Genome Wide Association Studies

- What is a Genome Wide Association Study?
- What is a Genome Wide Linkage Study?
- Linkage vs Association (Risch and Merikangas 1996)
- Study Design
- Different methods for detecting association

What is a Genome Wide Association Study?

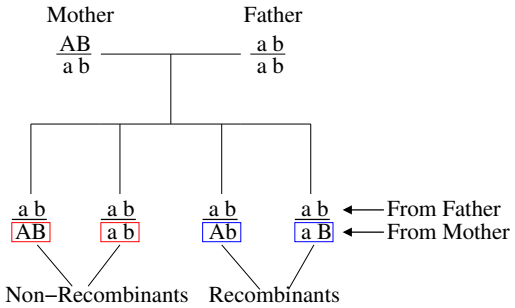
Goal Uncover the genetic basis of a given disease.
Basic Idea A rather vague idea of a study design that involves genotyping cases and controls at a large number ($10^4 \rightarrow 10^6$) of SNP markers spread (in some unspecified way) throughout the genome. Look for associations between the genotypes at each locus and disease status.

0 1 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

Why might this be a good idea?

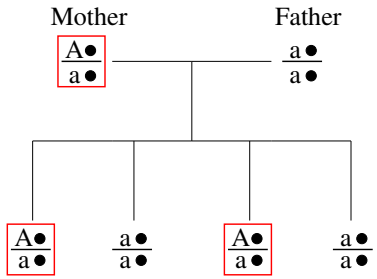
Linkage Mapping

Basic Idea Use the pattern of allele segregation in pedigrees (families) to estimate recombination fraction (θ) between a marker locus and an unobserved trait locus. (Sham (1998))



Out of 4 informative meioses 2 are recombinants $\Rightarrow \hat{\theta} = 1/2$
i.e. we can use pedigrees to estimate the "distance" between 2 markers.

Suppose marker B is the causative disease locus but the genotype information at this marker is missing. Instead we have disease status for the individuals in the pedigree. A is a marker we think is close to B.



Given a model of penetrance (i.e. how genotype dictates disease status) then for a given value of θ we can calculate the likelihood of the pedigree by summing over all the missing data configurations consistent with the data.

$$\begin{array}{rcccccc} L(\theta) & = & \sum_G & P(X|G) & \times & P(G_d|G_f;\theta) & \times & P(G_f) \\ \text{(Likelihood)} & = & \sum_G & \text{(Penetrance)} & \times & \text{(Transmission)} & \times & \text{(Founders)} \end{array}$$

Thus we can obtain the MLE, $\hat{\theta} \Rightarrow$ Likelihood ratio test $\frac{L(\hat{\theta})}{L(1/2)}$

Normally, the "lod score" is reported where $\text{lod}(\theta) = \log_{10}\left(\frac{L(\hat{\theta})}{L(1/2)}\right)$

When disease status has a high correlation with the genotype at marker A then this suggests the recombination fraction between markers A and B is small i.e. A is close to B.

Because recombination events are so rare we can use a sparse set of (Micro Satellite) markers spread throughout the genome to "map" the disease locus (usually between 200-500 markers).

Once areas of the Genome have been implicated Fine Mapping studies in candidate genes can be used to localize the causal locus.

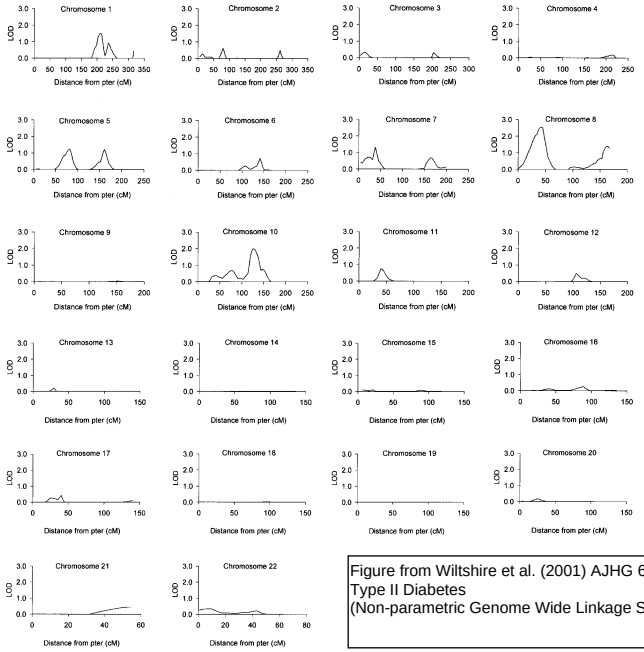


Figure from Wiltshire et al. (2001) AJHG 69:553-69
Type II Diabetes
(Non-parametric Genome Wide Linkage Study)

It's not quite as simple as I've made out (Sham (1998))

- Need to specify a penetrance function
- Founder haplotype probabilities
- Allow variable recombination rates in men and women?

