Introduction

- There are a wide variety of population-based association study designs, including candidate gene studies and whole genome association studies.
  - Ascertain samples of unrelated affected cases and unaffected controls. Important to collect data on exposure to potential non-genetic (environmental) risk factors.
- Basic analyses utilise standard epidemiological tools, regardless of study design, rather than specialised methods that have been developed for analysing more traditional pedigree and family studies.
  - Single-locus tests (contingency table analysis; logistic regression modelling).
  - Multi-locus methods (logistic regression modelling; haplotype-based analysis.

Genotype-based single-locus tests

- Assuming the sample to be typed at a SNP marker of interest, we can represent genotype data in a 2 x 3 contingency table.
- The usual χ^2 test for independence of rows and columns in contingency tables can be applied to test the null hypothesis of no disease-marker association

	Cases	Controls	Total
MM	n_{2A}	n_{2U}	$n_{2\cdot}$
Mm	n_{1A}	n_{1U}	$n_{1\cdot}$
mm	n_{0A}	n_{0U}	$n_{0\cdot}$
Total	$n_{\cdot A}$	$n_{\cdot U}$	$n_{\cdot\cdot}$

$$X^2 = \sum_{i=0,1,2} \sum_{j=A,U} \frac{\left(n_{ij} - E\left[n_{ij}\right]\right)^2}{E\left[n_{ij}\right]}$$

where

$$E\left[n_{ij}\right] = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$$

- X^2 has χ^2 distribution with 2 degrees of freedom under null hypothesis.

- Odds ratio for genotype MM relative to mm

$$\psi_{MM|mm} = n_{2A} n_{0U} / n_{0A} n_{2U}$$

- Affected individual $\psi_{MM|mm}$ times more likely to have marker genotype MM than mm.

- The test can be generalised to multi-allelic markers, with *g*-1 degrees of freedom, where *g* is the number of observed genotypes. However, with many genotypes, we encounter the usual problems of sparse data in contingency tables and a lack of power.
- Possible solutions:
  - Use prior information for associated genotypes from previous studies.
  - Pool together all but the most frequent genotypes.
  - Sequential pooling – successively grouping together genotypes that are at high-risk or at low-risk of disease.
- Alternatively, we can assume that alleles have ***independent*** effects on disease penetrance, i.e. a ***multiplicative disease model*** $\psi_{MM|mm} = \psi^2_{Mm|mm}$ using the ***Cochran-Armitage trend test***. Power is very often improved as long as the penetrance of the Mm genotype is intermediate between the two homozygote penetrances.

Single-locus Cochran-Armitage trend tests

- Assuming the sample to be typed at a SNP marker of interest, we can represent genotype data in a 2 x 3 contingency table.
- The Cochran-Armitage trend test of association between disease and the marker SNP is given by

$$X^2 = \frac{\left[\left(p_{2A} + \frac{1}{2} p_{1A}\right) - \left(p_{2U} + \frac{1}{2} p_{1U}\right)\right]^2}{\left(\frac{1}{n_{.A}} + \frac{1}{n_{.U}}\right)\left(\frac{1}{n_{..}^2}\right)\left[n_{..}\left(\frac{1}{4} n_{1.} + n_{2.}\right) - \left(\frac{1}{2} n_{1.} + n_{2.}\right)^2\right]}$$

where

$$p_{ij} = \frac{n_{ij}}{n_{.j}}$$

- X^2 has χ^2 distribution with 1 degree of freedom under null hypothesis.

	Cases	Controls	Total
MM	n_{2A}	n_{2U}	$n_{2\cdot}$
Mm	n_{1A}	n_{1U}	$n_{1\cdot}$
mm	n_{0A}	n_{0U}	$n_{0\cdot}$
Total	$n_{\cdot A}$	$n_{\cdot U}$	$n_{\cdot\cdot}$

- Odds ratio for allele M relative to allele m

$$\psi_{M|m} = \frac{\left(\frac{n_{1A} n_{0U}}{n_{0.} + n_{1.}}\right) + \left(\frac{n_{2A} n_{1U}}{n_{1.} + n_{2.}}\right) + \left(\frac{4 n_{2A} n_{0U}}{n_{0.} + n_{2.}}\right)}{\left(\frac{n_{0A} n_{1U}}{n_{0.} + n_{1.}}\right) + \left(\frac{n_{1A} n_{2U}}{n_{1.} + n_{2.}}\right) + \left(\frac{4\left[n_{2A} n_{2U} n_{0A} n_{0U}\right]^{1/2}}{n_{0.} + n_{2.}}\right)}$$

- Affected individual $\psi_{M|m}^2$ times more likely to have marker genotype MM than mm, and $\psi_{M|m}$ times more likely to have genotype Mm than mm.

