Statistical methods in genetic analysis

Heather J. Cordell

JDRF/WT Diabetes and Inflammation Laboratory

Department of Medical Genetics

University of Cambridge

Overview

- Allele sharing methods
  - Affected sib pairs
  - Affected relative pairs/sets
  - Sib pairs (quantitative phenotypes)
- Allele transmission methods
  - Transmission disequilibrium test (TDT)
  - Case/pseudo-control approach

Allele sharing methods

- Alleles in two or more individuals in a family are identical by descent (IBD) if inherited from the same common ancestor.
- Assuming no inbreeding, the prior probabilities of the IBD states for different relationships are:

	No. genes shared IBD		
	2	1	0
Relationship	f_2	f_1	f_0
Parent–Offspring	0	1	0
Half siblings	0	1/2	1/2
Full siblings	1/4	1/2	1/4
First cousins	0	1/4	3/4
Double 1st cousins	1/16	6/16	9/16
Second cousins	0	1/16	15/16
Uncle–nephew	0	1/2	1/2

- However, relatives who are phenotypically alike (e.g. both affeted with disease) will have inherited the same disease alleles from a common ancester.
- Hence they will share more alleles IBD at the disease locus (and at markers in the vicinity) than expected by chance.

Affected sib pair (ASP) studies

- Collect sample of sib pairs both affected with disease (plus parents if possible)
- Compare IBD sharing at specific locations in genome (e.g. candidate loci or at increments across genome) with null (0.25, 0.5, 0.25) values.
- Advantages:
  - Easy to collect
  - For early onset disease parents usually available
  - Specification of disease model not required (i.e. 'non-parametric'/model free)
- Disadvantages
  - Ignores other affected relatives if available
  - May be less powerful than parametric methods if true disease model is known.

Test statistics

- χ^2 goodness-of-fit test
  - Calculate the usual statistic

$$X = \sum_{i=0}^{2} \frac{(O_i - E_i)^2}{E_i}$$

where O_i = observed number of pairs (n_0, n_1, n_2) and E_i = expected number of pairs $(N/4, N/2, N/4)$ sharing i alleles IBD.

  - Only useful if IBD sharing known for each pair.

- Mean IBD test
  - Compares observed proportion of alleles shared by ASPs in sample, to that expected (0.5) under no linkage.
  - Test statistic is $[p - E(p)]/\sqrt{Var(p)}$.
  - Most powerful under wide range of genetic models.
  - Generalizable to situation when IBD sharing is uncertain via posterior probabilities $(\hat{f}_0,\ \hat{f}_1,\ \hat{f}_2)$
  - Hidden Markov models allow calculation of $(\hat{f}_0,\ \hat{f}_1,\ \hat{f}_2)$ at uninformative loci and at increments between loci (multipoint analysis).

- Likelihood ratio (LR) method:
  - Define unknown parameters $\underline{z} = (z_0, z_1, z_2)$ as

    z_i =P(affected pair share i alleles ibd)

    i.e. IBD probabilities conditional on affection status
  - Calculate likelihood $L(\underline{z})$

    (depends on family structure and genotypes).
  - Test null hypothesis $\mathrm{H}_0 : (z_0, z_1, z_2) = (0.25, 0.50, 0.25)$ using likelihood ratio test

    $$\mathrm{MLS} = \log_{10} \frac{L(\underline{\hat{z}})}{L(0.25, 0.50, 0.25)}$$

    where $\underline{\hat{z}} = (\hat{z}_0, \hat{z}_1, \hat{z}_2)$, the values of (z_0, z_1, z_2) that maximize the likelihood of the data.
  - Likelihood expressible as

    $$L(z) = \prod_j \left(\frac{z_0 \hat{f}_{0j}}{f_{0j}} + \frac{z_1 \hat{f}_{1j}}{f_{1j}} + \frac{z_2 \hat{f}_{2j}}{f_{2j}} \right)$$

    where f_{ij} is the prior probability and $\hat{f}_{ij}$ the posterior probability (given the observed marker data) that pair j share i alleles IBD. $(f_{0j}, f_{1j}, f_{2j}) = (0.25, 0.5, 0.25) \quad \forall j$

Distribution of test statistics

- Note test statistic defined in terms of $\log_{10}$ rather than $2\log_e$ (need to rescale: multiply by 4.6)
- Distribution also depends on whether maximization carried out subject to constraints on (z_0, z_1, z_2)
- Test statistics often called ‘lod’ regardless of number of free parameters, distribution, lack of correspondence with parametric lod score:

$$\text{LOD} = \log_{10} \frac{L(\hat{\theta})}{L(\theta = 0.5)}$$

where likelihood of data is expressed as function of recombination fraction θ between disease and marker loci, under assumed genetic model.

