

# Genetic Association Studies

## Statistics for Genomic Data Analysis

### Lecture 11

<http://moodle.epfl.ch/course/view.php?id=15271>

# What is a polymorphism ?

- A *polymorphism* is a difference in DNA sequence among individuals
- Genetic variations occurring in more than 1% of a population are considered useful for genetic *linkage analysis*
- *Single nucleotide polymorphisms* (*SNPs*) are the most common type of genetic variation among people
- Each SNP represents a difference in a single nucleotide : e.g. a SNP may replace a C with a T in a certain stretch of DNA
- Almost all common SNPs have only two alleles
- Within a population, the *minor allele frequency* of a SNP is the smaller of the two allele frequencies

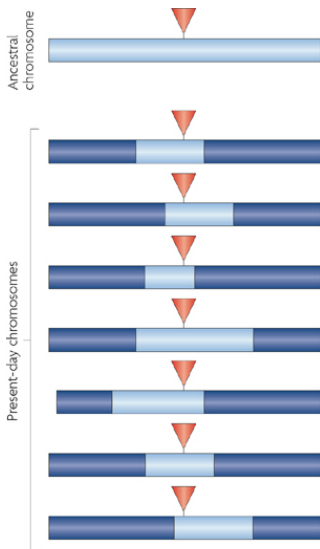
## A little history

- *1920s* : Blood group markers
- *1960s* : HLA markers
- *late 1970s – early 1980s* : RFLP markers
- *1980s* :
  - linkage map using RFLP markers [many not very polymorphic]
  - applications to Mendelian traits
  - Lander-Botstein (1986) : most human traits follow 'complex' modes of inheritance, suggested LD (association) mapping
- *Late 1980s – mid 1990s* : PCR-based microsatellite markers (SSRs) [highly polymorphic]
- LD mapping most often done as a followup to linkage to narrow the genomic region
- Most genome scans were linkage studies, but for complex traits usually difficult to find 'major genes' [genes with large effects] → *Wanted new approaches*

## Linkage disequilibrium (LD)

- A fundamental notion in *association mapping* is that of linkage disequilibrium (LD) between a genetic marker and the locus that affects the trait under study
- LD is the *nonrandom association of alleles* at two (or more) loci
- *Loci are in LD* when combinations of alleles occur either more or less frequently than expected from random formation of haplotypes based on marginal allele frequencies
- LD may be due to selection, random genetic drift, co-ancestry ...
- Not necessarily due to linkage (and therefore also sometimes referred to as *gametic disequilibrium*)
- There are several measures available to quantify the strength of LD

## LD around ancestral chromosome (Kruglyak Fig. 1)



## Example : LD in a bi-allelic system

Table: Frequencies of co-occurrence in a bi-allelic, two-locus system

	<i>B</i>	<i>b</i>	
<i>A</i>	$p_{AB}$	$p_{Ab}$	$p_A$
<i>a</i>	$p_{aB}$	$p_{ab}$	$p_a$
	$p_B$	$p_b$	

- Under independence,  $p_{AB} = p_A p_B$  (similarly for all combinations)
- Some measures of LD:
  - $\delta = p_{AB}p_{ab} - p_{Ab}p_{aB}$
  - $D = p_{AB} - p_A p_B$
  - $D' = D/D_{\max}$ , where  $D_{\max} = \min(p_A p_b, p_a p_B)$  for  $D > 0$ ,  
 $D_{\max} = \min(p_A p_B, p_a p_b)$  for  $D < 0$
  - $r^2 = \frac{D^2}{p_A p_a p_B p_b}$

## Continuing the stroll down memory lane ...

### 1990s

- *Risch-Merikangas, 1996* : power assoc > power linkage
- *Linkage* : good for low-frequency, large effects
- *Genomewide association* : good for high-frequency, small-effects
- *Lander* : Common disease - common variant hypothesis
- Direct vs indirect mapping approach
  - *Direct* : test all functional variants
  - *Indirect* : use a dense set of markers and do LD mapping modes of inheritance, suggested LD (association) mapping
- *late 1990s* : SNP Consortium (consortium of pharmas + Wellcome Trust) founded to identify at least 100,000 SNPs

