Statistics for Genomic Data Analysis

- Genetic linkage, crossing over, recombination
- Genetic markers
- Genetic association
- Population sub-structure
- Genome-wide association studies (GWAS)

[NOTE : the part at the end about multiple testing is REVIEW ; we saw this already in Lecture 5b]

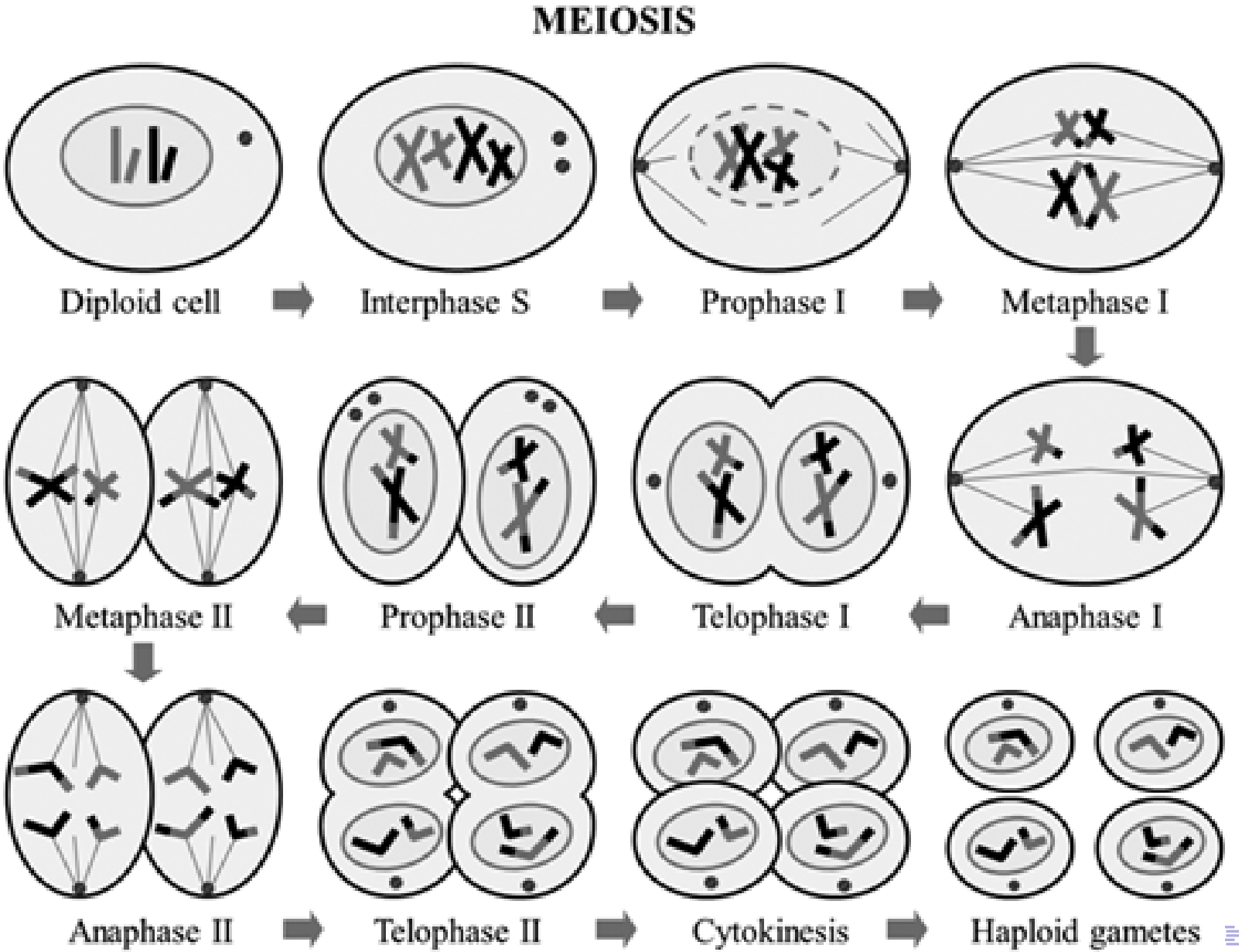
Genetic linkage

- Genes that are close together on the same chromosome are said to exhibit *linkage*
- Genes on nonhomologous chromosomes (also far apart on the same chromosome) *assort independently* during meiosis
- Linked genes, and hence the phenotypic characters they control, are *inherited together* because they are located on the same chromosome
- Modern understanding of genetic linkage came from the work of Thomas Hunt Morgan : showed that two recessive genes in *Drosophila melanogaster* : white eye (w) and miniature wing (m) are X-linked.
- Linkage is based on the frequency of *crossing over between the two genes*