It is also non-trivial to sum over all the missing data and considerable effort has been put into efficient calculation of the likelihoods (Abecasis et al. (2002)) [In much the same way that people have worked very hard to calculate likelihoods in population genetics models]

Extensions

- Multiple loci
- Other covariates e.g. environmental factors
- Continuous traits

Successes and Failures of Linkage Mapping

Linkage Mapping has been successful in identifying the genetic basis of many human diseases in which the disease penetrance resembles a simple Mendelian model

- Huntington's disease, Cystic Fibrosis, some forms of breast cancer

Risch and Merikangas (1996)

But "the literature is now replete with linkage screens for an array of common 'complex' disorders such as

- schizophrenia, manic depression, autism, asthma, type I and type II diabetes, Multiple Sclerosis, Lupus

Although many of these studies have reported significant linkage findings, none has lead to convincing replication"

Risch(2001)

What is a Complex Disease?

Not simply Mendelian!

Possible departures from the simple Mendelian model include

Allelic Heterogeneity Different alleles at the same locus increase disease susceptibility i.e. not just one allele. (linkage mapping is robust to this effect)

Locus Heterogeneity Mutations at different loci increase disease susceptibility (linkage mapping is not robust to this effect)

Gene-Gene Interactions multiple loci "interact" to increase disease susceptibility

Environmental factors e.g. diet, smoker

Gene-Environment Interactions

Linkage Mapping vs Association Studies

In a widely quoted paper

Risch and Merikangas (1996) The Future of Genetic Studies of Complex Human Diseases. Science **273**:1516-17

the authors pointed out that linkage studies had less power than association studies to detect weak genetic effects exhibited by the loci involved in complex diseases.

Not quite that general.

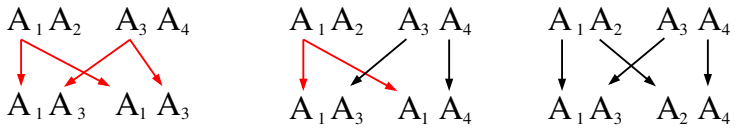
More specifically, the authors compared two specific methods of linkage and association that were popular at that time

- Non-parametric linkage mapping using an Affected Sib Pairs (ASP) design
- Family based association using the Transmission Disequilibrium/Distortion Test (TDT)

Non-Parametric Linkage

Design Two Affected Sibs and their parents (Micro Satellite Locus)

Basic Idea Count the number of alleles two affected sibs share Identical By Descent (IBD)



If the marker is linked to the disease locus the affected sibs will tend to share the disease allele more often than they would at a marker unlinked to the disease locus.

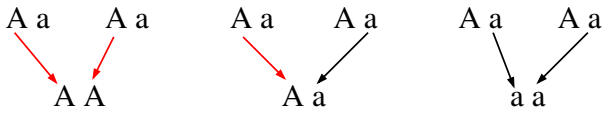
$\Rightarrow$ There will be a departure from the null IBD distribution of $\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$.

Note No distinction is made between *which* allele is shared IBD.

TDT (Transmission Disequilibrium Test)

Design Affected Child and their parents (SNP Locus)

Basic Idea Compare the distribution of the transmitted allele to the distribution of the non-transmitted allele from heterozygous parents.



If the marker is linked to the disease locus one of the alleles will tend to be transmitted more often than if the marker was unlinked to the disease locus.

$\Rightarrow$ There will be a departure from the null distribution of $\left(\frac{1}{2}, \frac{1}{2}\right)$.

Note This test *does* effectively focus on a specific allele.

Disease Model

Risch and Merikangas assumed that the disease locus was diallelic with the Genotypic Relative Risk (GRR) increasing in a multiplicative fashion

$$\frac{P(\text{Disease}|Aa)}{P(\text{Disease}|aa)} = \gamma \qquad \frac{P(\text{Disease}|AA)}{P(\text{Disease}|aa)} = \gamma^2$$

For both tests they assumed the marker locus is completely linked to the disease locus and calculated the IBD distribution and transmission distribution for a given values of γ.

