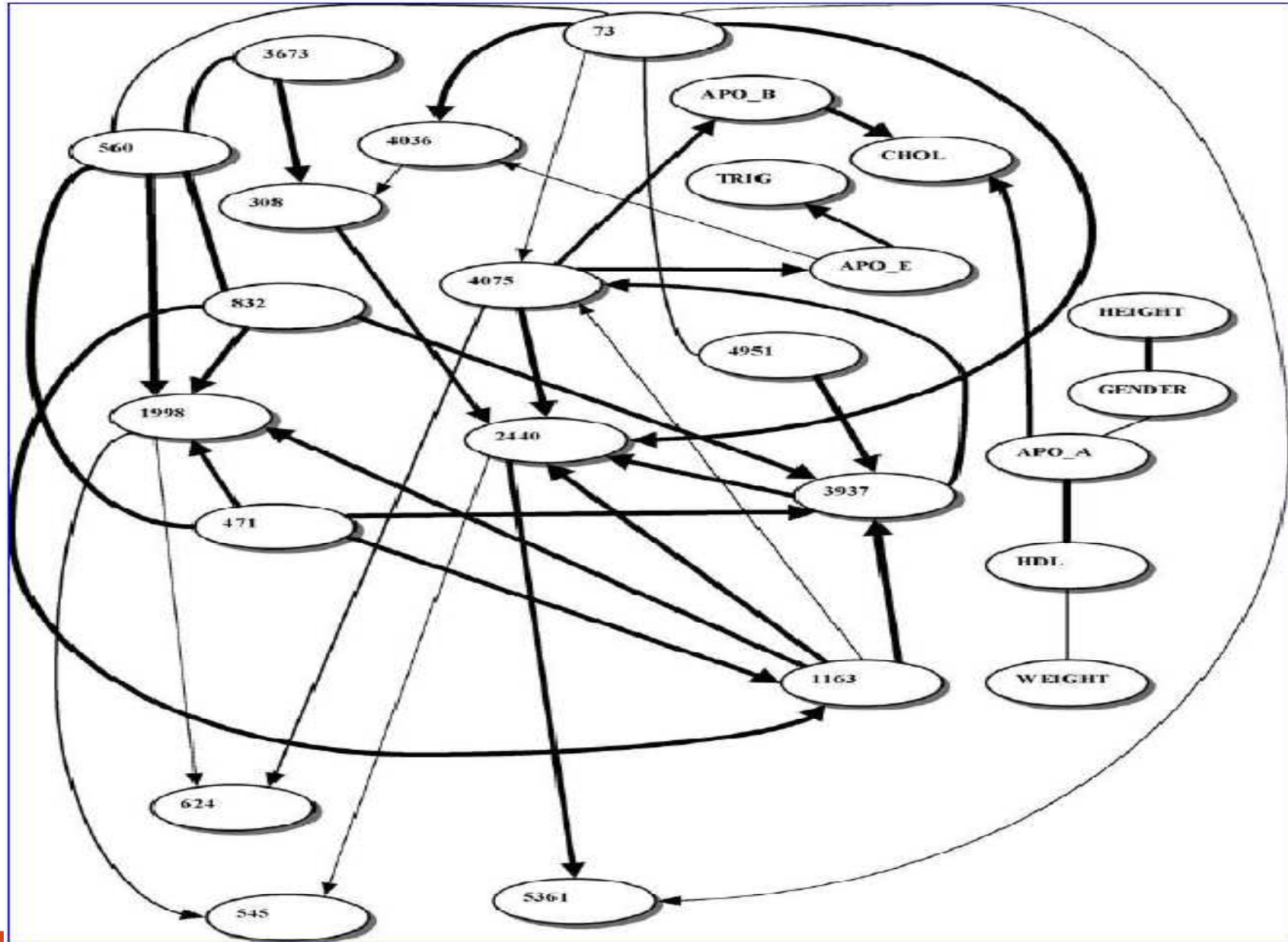
Biological validation of association

- Identification of risk-enhancing variant
- Examination of functional consequence of variant
- Determination of mechanism of risk-enhancement



Conclusion: The future

- To complex disease, complex analyses
- We still need powerful statistical methods that analyze many variants simultaneously for their individual effects and joint contribution to disease risk
- Some issues, such as stratification, will be banished with relatedness methods
- Bayesian methods and graphical bayesian models are becoming very attractive for GWAS data analysis



Recommended readings

- Nature Reviews Genetics
- Balding J, 2006. 7: 781-791
- Wang et al, 2005. 6: 109-117
- Stephens and Balding, 2009, 10: 681-690