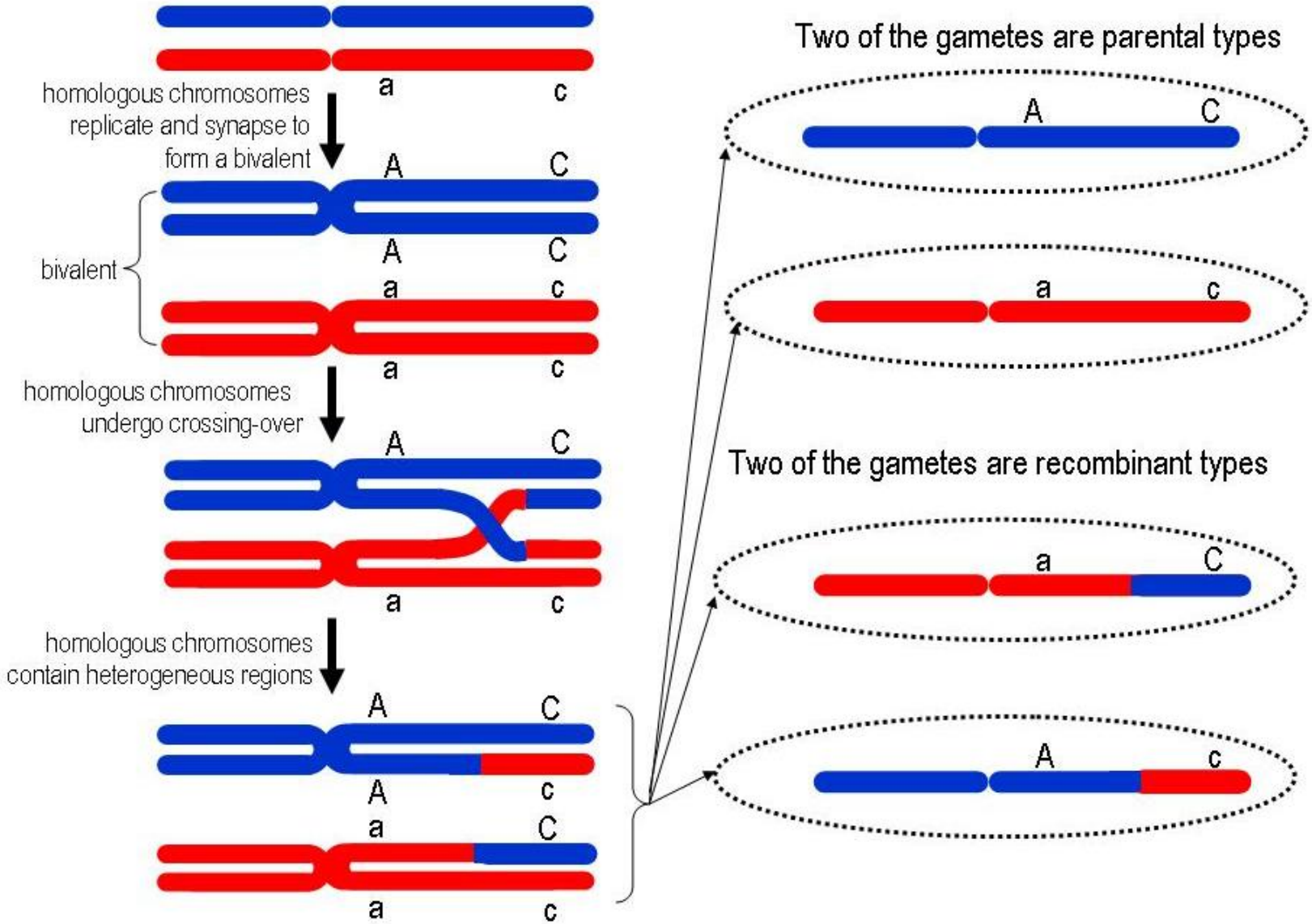
Crossing over

- Genes on the same chromosome travel through meiosis together : however, alleles of chromosomally linked genes can be *recombined by crossing over*
- During prophase of meiosis I, the double-chromatid homologous pairs (sister chromatids) of chromosomes cross over with each other and can exchange chromosome segments
- This process occurs at the *4-strand stage*
- Crossovers can result in *recombination and the exchange of genetic material* between the maternal and paternal chromosomes
- The recombinant gametes are those that differ from both haploid gametes that made up the original diploid cell (so differ from the parental gametes)
- Recombination creates *genetic diversity*

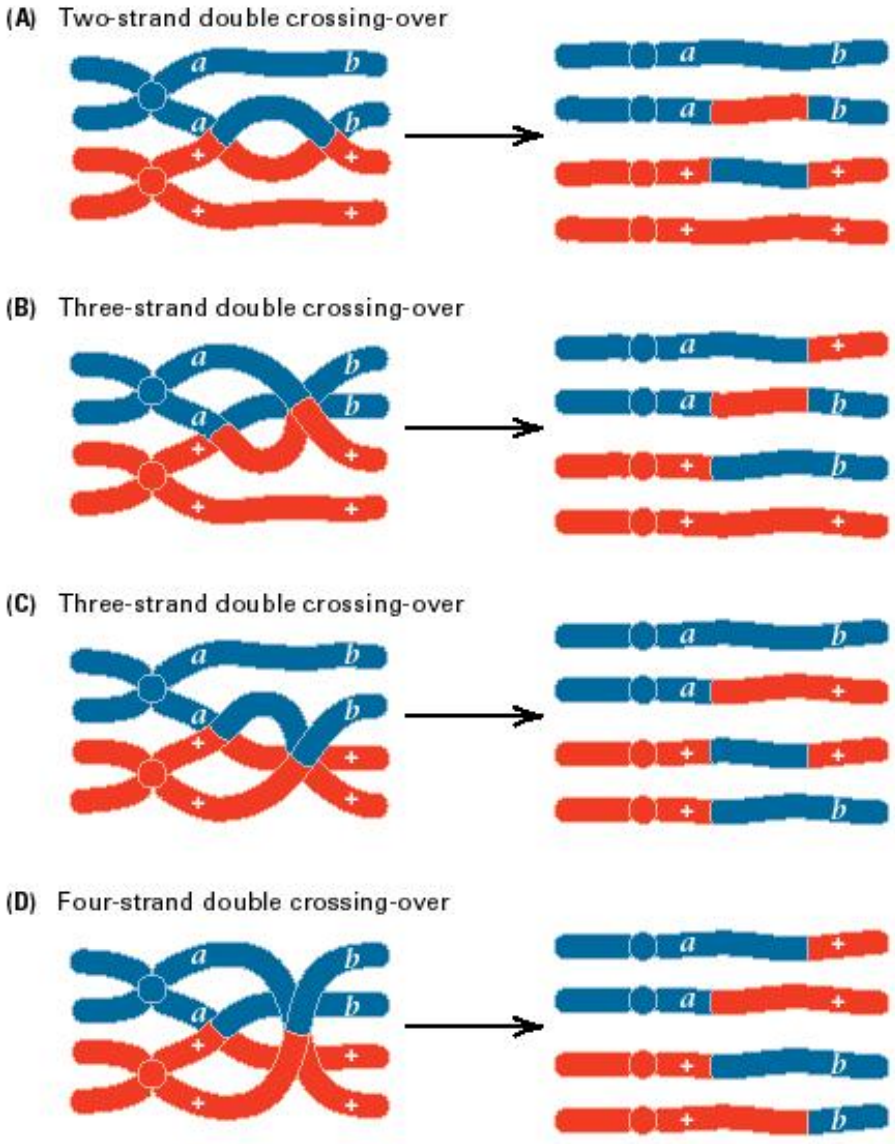
LD around ancestral chromosome (Kruglyak Fig. 1)



LD around ancestral chromosome (Kruglyak Fig. 1)



LD around ancestral chromosome (Kruglyak Fig. 1)