Q. How many families do you need for 80% power?

ASP They assumed they had 500 MS markers spread throughout the genome $\Rightarrow \alpha = 10^{-4}$

TDT They assumed they had 1,000,000 SNP markers spread throughout the genome $\Rightarrow \alpha = 5 \times 10^{-8}$

Power of ASP vs TDT

γ	Allele Frequency	ASP (N)	TDT (N)
2	0.01	296,710	5,823
2	0.1	5,382	695
1.5	0.01	4,620,807	19,320
1.5	0.1	67,816	2,218

Even with the much more stringent Type I error TDT is seen to have much more power to detect an effect.

Note 1 The power of both tests depends on allele frequency. If the disease allele is rare it reduces the number of heterozygous parents with the mutant allele.

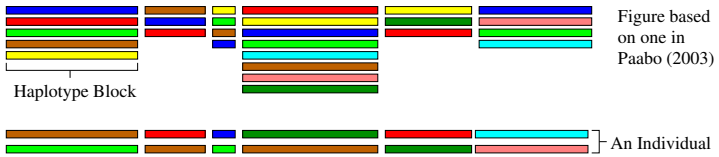
Note 2 Deeper families will have more power than the ASP design but there will also be a dependence on allele frequency.

Common variants or rare alleles?

The **common disease/common variant hypothesis** (CD/CV) holds that alleles at relatively high frequencies ($> 1\%$) represent a significant proportion of susceptibility alleles for common disease.

Their high frequency implies that association studies in large population cohorts will be fruitful for identifying risk alleles.

Based on recent empirical evidence some people have suggested that we may be able to characterize the variation in the human genome using a block like structure of common haplotypes.



If this is true then association studies may proceed by typing just those SNPs (Haplotype tagging SNPs or htSNP's) that code the common haplotypes.

The 'HapMap' project will investigate this issue (Nature Genetics (2001) **29**:353-4)

In opposition is the **common disease/rare allele hypothesis** (CD/RA) which holds that there is no reason to expect that most common genetic diseases result from common alleles.

Simulations from models based on empirical parameter estimates (Pritchard (2001)) suggest that this may be the case and that should expect extensive allelic heterogeneity.

Many people also expect extensive locus heterogeneity.

If this is the case then Haplotype Maps will be of limited use and that family studies in unusual populations (e.g. Iceland) may be the only way to go.

May be a mixture of both?

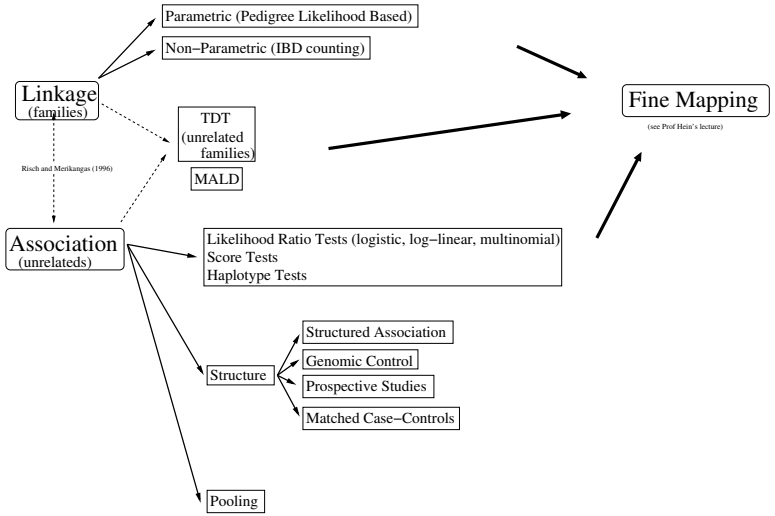
Genome Wide Association Studies in Practice

Risch and Merikangas (1996) says that to detect a disease allele with a frequency of 0.1 and GRR = 1.5 we need to genotype 2, 218 families at 1,000,000 SNP loci.

This isn't a solution. There are still lots of questions.

- Is this design practical? Costs? Pooling?
- Is TDT the best association design? Families vs Unrelated? MALD?
- How do both these approaches cope with population stratification?
- What if we don't type the causative marker but one in LD to the marker?
- How do allelic heterogeneity, locus heterogeneity, gene-gene interactions and gene-environment interactions impact association studies?

The data produced by the Haplotype Map project will have a large influence on association study design.