# Mapping by association (Kruglyak Figs. 2, 3)

## a Direct:

catalogue and test all functional variants for association

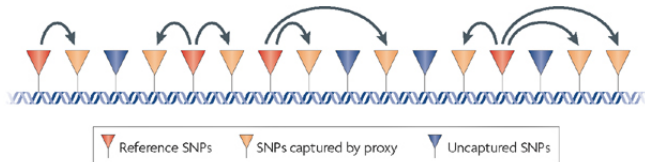


## b Indirect:

use a dense SNP map and test for linkage disequilibrium



Nature Reviews | Genetics



Nature Reviews | Genetics ▶



# Entering into this millenium !

- SNPs now preferred marker for LD mapping
  - not as polymorphic as microsatellites, *BUT* :
  - highly abundant, can create map with a density not achievable by other marker types
  - easy to genotype in a high throughput manner
  - low mutation rates

## Genetic association study

- Most commonly, compare genetic make-ups of cases and controls
  - *Candidate genes*, selected based on knowledge of PT
  - *Genetic markers*, e.g. SNPs
- Would like to find narrow regions of the genome associated with PT
- Markers must be spaced densely enough to be in LD with the (potentially disease-associated) variants that are not genotyped ( $\sim .5$ –1 million SNPs, depending on pop. origin)
- The essential idea is that a marker in strong LD with a disease locus is expected to be located nearby
- Reasons explaining observed associations :
  - *chance* or *artifact* (e.g. confounding, selection bias)
  - *LD* between marker locus and another locus that directly affects PT expression
  - *allele directly affects PT expression* (causal)

## NIH : GWAS definition

- A *genome-wide association study* (GWAS) is defined as any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition
- To meet the definition of a GWAS, the density of genetic markers and the extent of linkage disequilibrium should be sufficient to capture (by the  $r^2$  parameter) a large proportion of the common variation in the genome of the population under study, and the number of samples should provide sufficient power to detect variants of modest effect
- The trait can be qualitative or quantitative trait
  - different study designs
  - different analysis method

## Typical GWAS has 4 parts

- 1 *Select (many) individuals* with the disease/trait and suitable comparison group
  - 2 *DNA* isolation, genotyping, and quality assessment
  - 3 *Statistical tests for associations* between the SNPs passing quality thresholds and the disease/trait
  - 4 *Replication* of identified associations in an independent population sample and/or examination of *functional implications* experimentally
- 
- Common study designs :
    - Case-control design
    - Parent-offspring trio design
    - Cohort study
    - Multistage study

## Case-control study

- Usually *population-based*, *retrospective*
- Comparison between group with the phenotype (PT) of interest (e.g. disease) and group without
- Outcome is measured before 'exposure' (e.g. before genotyping)
- Controls selected on the basis of not having the PT
- Good for *rare PTs*
- Relatively inexpensive
- Smaller numbers required than for (prospective) cohort study
- Faster to complete than cohort study
- Prone to *bias*
  - selection bias
  - recall/retrospective bias
- Care needed to avoid *confounding*

## Parent-offspring trio study

(Parent-offspring) *trio design* : affected case plus both parents

- phenotype *offspring* (don't need parent phenotypes)
- genotype *trio*
- *TDT* : test for linkage in the presence of association (composite null)
- compare transmission of alleles from heterozygous parents to offspring
- not susceptible to effects of population stratification
- sensitive to genotyping error → need high quality genotyping

# Transmission Disequilibrium Test (TDT)

Table: *Combinations of transmitted and untransmitted marker alleles*

Transmitted	Non-transmitted		Total
	$M_1$	$M_2$	
$M_1$	$a$	$b$	$a + b$
$M_2$	$c$	$d$	$c + d$
Total	$a + c$	$b + d$	$2n$