Problems

- For complex traits (small effects), ASP methods give notoriously poor localisation of disease loci unless sample sizes large (500 or more pairs)
- True disease location may lie 15-20cM from linkage peak.
- Calculating accurate confidence intervals for location is still a (somewhat) unsolved problem: depends on which statistic is being used, marker informativity etc.
- Large consortia being established to generate sufficient data (but note problems of heterogeneity between study centres)

Extensions to ASP approaches

- Aim to improve informativity or power...
- Multilocus models: model joint IBD sharing at several loci.
- Incorporate IBD information, linkage information or alleleic association at one locus into test statistic for another locus.
- More generally, incorporate covariates into test statistic at a locus.

Multilocus models

- For two loci, define sharing probabilities z_{ij} $(i, j = 0, 1, 2)$

$$\begin{array}{lcc} & & \text{Locus 2} \\ \text{Locus 1} & & \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ & Z = & \begin{pmatrix} z_{00} & z_{01} & z_{02} \\ z_{10} & z_{11} & z_{12} \\ z_{20} & z_{21} & z_{22} \end{pmatrix} \end{array}$$

$$\mathrm{MLS} = \log_{10} \frac{L(\mathrm{data}|\hat{Z})}{L(\mathrm{data}|Z_{\mathrm{null}})}$$

- Likelihood formulation:

$$L \propto \prod_k \left(\sum_{i=0}^{2} \sum_{j=0}^{2} \frac{z_{ij}\hat{f}_{ijk}}{f_{ijk}} \right)$$

where z_{ij}, $\hat{f}_{ijk}$ and f_{ijk} refer to the probabilities that pair k shares i alleles at locus 1 and j alleles at locus 2 simultaneously.

- m locus model:

$$L \propto \prod_k \left(\sum_{i_1=0}^{2} \sum_{i_2=0}^{2} \cdots \sum_{i_m=0}^{2} \frac{z_{i_1 i_2 \ldots i_m} \hat{f}_{i_1 i_2 \ldots i_m k}}{f_{i_1 i_2 \ldots i_m k}} \right)$$

where $z_{i_1 i_2 \ldots i_m}$, $\hat{f}_{i_1 i_2 \ldots i_m k}$ and $f_{i_1 i_2 \ldots i_m k}$ refer to the same sharing probabilities but at the m loci simultaneously.

Null Hypotheses

- No linkage at either locus

$$\begin{array}{ll} & \text{Locus 2} \\ \text{Locus 1} & \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ Z = & \begin{pmatrix} 0.0625 & 0.125 & 0.0625 \\ 0.125 & 0.25 & 0.125 \\ 0.0625 & 0.125 & 0.0625 \end{pmatrix} \end{array}$$

(Can adjust to allow for linkage between loci)

- Only strongest locus linked

$$\begin{array}{ll} & \text{Locus 2} \\ \text{Locus 1} & \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ Z = & \begin{pmatrix} 0.25z_0 & 0.5z_0 & 0.25z_0 \\ 0.25z_1 & 0.5z_1 & 0.25z_1 \\ 0.25z_2 & 0.5z_2 & 0.25z_2 \end{pmatrix} \end{array}$$

  - Tests effect at locus 2 ‘taking into account’ effect at locus 1 (and any interaction between the loci)

- Specific ‘biological’ models for z_{ij}
  - heterogeneity (independent pathways)
  - multiplicative (epistatic)

- Express z_{ij} in terms of relative risk parameters λ_s, λ_{ij}.