Multinomial Likelihood

Data	Case	Control	
aa	$r_0 = 100$	$s_0 = 130$	n_0
aA	$r_1 = 400$	$s_1 = 390$	n_1
AA	$r_2 = 500$	$s_2 = 480$	n_2
	R	S	N

We want to test whether Case/Control status is associated with disease i.e. is there a difference in the distribution of genotypes between Cases and Controls.

Thus, we can write our null and alternative hypotheses in terms of the conditional probabilities of genotype given case/control status, parameterized by $\boldsymbol{p} = \{p_0, p_1, p_2\}$ and $\boldsymbol{q} = \{q_0, q_1, q_2\}$

	Case	Control
aa	p_0	q_0
aA	p_1	q_1
AA	p_2	q_2

In terms of these probabilities, we can write $H_0 : \boldsymbol{p} = \boldsymbol{q}$ vs $H_1 : \boldsymbol{p} \neq \boldsymbol{q}$

The likelihood is multinomial

$$L(\boldsymbol{p}, \boldsymbol{q}) = p_0^{r_0} p_1^{r_1} p_2^{r_2} q_0^{s_0} q_1^{s_1} q_2^{s_2}$$

Under H_0 the MLE's of $\boldsymbol{p}$ and $\boldsymbol{q}$ are both $\boldsymbol{v} = \{n_0/N, n_1/N, n_2/N\}$

Under H_1 the MLE's are $\hat{\boldsymbol{q}} = \{s_0/S, s_1/S, s_2/S\}$ and $\hat{\boldsymbol{p}} = \{r_0/R, r_1/R, r_2/R\}$

We can test H_0 vs H_1 using the log-likelihood ratio (LLR) statistic

$$-2 \log \left(\frac{L(\boldsymbol{v}, \boldsymbol{v})}{L(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})}\right)$$

The LLR statistic can be written as

$$\sum_{i=0}^{2} r_i \log \left(\frac{r_i}{\hat{r}_i}\right) + s_i \log \left(\frac{s_i}{\hat{s}_i}\right)$$

where $\hat{r}_i$ and $\hat{s}_i$ are the fitted frequencies under H_0.

For example, $\hat{r}_0 = N \times \frac{R}{N} \times \frac{n_0}{N}$

Under H_0 this statistic has a χ^2_2 distribution (asymptotically).

For the example on the previous slide LLR = 4.45899

Logistic Regression

Alternatively, we can fit a logistic regression model to the data.

Each subject in our sample consists of a (y_i, x_i) pair where y_i is case/control status (1/0) and $x_i \in \{0, 1, 2\}$ is the genotype at the typed locus.

The relationship between y and x is modelled using the likelihood

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)} \quad \text{where} \quad \eta_i = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

This statistical model is equivalent to the genetics model in which the odds of disease given genotype increase in a multiplicative fashion.

$$\text{Odds of Genotype aa} = \frac{P(D|aa)}{1-P(D|aa)}$$

	Genotype	Odds
(aa)	0	α
(Aa)	1	$\alpha(1+\theta)$
(AA)	2	$\alpha(1+\theta)^2$

$\pi_A = P(A)$

$p =$ Disease prevalence in the population

$$= \frac{\alpha}{1+\alpha}(1-\pi_A)^2 + 2\frac{\alpha(1+\theta)}{1+\alpha(1+\theta)}\pi_A(1-\pi_A) + \frac{\alpha(1+\theta)^2}{1+\alpha(1+\theta)^2}\pi_A^2$$

The relationship between the models is

$$\beta_0 = \log\alpha \qquad \beta_1 = \log(1+\theta)$$

There is no explicit formulae for the MLE of β. The Likelihood is maximized numerically using an algorithm called Iteratively Re-Weighted Least Squares (IRWLS). The algorithm is derived by using a Newton-Raphson algorithm to maximize the Likelihood.

Using this framework we can test the hypothesis that there is an association which takes the form of a multiplicative increase in the odds of disease.

Writing $\beta^{(0)}$ as the MLE of β under H_0 and $\beta^{(1)}$ as the MLE of β under H_1 the LLR statistic is

$$LLR = -2\log\left(\frac{L(\beta^{(0)})}{L(\beta^{(1)})}\right)$$

For the example used before LLR = 2.693173

The logistic regression framework allows us more flexibility to include other covariates into the linear predictor η_i e.g. gentotypes at other loci, environmental covariates, interaction terms etc.

The flexibility implies that the LLR test using the multinomial likelihood is a special case of the logistic regression model when we set the linear predictor to have the form

$$\eta_i = \beta_0 + \beta_1 x_i + \beta_2 y_i$$

where $y_i = I[x_i > 1]$ and $H_0 : \beta_1 = \beta_2 = 0$

Score Tests

Score tests can be thought of as approximate likelihood ratio tests.