Allele-based single-locus tests

- Each individual now contributes ***two*** counts to the contingency table, one for each allele in their marker genotype.
- Assuming the sample to be typed at a marker SNP of interest, we can represent genotype data in a 2 x 2 contingency table.
- To test the null hypothesis of no disease-marker association

$$X^2 = \sum_{i=0,1} \sum_{j=A,U} \frac{\left(n_{ij} - E\left[n_{ij}\right]\right)^2}{E\left[n_{ij}\right]}$$

where $E\left[n_{ij}\right] = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$

- X^2 has χ^2 distribution with 1 degree of freedom under null hypothesis.

	Cases	Controls	Total
M	n_{1A}	n_{1U}	$n_{1\cdot}$
m	n_{0A}	n_{0U}	$n_{0\cdot}$
Total	$n_{\cdot A}$	$n_{\cdot U}$	$n_{\cdot\cdot}$

- Odds ratio for allele M relative to m

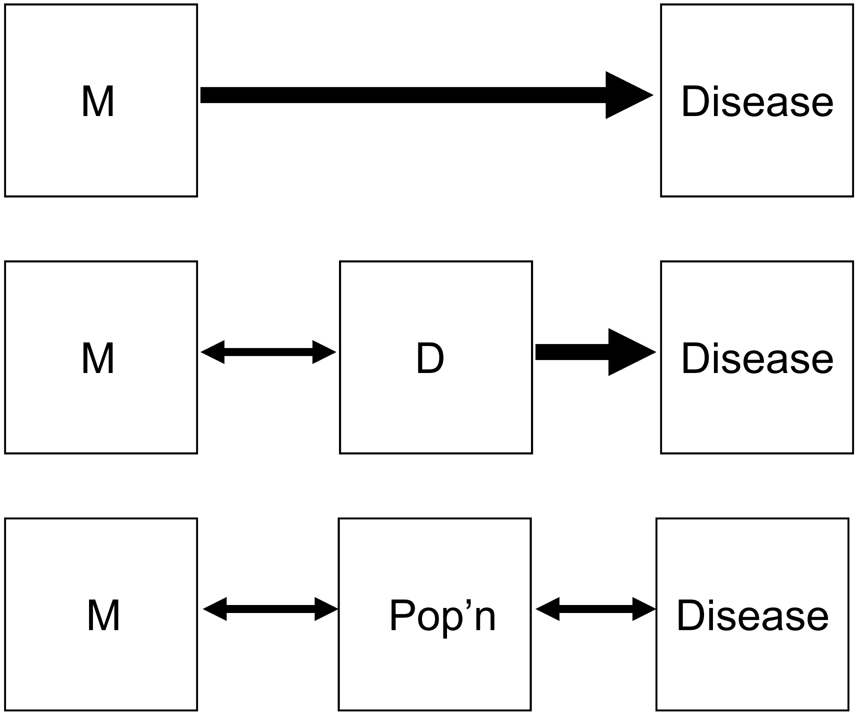
$$\psi_{M|m} = n_{1A} n_{0U} / n_{0A} n_{1U}$$

- Allele M is $\psi_{M|m}$ times more likely to be carried by an affected individual than allele m.
- Assumes multiplicative disease risks and Hardy-Weinberg equilibrium at SNP in cases and controls.

Interpretation

A significant result in a test of disease-marker association may imply:

- Marker locus is ***causative***, directly influencing disease risk: needs to be established via functional studies.
- Alleles at marker locus are ***correlated*** with alleles at the disease locus, but do not directly influence disease risk: ***linkage disequilibrium***.
- ***Population substructure*** not accounted for in the analysis, with different disease and marker allele frequencies in each subpopulation.
- ***False positive*** signal of association.



Example: type 2 diabetes T2D

- PPARGamma Pro12Ala polymorphism is a strong candidate for T2D.
- Ardlie *et al*. (2002) typed the variant in 500 cases and 500 controls in two populations: Poles and US with North European ancestry.
- Significantly reduced frequency of C allele in T2D cases in Polish sample (p<0.0001).
- No evidence of association with T2D in US sample.
- Population substructure among Poles or greater genetic heterogeneity among US sample?
- More recent studies suggest no evidence of substructure in Poles, and replicate the association in a more homogeneous US sample.