- TDT is just the *McNemar test*, a test on  $2 \times 2$  table for the difference between *paired proportions* (e.g., in studies where patients serve as their own control, or with 'before and after' design)
- The classic TDT statistic is

$$\chi^2_{TD} = (b - c)^2/n,$$

## Other study designs

### ■ *Cohort study*

- collect baseline information in a large number of individuals
- more expensive and time-consuming than case-control
- participants may be more representative than case-control
- often include a vast array of health-related characteristics and exposures for which genetic associations can be sought ('fishing expedition')

### ■ *Multistage designs*

- scan one set of individuals
- followup smaller number of loci in second set of individuals



## Selecting study participants

- Usual caveats apply
- Need careful trait assessment (avoid misclassification)
- Controls should be from same population as cases, and be at risk of developing the trait
- Not uncommon (but imho not good) for controls to be selected from blood donors
  - quality difficult to ensure
  - followup difficult/impossible
- Assessment of comparability of cases/controls
  - Adjust for important differences in the analysis where possible
- Assess population structure (crude but usual to test for deviation from HWE)
- Adjust/account for substructure

# BREAK

# Population substructure in case-control study

- Substructure introduces deviation from HWE (but testing for deviation from HWE not very powerful)
- To analyze case-control data, assume
  - *Genetic homogeneity* : all individuals have same risk
  - *Statistical independence* between individuals
- **Substructure violates analysis assumptions**
- Mini-controversy : *is the problem exaggerated ??*
  - one argument is that this is flawed epidemiological practice rather than just poor genetic matching
  - when addressed as part of study design, heterogeneity is usually not seen to be extensive
  - practically speaking : your grant app/paper/etc will be rejected if you don't address this issue !

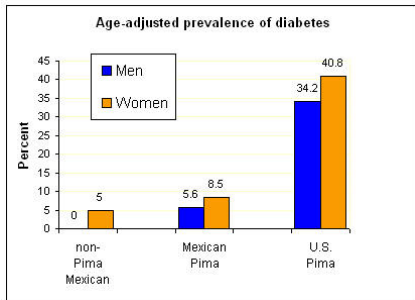
# Population substructure : why should we account for it ?

Violation of the assumptions can adversely affect inference :

- Can severely bias association studies
- Spurious (false positive) associations due to confounding
  - differences in disease prevalence between cases and controls along with variations in allele frequency between groups
- False negative associations (e.g. Simpson's paradox) – admixture can also appear to mask, change, or reverse true genetic effects
- *Simpson's Paradox* refers to the reversal of the direction of a comparison or an association when data from several groups are combined to form a single group

## Example of spurious association : NIDDM in Pima Indians

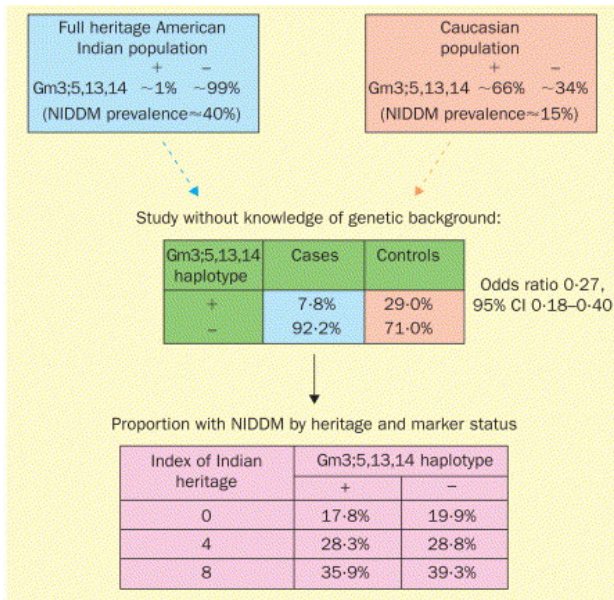
The U. S. Pima Indians have the highest reported prevalence of NIDDM of any population in the world