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

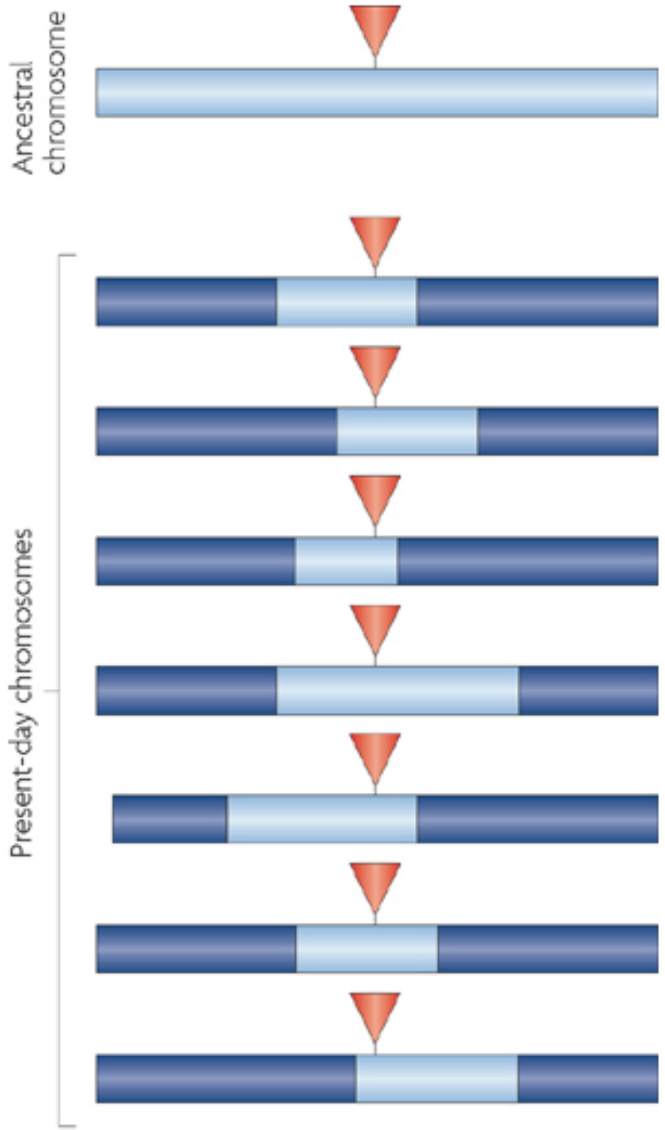
A little history

- *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*

	B	b	
A	p_{AB}	p_{Ab}	p_A
a	p_{aB}	p_{ab}	p_a
	p_B	p_b	

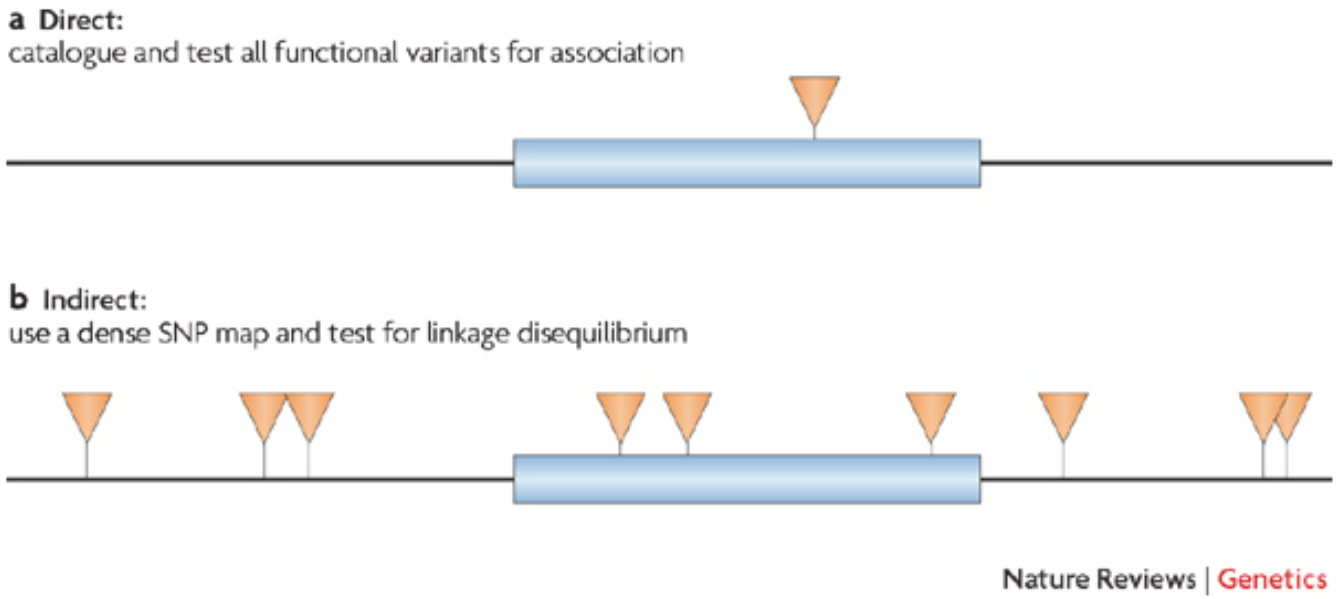
- Under independence, $p_{AB} = p_A p_B$ (similarly for all combinations)
- Some measures of LD:
  - $\delta = p_{AB}p_{ab} - p_{Ab}p_{aB}$
  - $D = p_{AB} - p_A p_B$
  - $D' = D/D_{\max}$, where $D_{\max} = \min(p_A p_b, p_a p_B)$ for $D > 0$, $D_{\max} = \min(p_A p_B, p_a p_b)$ for $D < 0$
  - $r^2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

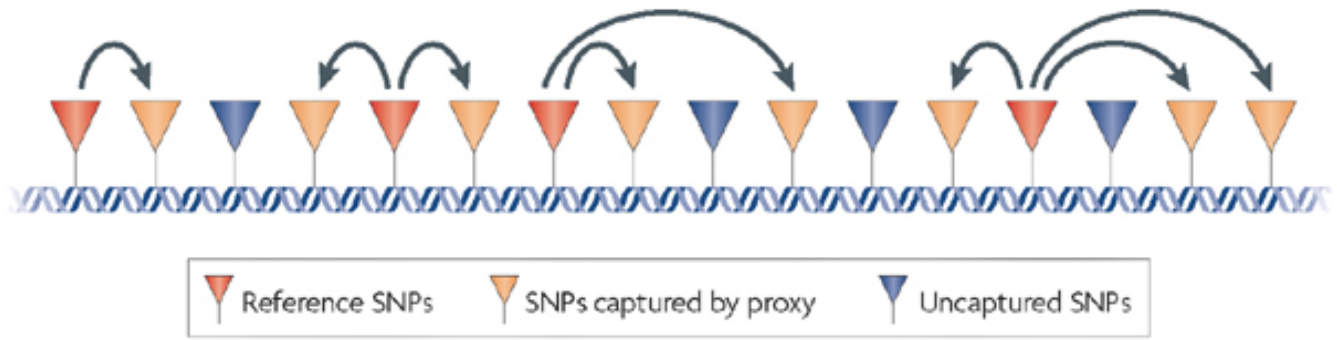
Continuing the stroll down memory lane ...

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





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*

	Non-transmitted		
Transmitted	M_1	M_2	Total
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 2×2 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

Hardy-Weinberg Equilibrium (HWE)

- Five assumptions :
  1. Random mating (with respect to the genetic locus)
  2. Infinite population size (*i.e.* no genetic drift, which reduces genetic variability)
  3. No mutations
  4. No genetic migration (permanent movement of alleles from one population to another, usually by dispersal of individuals)
  5. No natural selection
- Then for a two allele locus, with alleles A and a, where $p(A) = p$ and $p(a) = 1 - p = q$, the genotypes AA, Aa and aa occur in proportions p^2, $2pq$ and q^2

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 !