- Express λ’s in terms of covariance between sibs, hence in terms of underlying genetic additive and dominance variance parameters. Biological models imply certain restrictions on variance parameters.

Analysis of type 1 diabetes data set

- 356 ASPs (with parents) typed across genome

Table 1: Maximum MLS values and conditional MLS values (with p values) for selected chromosomes. Results are given for a stepwise procedure consisting of a single locus analysis followed by a two locus analysis conditional on *IDDM1*, and finally a three-locus analysis conditional on *IDDM1* and *IDDM10*.

Chromosome	Closest marker	Location	Single locus		Two-locus		Three-locus	
	(or IDDM locus)	on Figs			conditional		conditional	
		5 and 6	MLS	p value	MLS	p value	MLS	p value
3	*D3S1576*	180 cM	1.01	0.03	1.28	0.04	2.88	0.004
6	*D6S291* (*IDDM1*)	29 cM	34.7	HS	-	-	-	-
6	*D6S294-D6S286*	56 cM	19.4	HS	2.42^a	0.001	2.60	0.008
8	*D8S88*	111 cM	0.70	NS	1.62	0.03	2.25	0.01
10	*D10S220* (*IDDM10*)	51 cM	4.67	0.000004	5.02	0.000008	-	-
11	*TH/INS* (*IDDM2*)	3 cM	2.77	0.0003	4.14	0.00006	5.17	0.0002
11	*FGF3* (*IDDM4*)	81 cM	0.54	NS	2.04^a	0.002	1.97^a	0.003
14	*D14S75-D14S276*	43 cM	1.95	0.002	2.42	0.003	2.83	0.004
15	*CYP19-D15S125*	39-57cM	0.74	0.05	1.12^a	0.02	1.72^a	0.005
16	*D16S3098*	87 cM	3.24	0.0001	4.92	0.00001	5.02	0.0002
18	*D18S487*	72 cM	1.10	0.02	1.95^a	0.002	1.98^a	0.003
19	*D19S226*	24 cM	1.80	0.004	1.96	0.02	2.18	0.02
21	*D21S120*	5 cM	0.06	NS	0.95	0.07	1.59	0.04
Pseudo-autosomal	*DXYS154*	33 cM	1.23	0.02	1.65^a	0.005	1.12^a	0.02

[a] Multilocus results are given for the general model except those for those marked [a] which are for the additive model.

NS=not significant (p value > 0.05). HS=highly significant (p value < 0.000001)

Alternative methods

- Cox et al. (1999) Nat Genet 21: 213-215: weight families in a single-locus analyses of locus 2 according to their evidence of linkage at locus 1.
  - Optimal weighting depends on underlying genetic model
- Include IBD sharing or test statistic at first locus as covariate
  - Extends to including other covariates e.g. genotype (combination) at first locus, gender (combination), parent-of-origin, environmental effects.
  - Issues with choice of covariates, choice of coding scheme.

Covariate methods

- Rice (1997):
  - Model IBD sharing as

$$z_2 = p^2, \quad z_1 = 2p(1-p), \quad z_0 = (1-p)^2$$

  - Model

$$\log \frac{p}{1-p} = \alpha + \beta^T x$$

    where x is a vector of covariates. Null is $\alpha = 0$, $\beta = 0$.
  - Use usual Risch likelihood

$$L(z) = \prod_j \left(\frac{z_0 \hat{f}_{0j}}{f_{0j}} + \frac{z_1 \hat{f}_{1j}}{f_{1j}} + \frac{z_2 \hat{f}_{2j}}{f_{2j}} \right)$$

- Olson (1999)
  - Reparameterize Risch likelihood as

$$\prod_j \left(\frac{\hat{f}_{0j} + e^{\beta_1} \hat{f}_{1j} + e^{\beta_2} \hat{f}_{2j}}{f_{0j} + e^{\beta_1} f_{1j} + e^{\beta_2} f_{2j}} \right)$$

  - Incorporate covariates via 2 parameters δ_1, δ_2

$$\prod_j \left(\frac{\hat{f}_{0j} + e^{\beta_1 + \delta_1 x} \hat{f}_{1j} + e^{\beta_2 + \delta_2 x} \hat{f}_{2j}}{f_{0j} + e^{\beta_1 + \delta_1 x} f_{1j} + e^{\beta_2 + \delta_2 x} f_{2j}} \right)$$

Additional comments

- See also methods proposed by Greenwood and Bull (1999), Gauderman and Siegmund (2000), Goddard et al (2001).
- Properties of different methods have not been compared.
- Holmans (2002): compared utility of conditioning on linkage peak (e.g. via covariates based on IBD sharing) vs conditioning on disease-associated genotypes.
  - Disease-associated genotypes generally more useful.