Suppose the likelihood can be written as $L(\alpha, \beta)$ and $H_0 : \beta = 0$

These tests take the form $U^T V^{-1} U$ where U is the vector of first derivatives of the log-likelihood w.r.t to β evaluated at the MLE of α under H_0 and $\beta = 0$. V is covariance matrix of U under H_0.

Under H_0 this statistic has a χ^2_2 distribution (asymptotically).

Applying this method to the multinomial likelihood results in the statistic

$$\sum_{i=0}^{2}\left[\frac{(r_i - \hat{r}_i)^2}{\hat{r}_i} + \frac{(s_i - \hat{s}_i)^2}{\hat{s}_i}\right]$$

This is the "standard" Pearson's Chi-squared test statistic for a 2×3 contingency table.
For the example used before the Score test = 4.4478

Applying this method to the logistic regression model with $\eta_i = \beta_0 + \beta_1 x_i$ results in

$$\frac{N\{N(r_1 + 2r_2) - R(n_1 + 2n_2)\}^2}{R(N-R)\{N(n_1 + 4n_2) - (n_1 + 2n_2)^2\}}$$

which is equivalent to Armitage's Trend test statistic i.e. a popular test statistic used in testing for association at a given SNP marker.

For the example used before the Score test = 2.69179

Multiple Alleles

The tests/models can be extended to the case when we have more than two alleles at a given locus.

Genotype Models

Suppose the locus has K alleles then there will be $K(K+1)/2$ possible genotypes at the locus.

Thus we might think that each genotype has a different effect on the odds of disease.

In a logistic regression framework we can model this using a linear predictor of the form

$$\eta_{(i,g)} = \log\left(\frac{p_{(i,g)}}{1-p_{(i,g)}}\right) = \beta_g$$

where g codes for genotype.

This model effectively models interactions between the alleles.

We can test for an effect at the locus using a LLR test with $K(K+1)/2 - 1$ degrees of freedom.

When K is large the test will have low power.

Allelic Models

We can reduce the number of parameters in the model by assuming a specific form for the interaction between alleles at each locus i.e. they interact to increase the odds of disease in a multiplicative fashion.

$$\eta_{(i,j,k)} = \log\left(\frac{p_{(i,j,k)}}{1-p_{(i,j,k)}}\right) = \beta_0 + \beta_j + \beta_k$$

where j and k codes for the two alleles that make up a specific genotype.

We can test for an effect at the locus using a LLR test with $K-1$ degrees of freedom.

The log link form of this model can be generalized (Clayton (2001)).

Alternatively we could carry out K 1 df tests, each one focussing on a specific allele and combining all the other alleles.

If we do this we need to correct for the multiple testing involved. This can be achieved through simulation.

Sham and Curtis (1995) search over all possible 1 df comparisons obtainable by collapsing alleles into two groups and construct p-values through simulation.

Haplotype Tests

If we don't observe that causative locus directly we need to pick up the signal of association using a marker or markers in LD with the causative marker. It makes sense to try to combine the information in several SNP markers by considering the haplotype effects at these markers.

In practice, we will probably not observe the haplotypes of each individual at a given set of markers. There has been alot of recent work in the area of haplotype phase reconstruction so this is not a serious problem (Stephens et al. (2001))

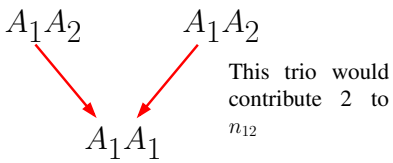
Approach 1 If we assume we know the haplotypes at a given set of markers one approach we can take is to treat the set of markers as one multi-allelic marker as discussed before.

Approach 2 We might think that haplotypes that are "similar" will have similar effects on disease susceptibility. This idea has lead to approaches that use ideas from spatial modelling of disease risk to impose this type of "prior" structure on the estimation of genotype and haplotype risks (Clayton and Jones (1999), Seaman et al. (2002))

TDT

The TDT test compares the distribution of transmitted and non-transmitted alleles by parents of affected offspring (Speilman et al. (1993))

		Non-transmitted allele		
		A_1	A_2	Total
Transmitted Allele	A_1	n_{11}	n_{12}	$n_{1\cdot}$
	A_2	n_{21}	n_{22}	$n_{2\cdot}$
	Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$2n$



This trio would contribute 2 to n_{12}

If the marker is unlinked to the causative locus then we expect $n_{12} = n_{21}$.