## Example, contd

- Association study of HLA haplotype (Gm3;5,13,14) and NIDDM in U. S. Pima Indians
- Confounding due to admixture of Caucasian and Pima Indian ancestry
- Failure to control for ethnic origin introduced bias :
  - Diabetes prevalence and frequency of the haplotype both much higher in individuals of American Indian ancestry than in those of European ancestry
  - Association of Gm3;5,13,14 haplotype with reduced risk of NIDDM attributable to *ancestral population of origin* rather than to LD between the disease and marker loci
- Observed association disappeared when analysis restricted to full-heritage Pima Indians

## Example of spurious association : NIDDM in Pima Indians



## Avoiding population stratification

- *Test* whether cases and controls differ at unlinked markers
- *Match cases and controls* on ethnicity, geographic location, etc.
- Control for stratification with *family-based controls*
  - Lower power
  - Can also be difficult to recruit enough family members
- Account for stratification in analysis
  - *Genomic control (GC)* (lower power)
  - Structured association (STRUCTURE) : define underlying subgroups based on a set of genomic markers, test for disease association by combining subgroup association results
  - *EIGENSTRAT* : based on PCA
  - *Population Stratification Association Test (PSAT)* : permutation test-based approach



## Genotyping and QC in GWAS

- GT error can be an important cause of spurious association
- (Sampson, Zhao : test using signal rather than called GT)
- Assess GT quality
  - per sample (individual)
  - per SNP
- Per sample checks
  - sample identity checks to avoid sample mix-ups
  - minimum rate of successfully called GTs (e.g. >90%)
- After bad samples removed, per SNP checks
  - high SNP call rate (e.g. > 95%)
  - minor allele frequency (MAF) > 1%
  - severe violations of HWE
  - Mendelian inheritance errors (in family studies)
  - concordance rates in duplicate samples (e.g. > 99.5%)
- Standard followup to use a different technology to re-genotype the most strongly associated SNPs

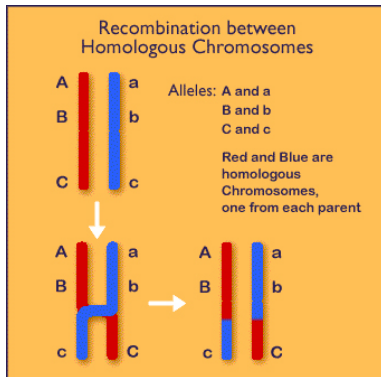
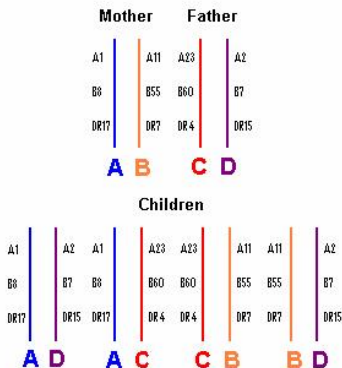
# GWAS analysis

- Most studies test each *single locus* one at a time
  - highly correlated test results
- **Avoid** assessing association between case/control status and *allele*
  - invalid when HWE is violated in a combined population
- **Do** assess association between case/control status and *genotype*
- (Cochran-)Armitage trend test most commonly done
- Multiple testing
  - many studies have used Bonferroni correction, although it is well-known to be conservative in the presence of LD (positive correlation between test results)
  - could use FDR instead
  - perm tests account for LD, can have high computational cost
  - another possibility is estimating the *effective number of independent tests* and Bonferroni-adjust for that number