Affected relative pairs (ARPs) or sets

- Affected sib pairs convenient sampling unit.
- If other types of relative available, makes sense to use them.
- Large pedigrees often collected for traditional linkage studies.
- Like ASP methods, idea is to compare observed IBD sharing with that expected under no linkage.
- Several of ASP methods extend quite naturally to ARPs.
  - Mean IBD test
  - MLS method
  - Olson/Rice covariates method

Extension of MLS method to ARPs

- Recall for ASPs we define $\underline{z} = (z_0, z_1, z_2)$ as

  z_i =P(affected pair share i alleles ibd)
- Test null hypothesis $\mathrm{H}_0 : (z_0, z_1, z_2) = (0.25, 0.50, 0.25)$ using likelihood ratio test:

$$\mathrm{MLS} = \log_{10} \frac{L(\hat{\underline{z}})}{L(0.25, 0.50, 0.25)} = \log_{10} \mathrm{LR}$$

- Formula for likelihood:

$$L \quad = \quad \prod_j \left(\frac{z_0 \hat{f}_{0j}}{f_{0j}} + \frac{z_1 \hat{f}_{1j}}{f_{1j}} + \frac{z_2 \hat{f}_{2j}}{f_{2j}} \right)$$

  f_{ij} = prior probability and $\hat{f}_{ij}$ = posterior probability (given the marker data) that pair j share i alleles IBD.
- Note that for sibs $\quad (f_{0j}, f_{1j}, f_{2j}) = (0.25, 0.5, 0.25) \quad \forall j$
- For ARP of arbitrary relationship, we use same formula, but z_i, f_{0j}, f_{1j}, f_{2j} vary depending on relationship.

  (z_i function of relationship and underlying additive and dominance variances of disease model, σ_a^2, σ_d^2).

- Multipoint f_{ij}, $\hat{f}_{ij}$ output from standard programs

  e.g. Genehunter, Allegro, Genibd, Simwalk2

- Similar extension can be used for Olson (and Rice?) covariate approaches.

- Pairs from same family not independent.

  - Evaluate LR test statistic using simulation.
  - Or use score test with robust variance estimation?

Non-parametric linkage (NPL) methods

- Generalization of mean IBD test.

- Based on variety of proposed scoring statistics (Whittemore and Halpern 1994)

- Idea is to produce score based on IBD sharing amongst affected individuals in a pedigree

  - Pairwise IBD sharing (NPL $_{\text{pairs}}$)
  - IBD sharing amongst whole set (NPL $_{\text{all}}$)

- Generate normalised score for each pedigree: combine to produce overall test statistic.

- Calculation of mean and variance of pedigree-specific scores under null hypothesis not trivial: requires ennumeration of all possible inheritance vectors.

- Initial packages (e.g. Genehunter) used ‘perfect data’ approximation: assumed IBD sharing unambiguous at every location.

Allele-sharing models (Kong and Cox 1997)

- Linear model: Construct likelihood assuming

$$P(\nu_i = \nu | \delta) = P(\nu_i = \nu)(1 + \delta w_i Z_i)$$

  ν_i denotes underlying inheritance vector for pedigree i

  w_i a pedigree-specific weight

  $Z_i(\nu_i)$ is the normalised score for pedigree i

  δ is parameter to be estimated representing magnitude of deviation from null sharing.