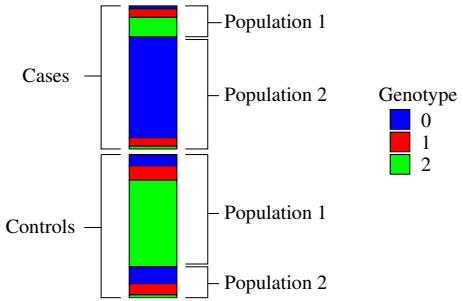
Thus we can test for a distortion in the transmission distribution using the test statistic

$$\frac{(n_{12}-n_{21})^2}{(n_{12}+n_{21})}$$

which is asymptotically χ^2_1.

Various extensions of this method exist i.e. multi allelic markers, multiple siblings, missing parental data

Population Structure



Spurious association is caused by the co-occurnce of 2 factors

- A difference in proportion of individuals from two (or more) subpopulations in cases and controls
- Subpopulations have differing allele frequencies at the locus

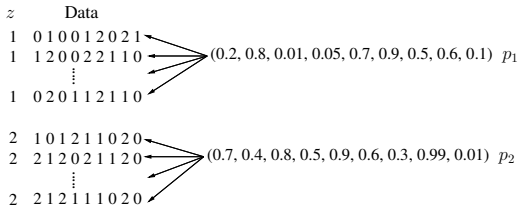
There are 2 general strategies for protecting against structure in case/control studies.

Structured Association

Basic Idea Try to infer (discover) the structure and then condition on the structure when testing for association.

Basic Model

- Assume we have data on N individuals at L (unlinked) loci.
- Assume that there are K underlying populations.
- Each individual has a parameter that indicates population membership, z_i.
- Each population has a set of parameters that specifies the allele frequencies at all of the markers, $\boldsymbol{p}_k = \{p_{k1}, \ldots, p_{kL}\}$, $k = 1, \ldots, K$



The model specifies $P(\text{Data}|Z, P, K)$

In practice we won't know the number of populations K so we want to calculate $P(K|\text{Data})$ which we can calculate (in principal) in the following way

$$P(K|\text{Data}) \propto P(\text{Data}|K) = \int P(\text{Data}|Z, P, K)\pi(Z)\pi(P)\text{d}Z\text{d}P$$

where $\pi(Z)$ and $\pi(P)$ are prior distributions on the parameters.

In practice, the integral is approximated using Markov Chain Monte Carlo (MCMC) techniques.

The program *structure* (Pritchard et al. (2000a)) uses a novel method of calculating this integral that seems to work very well on many data sets.

The actual model *structure* uses allows for admixture between populations.

More recently, the model has been extended to include a more realistic model of the correlations that occur between populations (Marchini and Cardon (2002))

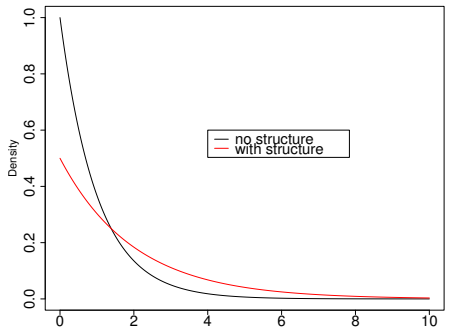
Once the number of populations has been inferred we can obtain an estimate of the structure Z.

Pritchard et al. (2000b) propose a likelihood ratio test (STRAT) to test the null hypothesis that the subpopulation allele frequencies are independent of the disease phenotype at a given marker.

Genomic Control

Basic Idea Correct the null distribution of the Chi-squared statistic for the effects of structure.

If we use the logistic regression score statistic to test association at a given locus population structure in or data will tend to skew the null distribution.



Devlin and Roeder (1999) show that the null distribution of this statistic in the presence of structure is

$$\lambda\chi^2_1$$

where the constant λ is a function of the structure present in the data.

These authors suggest estimating λ from the empirical distribution of the statistics across all loci tested.

Pritchard and Donnelly (2001) show that

$$\lambda \approx 1 + NF_{ST}\sum_{k=1}^{K}(d_k - c_k)^2$$

where d_k and c_k are the fractions of cases and controls from subpopulation k, N is the number of cases and controls and F_{ST} is a statistic that measures the level of structure between subpopulations.

Note λ increases with sample size.