Polish sample:

	C	G	Total
Cases	124 (13%)	838	962
Controls	183 (20%)	743	926

US sample:

	C	G	Total
Cases	96 (10%)	894	990
Controls	102 (10%)	890	992

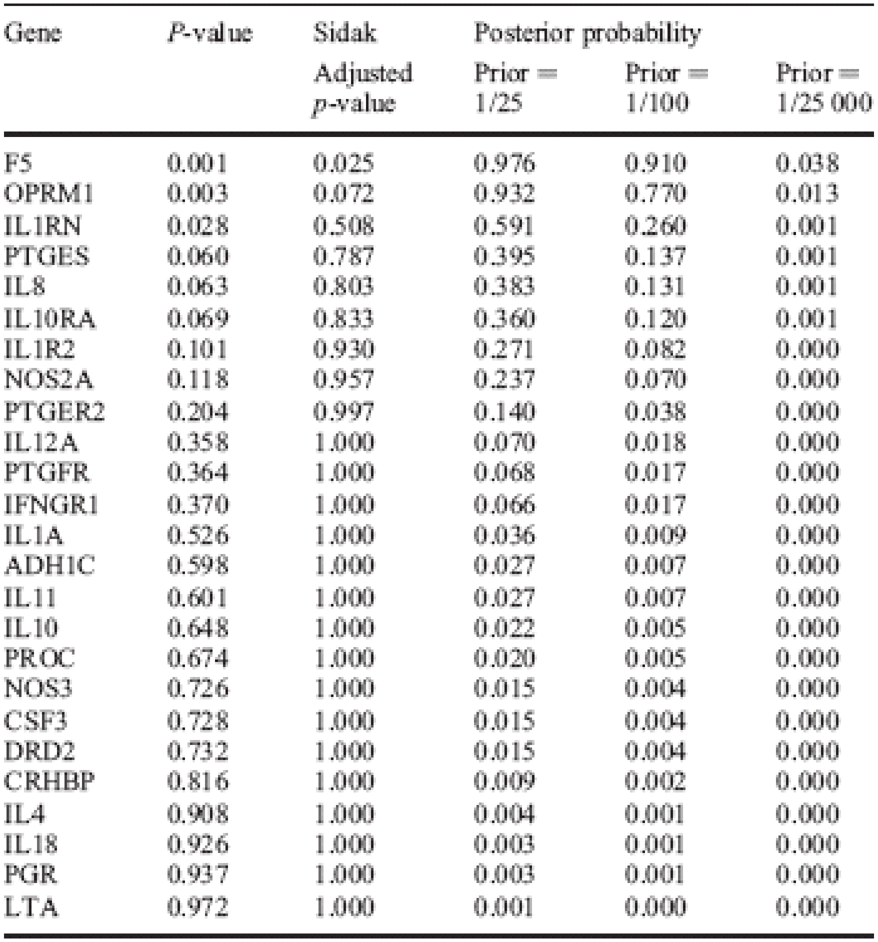
Multiple testing

- A type I error occurs when we reject the null hypothesis of no association, when in fact the null hypothesis is true.
- Specify type I error rate – or significance level – at the design stage of the analysis.
  - Lower type I error rate reduces the probability of detecting a false positive association, but with the penalty of reducing the power to detect association when it truly exists.
- If we choose a 5% significance level, if we test 20 independent SNPs for association with disease, we expect one of them to show significant evidence of association, even if none of them are truly associated with disease.
- It is important to correct for multiple testing to maintain the type I error rate for the experiment overall (i.e. all the SNPs tested in the association study).
- ***Replication*** is necessary to confirm association.

- ***Bonferroni correction***. Treat each test as independent and adjust point-wise significance level to achieve overall experiment-wise type I error rate of 100α%.
  - When testing *N* SNPs, use significance level of 100α/*N*% for rejecting null hypothesis at each SNP.
  - Sidak *p*-value: $1-(1-p)^N$.
  - Assumes all tests are independent, and thus will be conservative if there is linkage disequilibrium between SNPs.
- ***False discovery rate***. Expected number of false positive signals among significant associations.
  - FDR using pointwise *p*-value of 100α% given by $N\alpha/k$, where *k* is the number of SNPs with $p \leq \alpha$.
- ***Permutation procedures***. Generate the distribution of experiment-wise test statistics under the null hypothesis of no association by creating permuted data sets by randomly exchanging case and control labels.
  - Compare observed test statistics with maximum test statistic from each permuted data set.
- ***Bayesian approaches***. Assign prior probability that each SNP is associated with disease.
  - Choice of prior probability is subjective.
  - Can incorporate information about functional relevance of the SNP for the disease under investigation.