- Exponential model:

$$P(\nu_i = \nu | \delta) = P(\nu_i = \nu) r_i(\delta) \exp(\delta w_i Z_i)$$

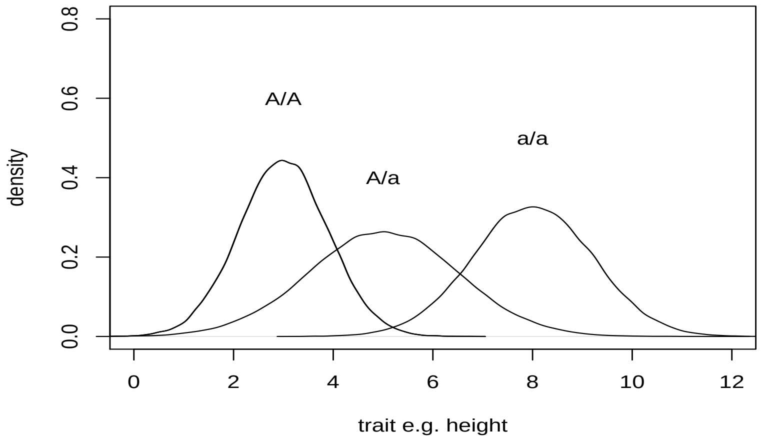
  where $r_i(\delta)$ is normalization constant.

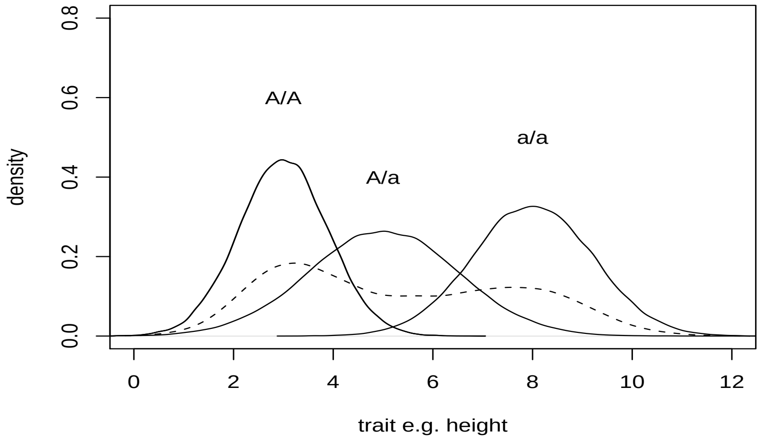
- Score test from these models = NPL statistic (when data fully informative).
- Kong and Cox propose using LR test of null hypothesis $\delta = 0$.

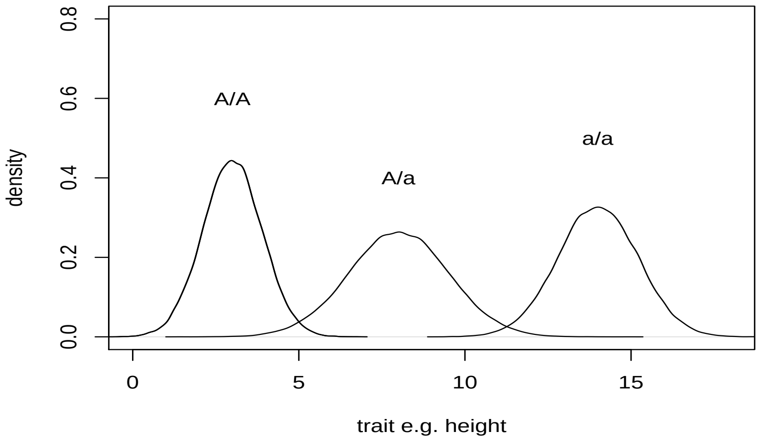
- ASM statistics less conservative than NPLs from Genehunter
- Implemented in Genehunter-Plus, Allegro, Merlin.
- Pedigree specific weights allow ASM (and NPL) methods to weight tests at one locus according to IBD sharing, genotypes at other locus (or according to other covariates).