Accuracy of Genomic Control

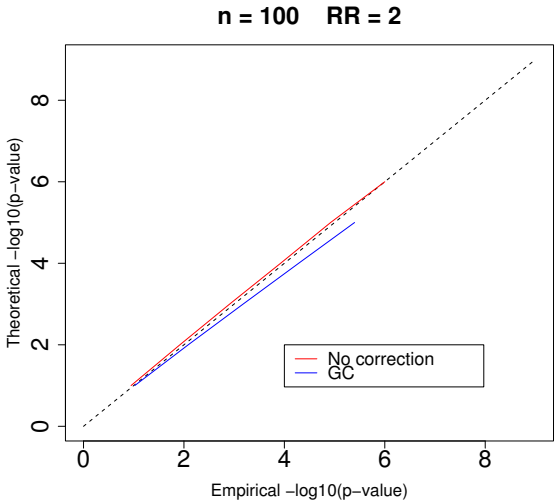
The accuracy of GC will depend on 2 factors

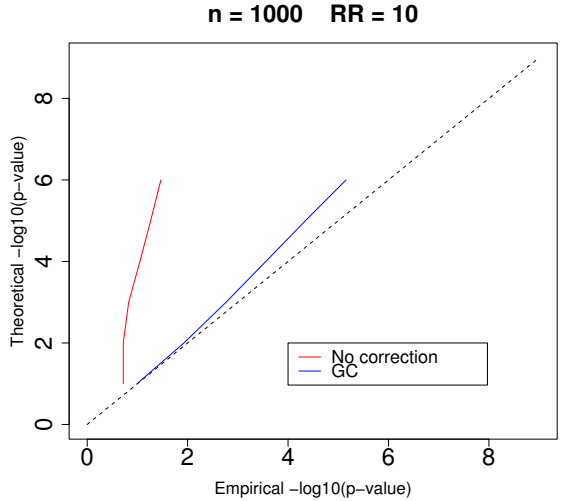
- how good the $\lambda\chi^2_1$ approximation is to the true null distribution.
- bias and variance in the estimation of λ

We have compared the empirical null distribution of GC to it's theoretical distribution using simulations from models fitted to real datasets.

We've found that the accuracy of GC can go both ways.

- In small samples the correction can be quite conservative.
- In large samples the correction can be quite liberal.





Power : STRAT vs GC vs TDT

R_1, R_2, R_3	L	P-value	STRAT	GC	TDT
1.5, 1.5, 1.5	200	.01	.29	.35	.45
	1000	.01	.30	.35	
	200	.001	.12	.14	.21
	1000	.001	.12	.14	
1.0, 1.0, 2.0	200	.01	.36	.22	.30
	1000	.01	.36	.22	
	200	.001	.16	.07	.12
	1000	.001	.16	.07	
0.5, 0.5, 1.5	200	.01	.64	.04	.06
	1000	.01	.66	.04	
	200	.001	.40	.008	.01
	1000	.001	.43	.005	

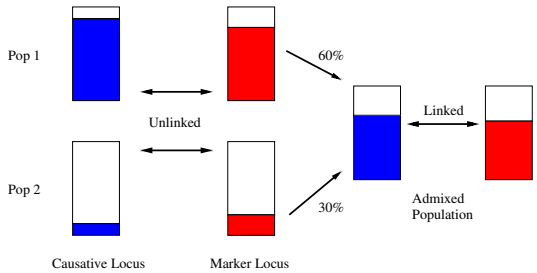
Table copied from Pritchard and Donnelly (2001)

MALD

Structure can be used to help map disease genes (Collins-Schramm et al. (2002))

Mapping by Admixture Linkage Disequilibrium (MALD) relies on recent admixture between 2 populations to create LD between markers that differ in allele frequency between the populations.

If the causative locus differs in allele frequency between 2 populations we can map the locus using a set of Ethnic Difference Markers (EDMs) spread throughout the genome.



Pros and Cons

	Pros	Cons
TDT	Protects against structure Need fewer markers than Case-Control	Need Parents Poor localization Doesn't use homozygous parents
MALD	Better power than TDT Need less markers than Case-Control	Relies on difference in disease between 2 populations Poor localisation
Unrelated Case-Control	Good localization	Need lots of markers Need to protect against structure

Other designs that protect against structure

Matched Controls We can protect against structure by selecting the controls to have similar ethnic and genetic backgrounds. Often ethnicity is self-reported and may not be accurate thus there may still be 'cryptic' structure present in the dataset.