Multiple testing: example

- Hao et al. (2004) tested for association between preterm delivery and SNPs in 25 candidate genes.
- Using a Bonferonni correction for 25 tests, only the strongest association (F5) remains significant at an experiment-wise 5% level.
- Using a pointwise p-value of 0.3%, genes F5 and OPRM1 are significantly associated with preterm delivery: we would expect 3.75% of these signals to be false positives.
- Posterior probability of association depends on prior probability assigned to each gene.
  - Prior 1/25: testing only 25 candidate genes.
  - Prior 1/100: testing first 25 of 100 candidate genes.
  - Prior 1/25000: expect only one associated gene in genome, and 25 tested are not strong candidates.



Gene	*P*-value	Sidak Adjusted *p*-value	Posterior probability Prior = 1/25	Prior = 1/100	Prior = 1/25 000
F5	0.001	0.025	0.976	0.910	0.038
OPRM1	0.003	0.072	0.932	0.770	0.013
IL1RN	0.028	0.508	0.591	0.260	0.001
PTGES	0.060	0.787	0.395	0.137	0.001
IL8	0.063	0.803	0.383	0.131	0.001
IL10RA	0.069	0.833	0.360	0.120	0.001
IL1R2	0.101	0.930	0.271	0.082	0.000
NOS2A	0.118	0.957	0.237	0.070	0.000
PTGER2	0.204	0.997	0.140	0.038	0.000
IL12A	0.358	1.000	0.070	0.018	0.000
PTGFR	0.364	1.000	0.068	0.017	0.000
IFNGR1	0.370	1.000	0.066	0.017	0.000
IL1A	0.526	1.000	0.036	0.009	0.000
ADH1C	0.598	1.000	0.027	0.007	0.000
IL11	0.601	1.000	0.027	0.007	0.000
IL10	0.648	1.000	0.022	0.005	0.000
PROC	0.674	1.000	0.020	0.005	0.000
NOS3	0.726	1.000	0.015	0.004	0.000
CSF3	0.728	1.000	0.015	0.004	0.000
DRD2	0.732	1.000	0.015	0.004	0.000
CRHBP	0.816	1.000	0.009	0.002	0.000
IL4	0.908	1.000	0.004	0.001	0.000
IL18	0.926	1.000	0.003	0.001	0.000
PGR	0.937	1.000	0.003	0.001	0.000
LTA	0.972	1.000	0.001	0.000	0.000

Taken from Farrall and Morris (2005).

Power calculations

- The GENETIC POWER CALCULATOR can be used to calculate power of simple case-control studies. Website found at http://statgen.iop.kcl.ac.uk/gpc.
- Specification of model parameters:
  - Disease model: disease prevalence, disease allele frequency, disease genotype relative risks.
  - Marker allele frequencies, linkage disequilibrium between disease locus and marker locus (D').
- Specify point-wise significance level and power, and case/control ratio to obtain required sample size.
- For low disease genotype relative risks, we can maximise power by matching marker allele and disease allele frequencies, with strong linkage disequilibrium between loci.
- Low power for rare disease variants, regardless of marker allele frequency and linkage disequilibrium.

Design issues

We may have nuclear family or pedigree data available from previous linkage studies. Is it cost effective to make use of these individuals in a subsequent association study?

- We can form internal control genotypes from the pair of alleles not transmitted from parents to affected offspring: matched analysis protects against population stratification. However this approach requires more genotyping than unmatched analysis with unrelated population controls.
- We can achieve greater power by using multiple affected sibs from the same family than the equivalent number of unrelated population cases.
    - There is greater probability that related cases are affected due to shared underlying genetic factors.
    - However, it is important to remember to allow for correlated observations in the analysis.

Logistic regression model

- We can model the case/control status of an individual within a logistic regression framework, parameterised in terms of log-odds of disease, **β**, for marker genotypes.
- Straightforward to incorporate covariates, **x**, which may include non-genetic risk factors, polygenic effects, or ancestrally informative markers to allow for population stratification.
- Let π_i denote the probability that individual *i* is a case, given their genotype G_i. The ***logit link*** function

$$\pi_i = \Pr\left(i \text{ is case} \,\middle|\, G_i, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \frac{\mathbf{exp}[\eta_i]}{1 + \mathbf{exp}[\eta_i]}$$

where $\eta_i = \beta_0 + \sum_j \gamma_j x_{ij} + \beta_A z_{(A)i} + \beta_D z_{(D)i}$, and **γ** denote covariate regression coefficients.