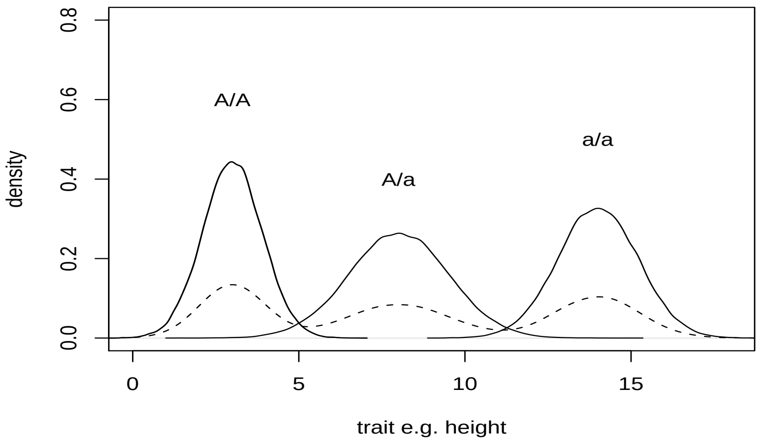
Sib pair methods for quantitative traits

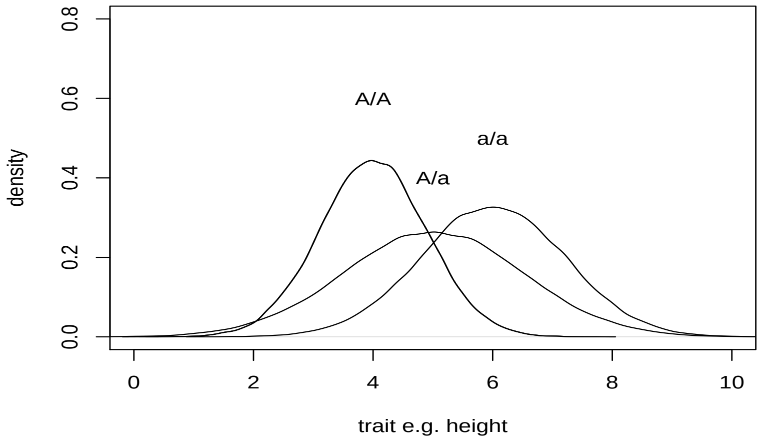
- Affected sib pairs: dichotomous trait (affected/unaffected)
- Suppose instead we are interested in genes influencing a continuous (quantitative) trait
  - Blood pressure
  - Height
  - Obesity/BMI
  - Immune response
  - Age of onset of disease (survival methods?)
- Idea is that genotype at one or more loci influences mean (and possibly variance) of trait distribution.

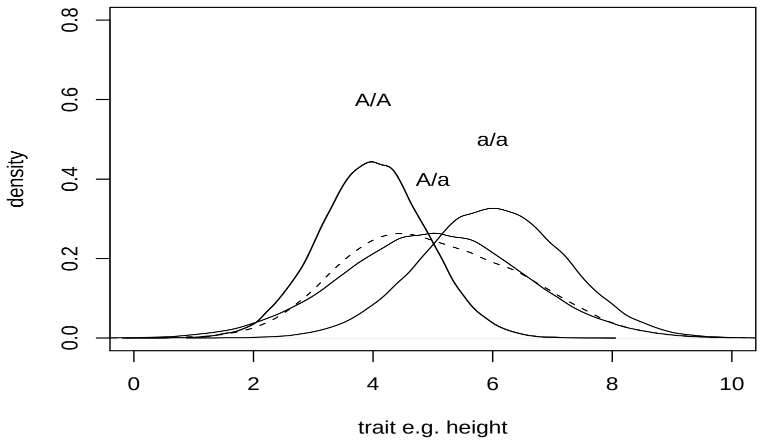





A/A
A/a
a/a
density
0.0
0.2
0.4
0.6
0.8
0
5
10
15
trait e.g. height


A/A
A/a
a/a
density
0.0
0.2
0.4
0.6
0.8
0
5
10
15
trait e.g. height

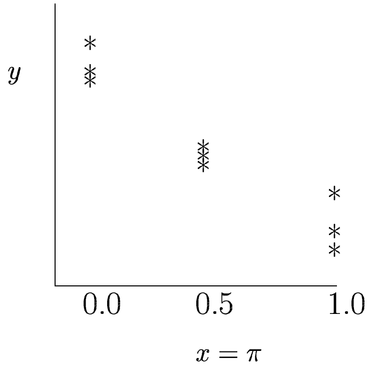

A/A
a/a
A/a
density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
trait e.g. height