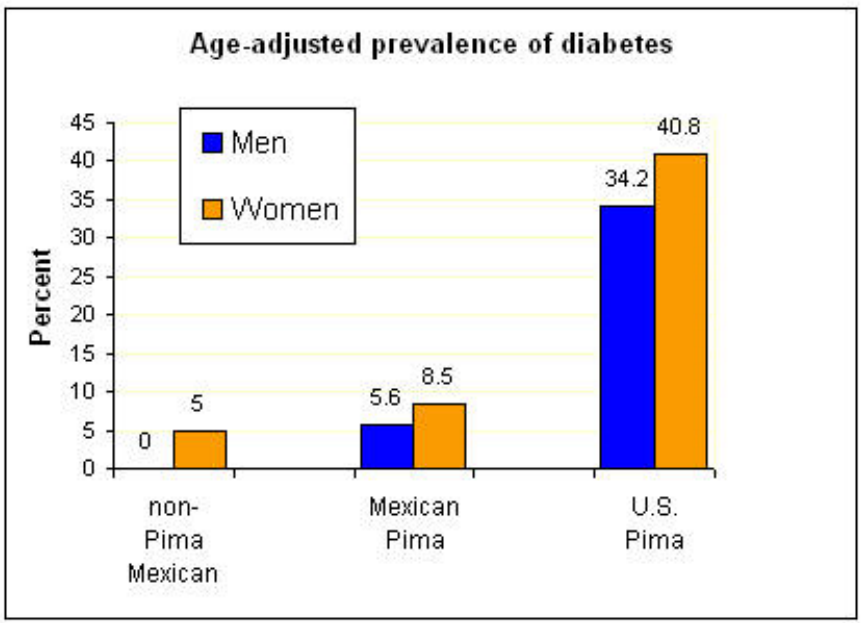
Population substructure : why should we account for it ?

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

Full heritage American Indian population	+	–
Gm3;5,13,14	~1%	~99%
(NIDDM prevalence≈40%)		

Caucasian population	+	–
Gm3;5,13,14	~66%	~34%
(NIDDM prevalence≈15%)		

Study without knowledge of genetic background:

Gm3;5,13,14 haplotype	Cases	Controls
+	7·8%	29·0%
–	92·2%	71·0%

Odds ratio 0·27, 95% CI 0·18–0·40

Proportion with NIDDM by heritage and marker status

Index of Indian heritage	Gm3;5,13,14 haplotype	
	+	–
0	17·8%	19·9%
4	28·3%	28·8%
8	35·9%	39·3%

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

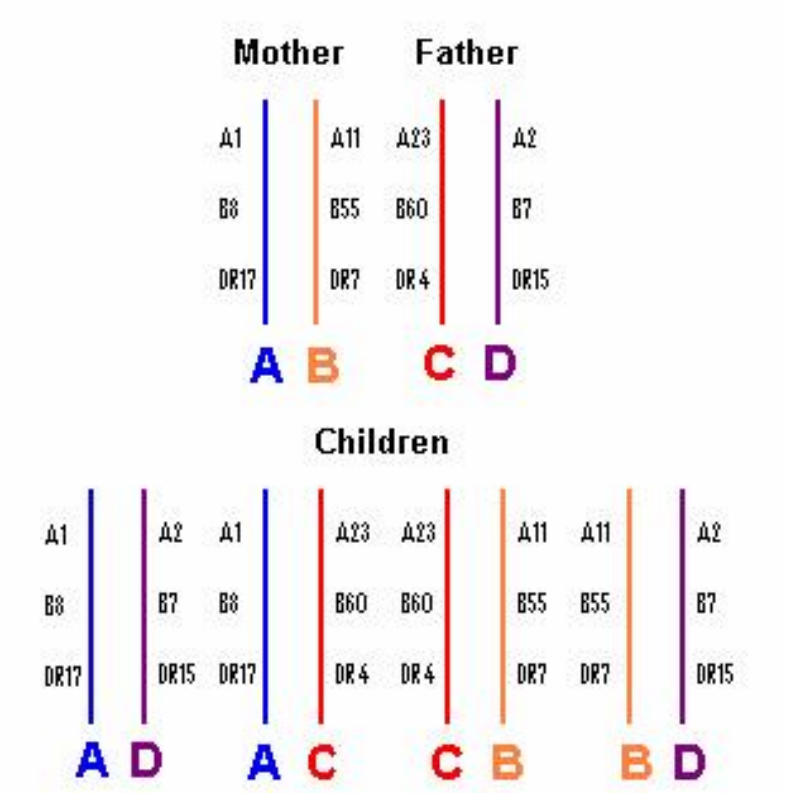
GWAS analysis

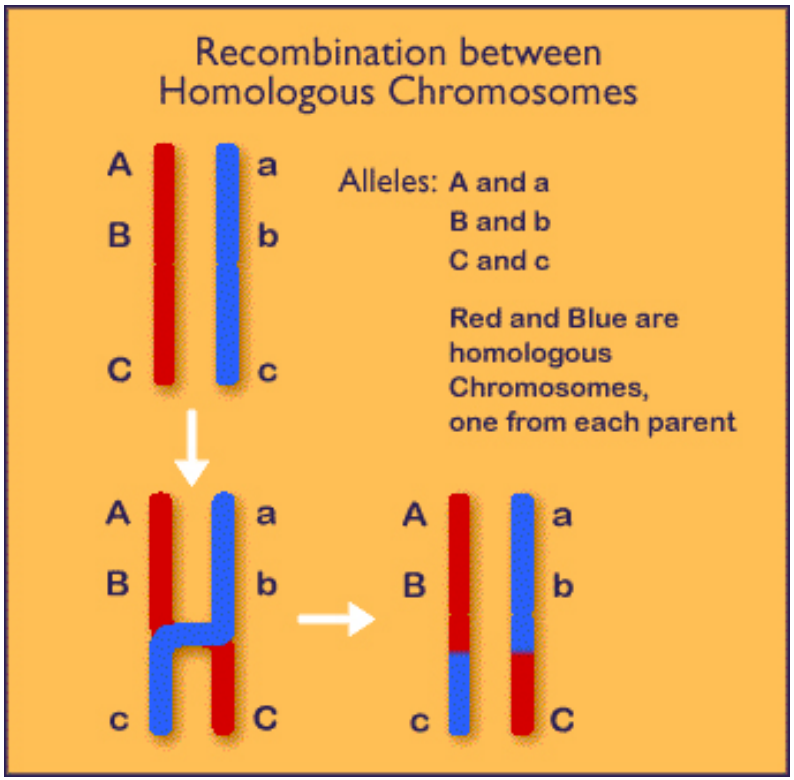
- Most studies test each *single locus* one at a time
  - highly correlated test results
- **Avoid** assessing association between case/control status and *allele*
  - invalid when HWE is violated in a combined population
- **Do** assess association between case/control status and *genotype*
- (Cochran-)Armitage 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