- It is common to parameterise genotype effects in terms of ***additive*** effects (multiplicative contribution to disease risk), β_A, and ***non-additive*** effects (any non-multiplicative contribution to disease risk), β_D.
- SNP genotype of individual *i* coded by two indicator variables, $Z_{(A)i}$ and $Z_{(D)i}$, representing additive and non-additive effects.
- Log-likelihood for specified model given by

$$l(\boldsymbol{y}|\boldsymbol{G},\mathbf{x},\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_i \left[y_i \, \boldsymbol{ln}\, \pi_i + (1-y_i)\boldsymbol{ln}(1-\pi_i)\right]$$

where y_i denotes the disease phenotype of individual *i* ($y_i = 1$ if case and $y_i = 0$ if control).

- Compare models by ***analysis of deviance***, having approximate χ^2 distribution with degrees of freedom given by the difference in the number of parameters.
  - Additive effect: $2[l(\boldsymbol{y}|\boldsymbol{G},\boldsymbol{x},\beta_A,\beta_0,\boldsymbol{\gamma})-l(\boldsymbol{y}|\boldsymbol{G},\boldsymbol{x},\beta_0,\boldsymbol{\gamma})]$
  - Genotypic effect: $2[l(\boldsymbol{y}|\boldsymbol{G},\boldsymbol{x},\beta_D,\beta_A,\beta_0,\boldsymbol{\gamma})-l(\boldsymbol{y}|\boldsymbol{G},\boldsymbol{x},\beta_0,\boldsymbol{\gamma})]$
  - Non-additive effect: $2[l(\boldsymbol{y}|\boldsymbol{G},\boldsymbol{x},\beta_D,\beta_A,\beta_0,\boldsymbol{\gamma})-l(\boldsymbol{y}|\boldsymbol{G},\boldsymbol{x},\beta_A,\beta_0,\boldsymbol{\gamma})]$

Genotype	$Z_{(A)i}$	$Z_{(D)i}$
mm	-1	0
Mm	0	1
MM	1	0

- Affected individual is $\exp[2\beta_A]$ times more likely to have genotype MM than genotype mm.
- Affected individual is $\exp[\beta_A+\beta_D]$ times more likely to have genotype Mm than genotype mm.

Alternative association models

- May have evidence from previous linkage and/or association studies that disease risk can best be described by means of a recessive, dominant, or heterozygote advantage model.
- Set up indicator variables to represent these models, and perform single degree of freedom association test by comparison of deviance with the null model.

Genotype	M Recessive	M Dominant	Heterozygote advantage
mm	0	0	0
Mm	0	1	1
MM	1	1	0

Quantitative traits

- The methodology described here generalises to quantitative (continuous) traits. It is straightforward to compare the mean response for each marker genotype by analysis of variance, assuming a normally distributed trait, within the standard linear regression framework.
- A powerful strategy is to ascertain individuals from the extremes of the quantitative trait distribution: cases and hyper controls.
  - We can analyse trait values by linear regression, although this leads to biased estimates of mean trait values for marker genotypes.
  - We can ignore the trait values, and analyse as a standard case-control sample.
  - Are hyper controls representative, or are there polygenic effects involved?
  - This strategy may not be cost effective if phenotyping is expensive relative to genotyping.

Software

- Contingency table analysis and generalised linear modelling can be performed using standard statistical software.
    - Define indicator variables for specific genetic models from the observed SNP genotype data.
- Some statistical software packages include specific libraries of routines to perform genetic analyses (R, STATA)
- Specialised genetic analysis software:
    - ***PLINK***. Whole genome association analysis toolset designed to perform a range of basic, large-scale analyses. Allows for data management and basic QC analyses. Performs simple case-control tests of association.
    - ***SNPTEST***. Designed for analysis of whole genome association studies. Allows for flexible single-locus analysis of genotype data allowing for covariates.