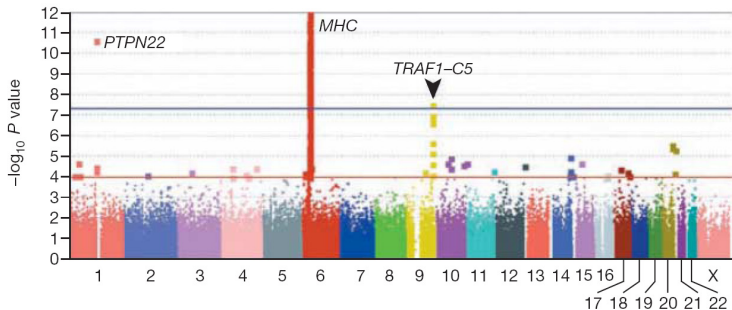
## More on analysis

- Several statistical methods for association mapping (including GLMs like logistic regression) require *genetic model of inheritance*
- For instance, in a Cochran-Armitage test or score statistics from logistic regression, an additive model can be imposed by giving genotype *weights* 0, 1 and 2, depending on the number of copies of the minor allele
  - assuming a model is more powerful when (at least approximately) true
  - can have very low power when the true model is different
- *Haplotype-based tests*
  - usually done as followup for highly associated regions
  - haplotype block as the smallest unit
  - can use tagging SNPs as surrogates for other markers in a block
  - reduced number of markers needed
- Emerging area : data mining for *interaction analysis*

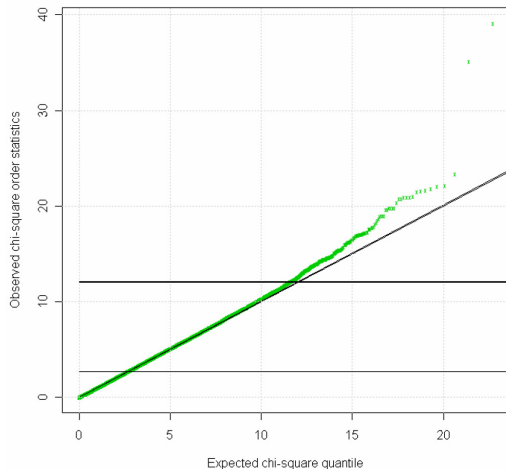
# Haplotype



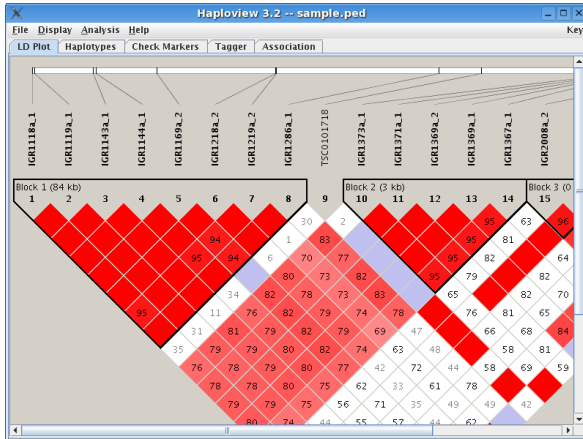
## Presentation of results : Manhattan plot



## Presentation of results : QQ plot for $p$ -values



# Presentation of results : LD display



## Replication : criteria

- *Sample large enough* to distinguish proposed effect from no effect
- Carry out in *independent data sets*
- Analyze same PT (or very similar one)
- Study *similar population*, describe important differences
- *Similar magnitude of effect and significance* should be demonstrated, in the same direction, with the same SNP /SNP high LD ( $r^2 \approx 1$ )
- Statistical significance should first be obtained using the genetic model reported in the initial study
- When possible, a joint or combined analysis should lead to a *smaller p-value* than the original
- *Strong rationale* for selecting replication SNPs, including LD structure, putative functional data or published literature
- Replication reports should be as *detailed* as initial study report



# Replication and functional studies

- Try to replicate result in independent population
- Usually start out trying to be as similar as possible to the original study
- Branch out : different phenotypes, population, study design
- *Lack of reproducibility*
  - population stratification
  - phenotype differences
  - selection biases
  - *genotyping errors*
  - ...
- *Functional studies* to elucidate disease mechanisms

## Limitations of GWA studies

- Potential for false-positive results
- Lack of information on gene function
- Insensitivity to rare variants and structural variants
- Requirement for large sample sizes
- Possible biases due to case and control selection and genotyping errors
- Difficulty identifying  $G \times E$  interactions due to limited information on environmental exposures and other non-genetic risk factors
- Difficult to assess reports (due to journal page limits, incomplete reporting)
- **false negatives**
- Clinical applications are still a (long ?) way off