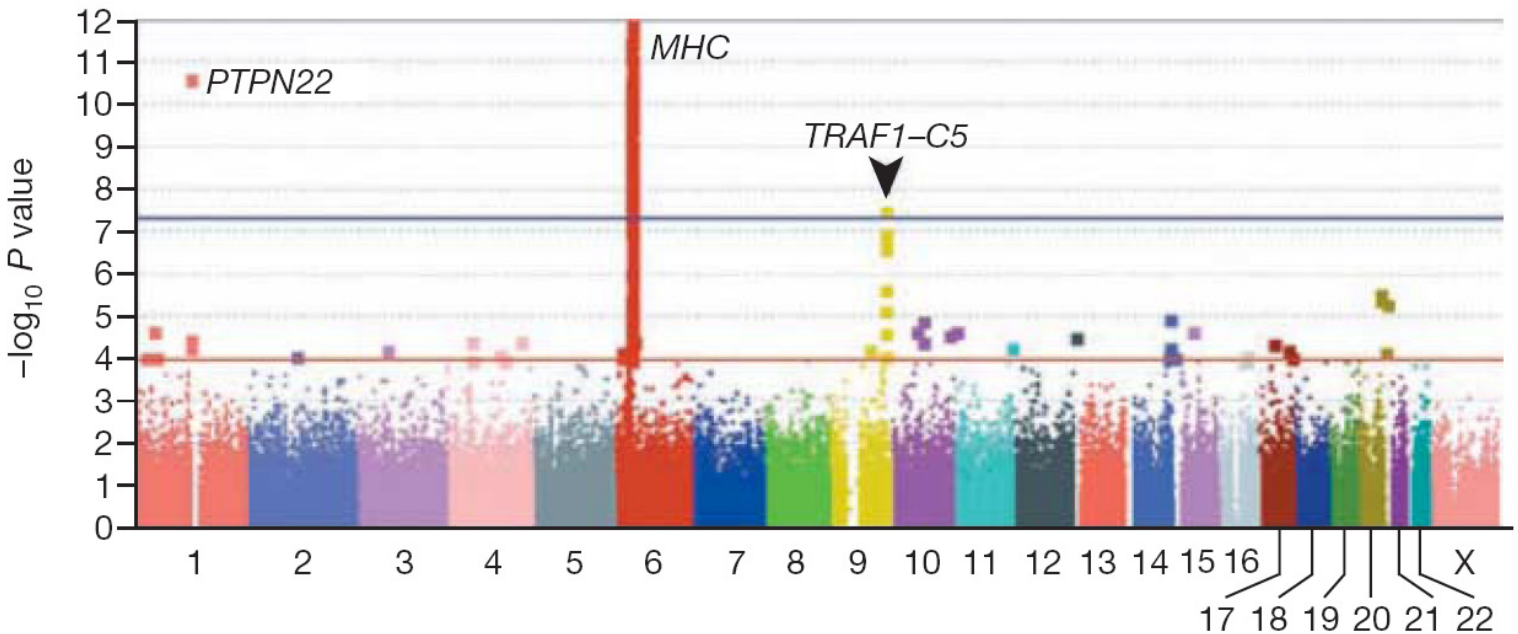
- Several statistical methods for association mapping (including GLMs like logistic regression) require *genetic model of inheritance*
- For instance, in a Cochran-Armitage test or score statistics from logistic regression, an additive model can be imposed by giving genotype *weights* 0, 1 and 2, depending 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

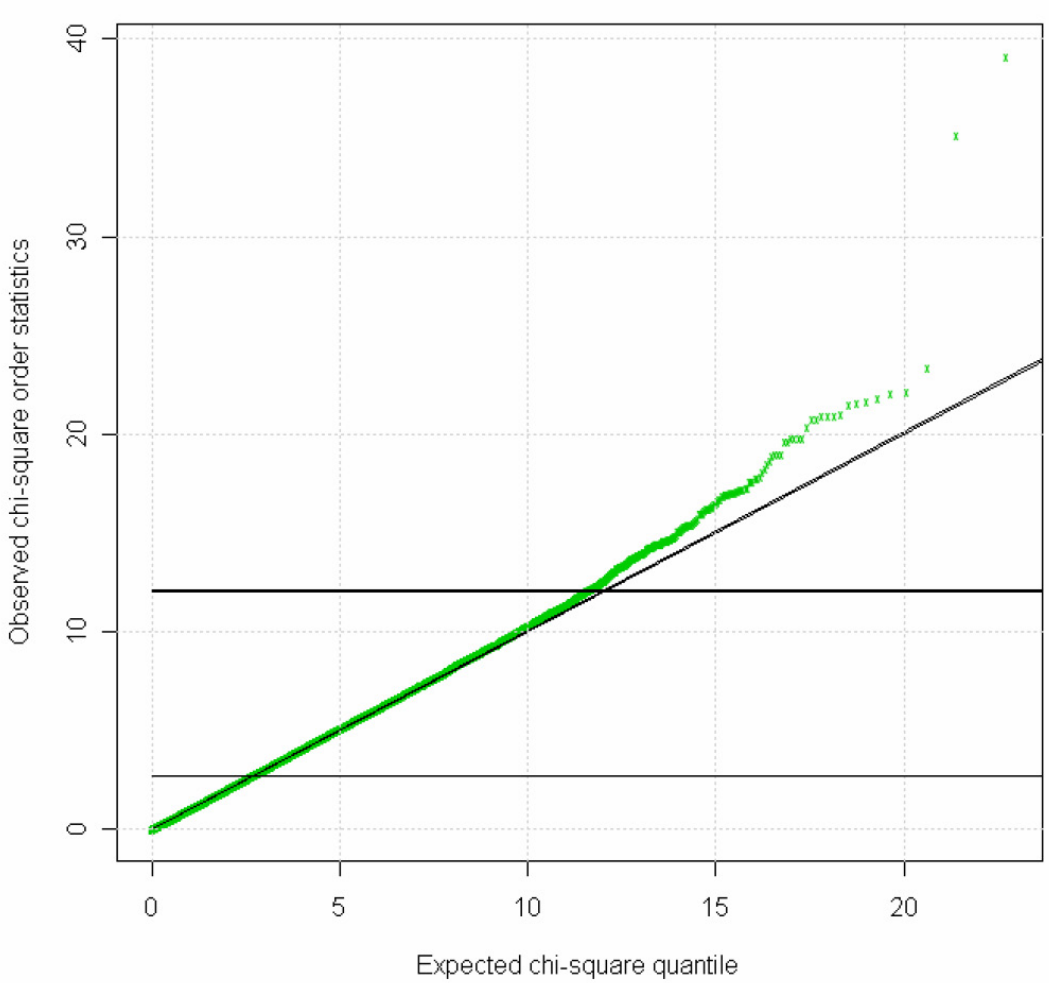

Mother
Father
A1
B8
DR17
A11
B55
DR7
A23
B60
DR4
A2
B7
DR15
A B C D
Children
A1 A2 B8 B7 DR17 DR15
A1 A23 B8 B60 DR17 DR4
A23 A11 B60 B55 DR4 DR7
A11 A2 B55 B7 DR7 DR15
A D A C C B B D


Recombination between
Homologous Chromosomes
A
B
C
a
b
c
Alleles: A and a
B and b
C and c
Red and Blue are
homologous
Chromosomes,
one from each parent