Prospective Studies This design samples a large number of individuals and follows them through until disease onset. The long time span of such studies allows the collection of detailed environmental information that can help to elucidate the genetic/envirnomental basis of the disease. The large number of subjects followed potentially allows careful matching of controls to cases.

Probably the biggest such study is the UK BioBank project (http://www.ukbiobank.ac.uk/).

"Up to half a million participants aged between 45 and 69 years will be involved in the study. They will be asked to contribute a blood sample, lifestyle details and their medical histories to create a national database of unprecedented size. This combination of information from participants will create a powerful resource for biomedical researchers."

Pooling

Genotyping costs can quickly escalate in case-control designs.

2000 individuals at 1,000,000 SNP's will cost at least $20,000,000.

Pooling is a method that can estimate the sample allele frequency at a locus using a pooled sample of genetic material from a group of individuals

This method can be used to estimate allele frequency differences between groups of cases and controls.

Pooling has already been used successfully to map some disease genes (Risch and Teng (1998))

A disadvantage of this method is that we cannot investigate interaction effects between loci.

Multiple Comparisons

If we test 100,000 loci for association at the 5% level we should expect $\approx 5,000$ false positives. Thus, if we want the overall Type I error to remain at 5% we need to lower the significance level at each locus (Hochberg and Tamhane(1987))

Bonferroni Bounds Simply divide the significance level at each locus by the number of tests. If the tests are independent then the Bonferroni bound provides a slightly conservative bound. If the tests are correlated then the bound becomes more conservative.

Simulation We expect the loci to be correlated. If we can estimate this correlation then simulation based methods can be used to calculate appropriate thresholds.

Step down tests Test the ordered p-values against a sequence of increasing bounds that preserve the overall Type I error.(Hochberg and Tamhane(1987))

False Discovery Rate (FDR) The False Discovery Rate is the proportion of positive tests that are false. FDR procedures attempt to bound the expected False Discovery Rate. Such procedures often lead to lower thresholds. (Benjamini & Hochberg (1995))

Bayesian Modelling In a Bayesian framework the problem reduces to estimating the location of the disease loci rather than testing a hypothesis that there is a disease locus. A natural approach to take is to try to model the distribution of effects at the disease loci using a mixture modelling approach. As we expect only a few loci to be involved in the disease and we have limited prior information on the effect size this is non-trivial.

Searching for interaction effects

All the methods above implicitly assume that we can discover the loci involved in a complex disease by looking at their marginal effects at each locus.

Genetic models exist that have no marginal effect at the contributing loci (Culverhouse et al. (2002)) so this may not be sensible.

Less extreme genetic models of interactions predict a substantial decrease in the marginal effect size so it may well be worth search explicitly for the interaction effects.

Some References

Risch and Merikangas (1996) The Future of Genetic Studies of Complex Human Diseases. Science **273**:1516-17
Sham (1998) Statistics in Human Genetics. Arnold
Abecasis GR, Cherny SS, Cookson WO and Cardon LR (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97-101.
Risch (2001) Searching for genetic determinants in the new millenium. Nature 405, 847-56
Pritchard (2001) Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. 69:124-137
Pritchard et al. (2000a) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.
Pritchard et al. (2000b) Association mapping in structured populations. Am J. Hum Genet. 67:170-181.
Pritchard and Donnelly (2001) Case-control studies of association in structured or admixed populations. Theor. Pop. Biol. 60:227-237
Clayton (2001) Population Association. Chapter 19, Handbook of Statistical Genetics.
Sham and Curtis (1995) Monte Carlo tests for association between disease and alleles at highly polymorphic loci. Annals of Human Genetics 59, 97-105
Stephens et al. (2001) A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet., 68:978-989
Clayton and Jones (1999) TDT for extended marker haplotypes. Am. J. Hum. Genet. 65, 1161-1169
Seaman et al. (2002) A bayesian partition model for case-control studies on highly polymorphic candidate genes. Genetic Epidemiology 22:356-368
Speilman et al. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. Am. J. Hum. Genet. 52:506-516
Devlin and Roeder (1999) Genomic Control for association studies. Biometrics 55:997-1004
Risch and Teng (1998) The relative power of family-based and case-control designs for LD studies of complex human diseases I. DB+NA pooling. Genome Research 1273-88
Hochberg and Tamhane (1987) Multiple Comparison Procedures. Wiley
Benjamini and Hochberg (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. JRSS(B) 57:289-300
Culevrhouse et al. (2002) A perspective on epistasis: limits of models displaying no main effect. Am. J. Hum. Genet. 70:461-471