Multi-locus association models

- Typically, many SNPs will be genotyped in a candidate gene, or high-density SNPs will be genotyped in a whole genome association study.
- Single-locus tests may lack power to detect association with the disease:
  - each individual SNP provides relatively little information about linkage disequilibrium with the disease variant;
  - Bonferroni correction for multiple testing is likely to be conservative.
- Greater power is expected by ***joint analysis*** of all markers in the same gene or region simultaneously by considering ***multi-locus models*** of association.
- We expect strong correlation between markers in the same gene due to linkage disequilibrium. Consequently, the information at one marker may become redundant given the genotypes at additional loci, and the effects of linked markers may be strongly correlated with each other.

- The logistic regression framework provides a natural hierarchy of multi-locus association models, allowing for:
    - ***main effects***, reflecting differences in multi-locus genotype frequencies in cases and controls;
    - ***interactions***, reflecting differences in multi-locus genotype frequencies in cases and controls, over and above main effects.
- Additive and dominance main effects of each marker coded as for a single-locus analysis.
- Potentially four contributions to interaction between each pair of SNPs: additive and non-additive at each SNP.
    - Indicator variables for each interaction term are given by the product of indicator variables for the corresponding main effects. For example, the additive-additive contribution to the interaction for individual *i* is coded by $Z_{(A1)i}{}^{*}Z_{(A2)i}$.

- Consider two SNP markers: there are three distinct genotypes at each marker, and consequently nine two-locus genotypes.
- There is a natural hierarchy of allelic (A) and genotype (G) models, compared via analysis of deviance.

Multi-locus model		Parameters
Null	0	β_0
Locus 1 only	A1	β_0, β_{A1}
	G1	β_0, β_{A1}, β_{D1}
Locus 2 only	A2	β_0, β_{A2}
	G2	β_0, β_{A2}, β_{D2}
Main effects	A1+A2	β_0, β_{A1}, β_{A2}
	A1+G2	β_0, β_{A1}, β_{A2}, β_{D2}
	G1+A2	β_0, β_{A1}, β_{D1}, β_{A2}
	G1+G2	β_0, β_{A1}, β_{D1}, β_{A2}, β_{D2}
Interaction	A1*A2	β_0, β_{A1}, β_{A2}, β_{A1*A2}
	A1*G2	β_0, β_{A1}, β_{A2}, β_{D2}, β_{A1*A2}, β_{A1*D2}
	G1*A2	β_0, β_{A1}, β_{D1}, β_{A2}, β_{A1*A2}, β_{D1*A2}
	G1*G2	β_0, β_{A1}, β_{D1}, β_{A2}, β_{D2}, β_{A1*A2}, β_{A1*D2}, β_{D1*A2}, β_{D1*D2}

- For multiple markers, we can employ standard model selection techniques, for example forward selection, backward elimination, and stepwise selection. However, this approach may increase the false positive error rates for testing for disease-marker association in the candidate gene.
- Alternatively, we can estimate additive and dominance main effects and interactions in a ***Bayesian model averaging*** framework, which takes account of uncertainty in the true underlying model.
- High-order interaction terms are likely to be difficult to estimate and complex to interpret in terms of genetic effects.
- To fit main-effects only models, we require marker locus genotype frequencies over the whole sample, rather than the multi-locus genotype of each individual. As a result, we can utilise ***DNA pooling***, reducing costs relative to standard SNP genotyping methods.

Replication and multi-stage designs

- To confirm positive association signals from an initial study, it is essential to replicate the result in independent samples from the same and/or different populations.
- Important to define what is meant by replication:
  - Association of same SNP or haplotype, with same high-risk variant(s) identified.
  - Association of same SNP or haplotype, with different high-risk variants identified???
  - Association of different SNP in the same gene???
- Replication of positive association signals has not proved to be easy: will depend on power of both initial and replication studies.
- For genome-wide association studies, multi-stage designs have been proposed as an efficient approach to allow for the possibility of replication.
  - Initial screen of the whole genome to identify positive signals of association.
  - Follow up the top *K* signals in an independent sample, reducing the number of tests performed: goal here is to identify which of the positive signals from the initial screen are false positives, and which might be carried forward for further testing.

Summary

- Standard statistical procedures available for the analysis of genotype data from genetic association studies.
- Logistic regression provides a flexible framework for modelling complex disease risk.
  - Can incorporate multi-locus models of association and covariates to allow for non-genetic risk factors, polygenic effects, and indicators of population stratification.
- Multi-locus association models take account of the correlation between proximal SNPs due to background patterns of linkage disequilibrium.
- Adjustment must be made for multiple testing, either by means of simple correction factors, or via permutation procedures.