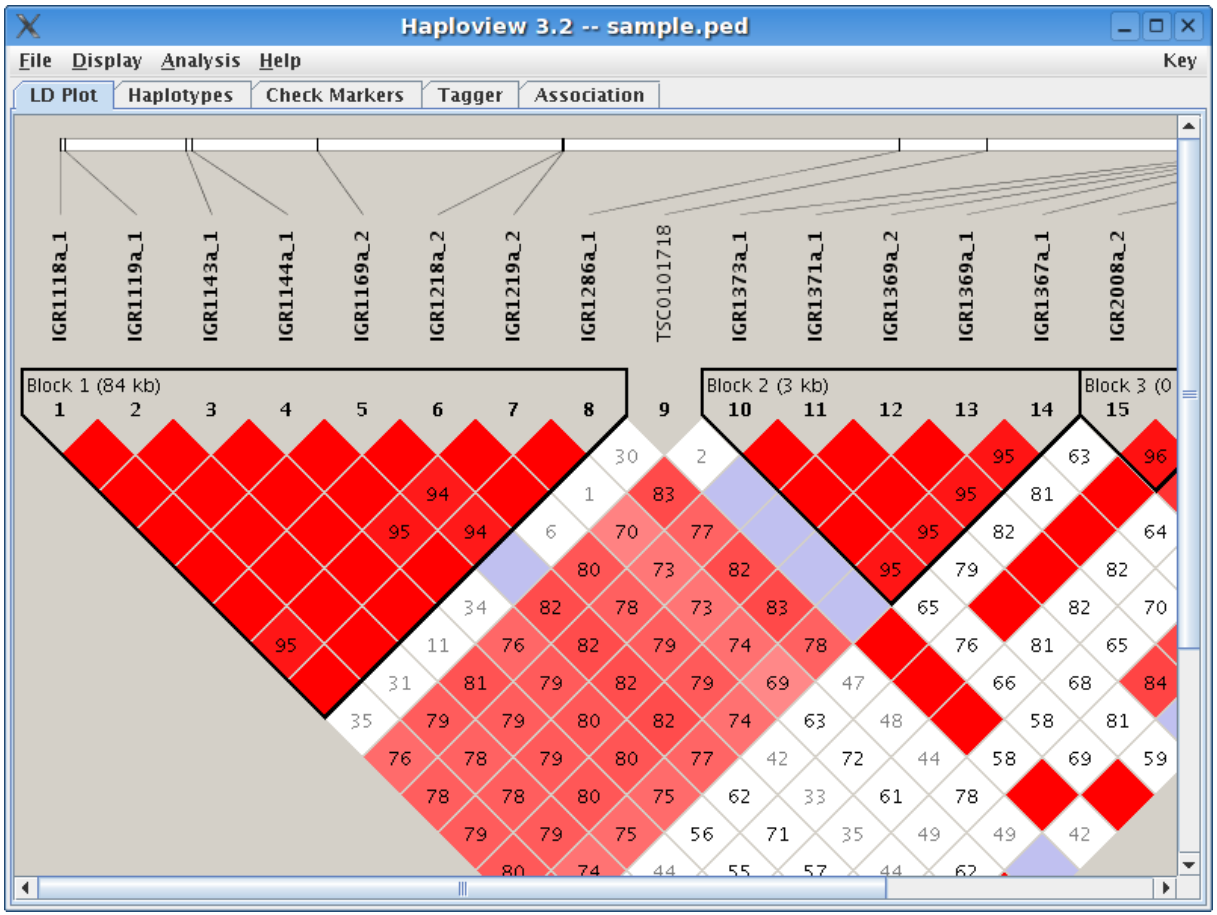
Presentation of results : Manhattan plot



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

BREAK

Locating a point in the plane

- We can describe the location of a point in the plane by saying how much we move in the horizontal (X) direction, then how much we move in the vertical (Y) direction
- As an example, think of describing how to get to some particular place from where you are (for example, how to get to CE 105 from MA 11)
- One way to do this is to say how far you go NORTH, then how far you go EAST

Variance-Covariance matrix

- Consider a data set consisting of p variables measured on n cases
- How the variables change together is summarized by the variance-covariance matrix (or by the correlation matrix)
- For a simple example (just 2 variables) :

```
> cov(head)
          [,1]     [,2]
[1,] 96.95061 54.48939
[2,] 54.48939 49.57918
```

```
> cor(head)
          [,1]    [,2]
[1,] 1.0000   .7859
[2,] 0.7859   1.0000
```

Principal Component Analysis (PCA)

- One aim of principal component analysis (PCA) is to *reduce the dimensionality* from p variables
- Try to explain the variance-covariance structure through *linear combinations (principal components)* of the (original) variables
- Another aim is to interpret the first few principal components in terms of the original variables to give greater insight into the data structure

More on PCA

- Each principal component (PC) accounts for a certain amount of the variation in the data
- The 1st PC is the linear combination that accounts for ('explains') the *most variation*
- Subsequent PCs account for as much as possible of the remaining variation, while being *uncorrelated* with earlier PCs
- *Aubergine*
- Where do these come from ?

What does this have to do with PCA ?

- Consider the variance-covariance matrix A
- The eigenvectors of A provide sets of coefficients defining p linear functions of the original variables
- **These functions are the PCs**
- If A has eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_p$, then the PCs have variances $\lambda_1, \lambda_2, \ldots \lambda_p$ and zero covariances

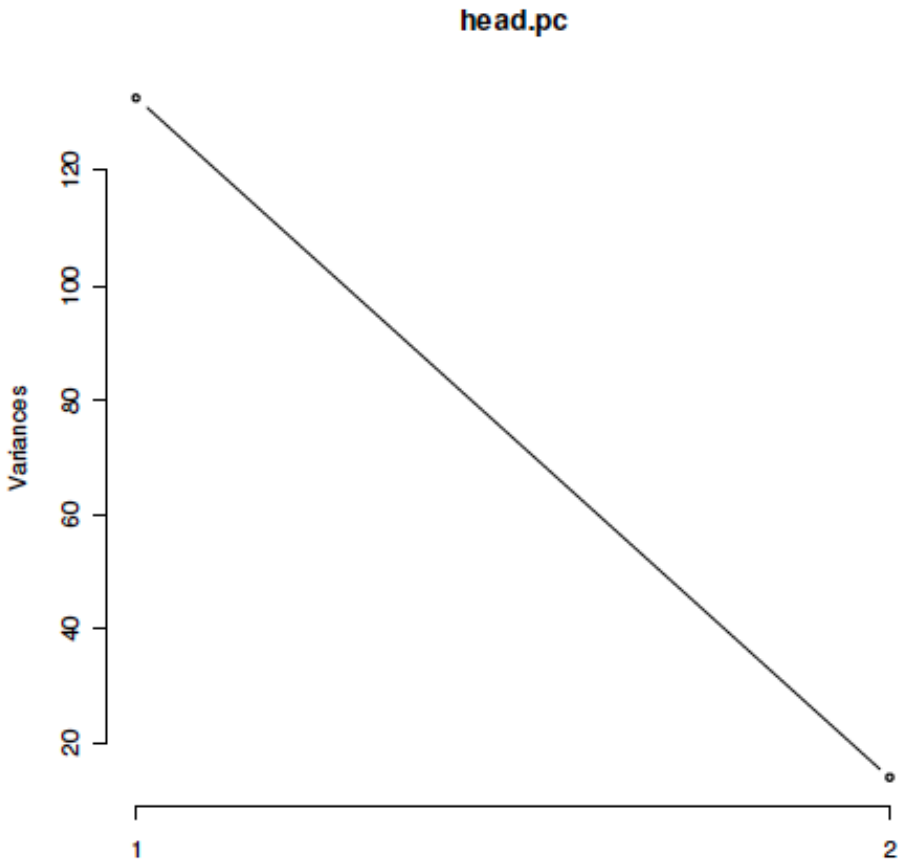
Cautions

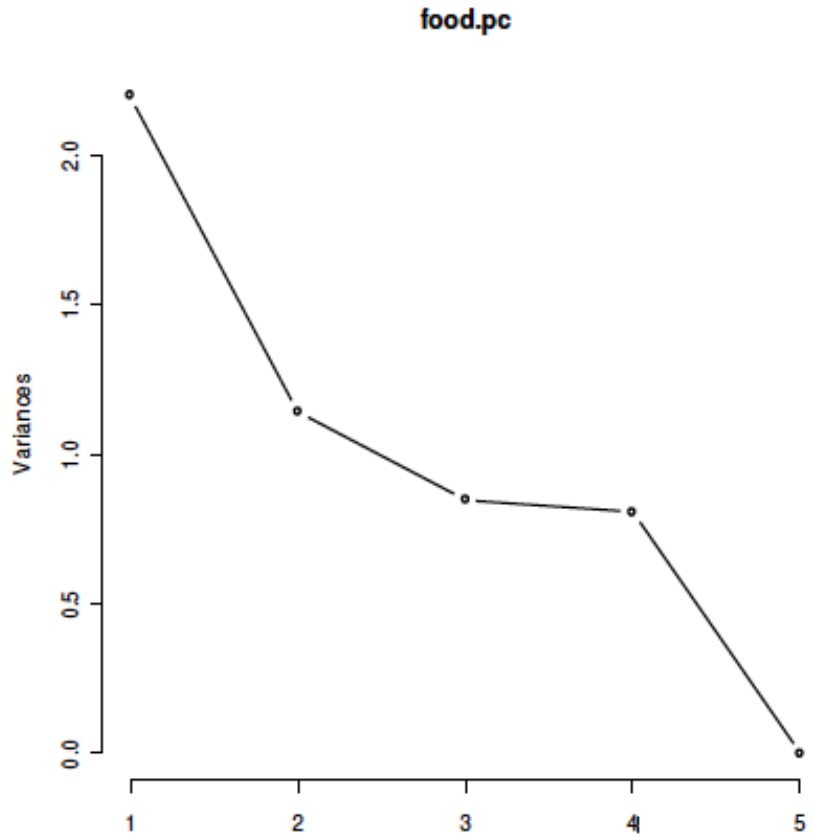
- Sometimes used as a method for *simplifying data* because PCs associated with smaller eigenvalues have smaller variances and might therefore be 'ignored'
- *This assumption requires caution*
- When variables are on *different scales*, it is customary to use the *correlation matrix* (rather than the covariance matrix)
- *These two formulations give different results :* the eigenvalues for the two matrices are not related in a simple way
- Theory not simple for correlation-based PCA

How many PCs ?

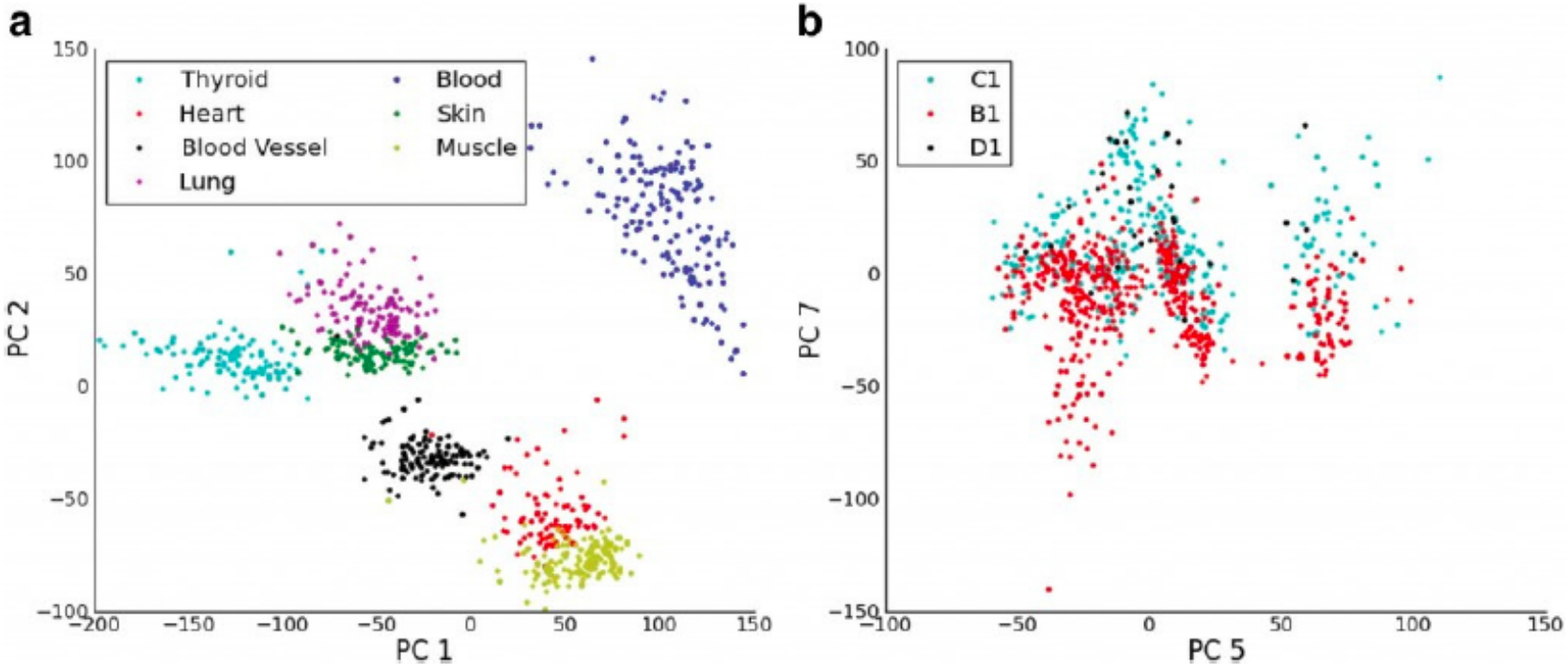
- There are a few ways to decide how many PCs to retain
- Some common methods are :
  - retain the number required to explain some percentage of the total variation (e.g. 90%)
  - number of eigenvalues > average (1 if correlation matrix is used)
  - look for 'elbow' in scree plot
  - compromise between these
- The scree plot shows proportion of variance (or just variance) explained by each component

R : scree plots





PCA to assess data quality



(a) *RNA-seq data projected onto PCs 1&2, where spot corresponds to a sample and color to tissue type. Samples from the same tissue cluster together.*

(b) *RNA-seq data projected onto PCs 5&7, now colored by enrollment center (C1, B1, D1). There is an obvious relation between PC 7 and center.*

Multiple testing problem

- Simultaneously test m null hypotheses
- Null hypothesis H_j : *no association between outcome measure j and the covariate*
- For high-dimensional data analysis (*e.g.* genomic data), there is a *large multiplicity issue*
- Increased chance of *at least one false positive*
- Would like some sense of how 'surprising' the observed results are

Hypothesis Truth vs. Decision : m tests

Truth \ Decision	# not rejected	# rejected	Totals
# true H	U	V (F +) Type I error	m_0
# non-true H	T Type II error	S	m_1
totals	W (= m - R)	R	m

Random Variables constants

Type I (false positive) error rates

- *Per-family Error Rate*

$$PFER = E(V)$$

- *Per-comparison Error Rate*

$$PCER = E(V)/m$$

- *Family-wise Error Rate*

$$FWER = p(V \geq 1)$$

- *False Discovery Rate*

$$FDR = E(Q), \text{ where}$$
$$Q = V/R \text{ if } R > 0\,; \quad Q = 0 \text{ if } R = 0$$

Strong vs. weak control

- All probabilities are *conditional* on which hypotheses are true
- *Weak control* refers to control of the Type I error rate only under the *complete null hypothesis* (*i.e. all* nulls true)
- *Strong control* refers to control of the Type I error rate under *any combination* of true and false nulls
- In general, *weak control* without other safeguards is *unsatisfactory*

Adjusted p-values (p^*)

- *Test level* (*e.g.* 0.05) does not need to be determined in advance
- Some procedures *most easily described* in terms of their adjusted p-values
- Usually *easily estimated using resampling*
- Procedures can be *readily compared* based on the corresponding adjusted p-values

A little notation

- For hypothesis $H_j, j = 1, \ldots, m$
  - observed test statistic : t_j
  - observed *unadjusted* (nominal) p-value : p_j
- Ordering of observed (absolute) t_j : $\{r_j\}$ such that
  - $| t_{r_1} | \geq | t_{r_2} | \geq \ldots \geq | t_{r_m} |$
- Ordering of observed (absolute) p_j : $\{r_j\}$ such that
  - $| p_{r_1} | \leq | p_{r_2} | \leq \ldots \leq | p_{r_m} |$
- Denote corresponding RVs by upper case letters (T, P)

Methods for obtaining p^*

- *Single-step* adjustment
  - p-values compared to a *predetermined value*
  - *same adjustment* for every p-value
- *Step-down* adjustment
  - p-values adjusted from smallest to largest
  - when find 'large' p-value, that null and all nulls with larger p-values are not rejected
- *Step-up* adjustment
  - p-values adjusted from largest to smallest
  - when find 'small' p-value, that null and all nulls with smaller p-values are rejected

Control of the FWER

- *Bonferroni single-step* adjusted p-values

$$p_j^* = \min(mp_j, 1)$$

- *Sidak single-step (SS)* adjusted p-values

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

- *Sidak free step-down (SD)* adjusted p-values

$$p_{(j)}^* = 1 - (1 - p_j)^{m-j+1}$$

Control of the FWER

- *Holm (1979) step-down* adjusted p-values

$$p^*_{r_j} = \max_{k=1,\ldots,j}\{\min((m-k+1)\,p_{r_k}, 1)\}$$

  - **Intuitive explanation** : once $H_{(1)}$ rejected by Bonferroni, there are only $m-1$ remaining H_0 that might still be true (then do another Bonferroni correction, *etc.*)

- *Hochberg (1988) step-up* adjusted p-values (Simes inequality)

$$p^*_{r_j} = min_{k=j,\ldots,m}\{min((m-k+1)\,p_{r_k}, 1)\}$$

Control of the FWER

- *Westfall and Young (1993) step-down minP* adjusted p-values

$$p^*_{r_j} = \max_{k=1,\ldots,j}\{P(\min_{l\in\{k,\ldots,m\}} P_{r_l} \leq p_{r_k} \mid H_0^C)\}$$

- *Westfall and Young (1993) step-down maxT* adjusted p-values

$$p^*_{r_j} = max_{k=1,\ldots,j}\{P(max_{l\in\{k,\ldots,m\}} \mid T_{r_l} \mid \geq \mid t_{r_k} \mid \mid H_0^C)\}$$

Control of the FDR

- *Benjamini and Hochberg (1995) : step-up* procedure that controls the FDR under *some* dependency structures

$$p^*_{r_j} = \min_{k=j,\ldots,m}\{\min([m/k]\, p_{r_k}, 1)\}$$

- *Benjamini and Yuketieli (2001) : conservative stepup* procedure that controls the FDR under *general dependency* structures

$$p^*_{r_j} = \min_{k=j,\ldots,m}\{\min(m \sum j = 1^m [1/j]/k]\, p_{r_k}, 1)\}$$

- *Yuketieli and Benjamini (1999) : resampling-based adjusted p-values* for controlling the FDR under *certain types* of dependency structures

R : multiple testing

- The BioConductor package `multtest` has implemented a number of adjustments for multiple hypotheses
- The package vignette reviews the functionality
- For gene expression data, `limma` also adjusts for multiplicity
- The BioConductor package `qvalue` computes q-values and various plots `http://www.bioconductor.org`

What about pre-screening ?

- To get around the problem of loss of power when adjusting, some have recommended *'prescreening'* outcome values and only testing those showing sufficient variation
- This is an example of *'data snooping'* : looking at the data before deciding what to test
- Unless the screening statistic is *independent of the test statistic under the null*, the Type I **error rate will not be correct**
- In addition, any p-value for the test may be *difficult to interpret*

Controversies

- *Whether* multiple testing methods (adjustments) should be applied at all
- *Which tests* should be included in the *family* (*e.g.* all tests performed within a single experiment ; define 'experiment')
- Alternatives
  - Bayesian approach
  - Meta-analysis

Situations where inflated error rates are a concern

- It is plausible that *all nulls may be true*
- A *serious claim* will be made whenever any $p < 0.05$ (say) is found
- *Much data manipulation* may be performed to find a 'significant' result
- The analysis is planned to be *exploratory* but wish to claim 'significant' results are real
- Experiment *unlikely to be followed up* before serious actions are taken

Pitfalls in hypothesis testing

- Even if a result is 'statistically significant', *it can still be due to chance*
- Statistical significance is *not* the same as practical importance
- A test of significance does not say how *important* the difference is, or *what caused it*
- A test *does not check the study design*
- If the test is applied to a *nonrandom sample* (or the whole population), the p-value may be *meaningless*
- *Data-snooping* makes p-values hard to interpret