A/A
a/a
A/a
density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
trait e.g. height

Haseman-Elston (H-E) Method

- Haseman and Elston (1972) Behav Genet 2:3-19
- Idea is to look at trait difference squared for pairs of sibs.
- E.g. Sib 1 has trait value 14.5, sib 2 has trait value 10.2.
- Difference = 4.3, difference squared = $4.3^2 = 18.49$.
- Difference squared is a measure of how phenotypically dissimilar the two sibs are.
- If a genetic locus is responsible for trait
  - Sibs with similar trait values likely to have inherited same allele(s) at this locus from parents.
  - Sibs with differing trait values likely to have inherited different allele(s) at this locus from parents.

- Small values of difference squared suggest
  - Sibs have similar trait values
  - Inherit same alleles from parents at trait locus
  - Share more alleles IBD than expected at trait locus and at linked markers in surrounding region.
- Large values of difference squared suggest
  - Sibs have differing trait values
  - Inherit different alleles from parents at trait locus
  - Share less alleles IBD than expected at trait locus and at linked markers in surrounding region.
- Look at relationship between sib-pair difference squared, and number or proportion of alleles shared IBD, in large sample of sib pairs.



- $y =$ sib pair difference squared
- $x = \pi =$ proportion of alleles shared IBD

$$\begin{aligned} \pi &= 0 && \text{0 alleles IBD} \\ \pi &= 0.5 && \text{1 allele IBD} \\ \pi &= 1.0 && \text{2 alleles IBD} \end{aligned}$$

- To test for linkage, fit regression line $y = mx + c$
  - Under null, slope $m = 0$.
  - Under alternative, slope $m < 0$.
- Test using standard stats/genetics package.

Mathematical details

- Let x_{1j} and x_{2j} be the trait values for sib pair j. We assume

$$\begin{aligned} x_{1j} &= \mu + g_{1j} + e_{1j} \\ x_{2j} &= \mu + g_{2j} + e_{2j} \end{aligned}$$

  where μ is the overall mean, g_{ij} and e_{ij} are genetic and environmental effects.

- Suppose single diallelic locus involved, $g_{ij} = a, d, -a$ for BB, Bb, bb individuals.
- Genetic variance $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$ where under random mating

$$\sigma_a^2 = 2pq[a - d(p - q)]^2, \quad \sigma_d^2 = 4p^2q^2d^2$$

  and p, q are allele frequencies of B and b.

- Let $e_j = e_{1j} - e_{2j}$, $E(e_j) = 0$, $E(e_j^2) = \sigma_e^2$.
- Let $y_j = (x_{1j} - x_{2j})^2$, the sib-pair difference squared.
- Let $\pi_i = 0, 0.5, 1$ be the proportion of alleles shared IBD at the trait locus.

- Haseman and Elston (1972) show that

$$E(y_j|\pi_j) \approx (\sigma_e^2 + 2\sigma_g^2) - 2\sigma_g^2\pi_j = \alpha + \beta\pi_j$$

- Approximation exact if no dominance.
- Note similarity to regression equation $y = mx + c$ where

$$\begin{aligned} y &\equiv (x_1 - x_2)^2 \\ m &\equiv \beta \\ x &\equiv \pi \\ c &\equiv \alpha = (\sigma_e^2 + 2\sigma_g^2) \end{aligned}$$

- Null hypothesis of no linkage can be tested by performing linear regression, testing whether $\beta \equiv -2\sigma_g^2 = 0$.
- Test statistic: use t statistic $\frac{\hat{\beta}}{sd(\hat{\beta})} \sim N(\beta, 1)$.
- σ_g^2 estimated by $-\hat{\beta}/2$.
- Test can be generalized to use $\hat{\pi}_j$, the *estimated* proportion of alleles shared IBD, instead of π_j.

Some extensions

- Qualitative traits ⇒ code 0/1 for unaffected/affected
- *Estimation* of genetic parameters assumes underlying normality, random ascertainment. *Test* of null of no linkage valid without these assumptions.
- H-E revisited (Elston et al. 2000, Genet Epid 19:1-17): use combination of mean-corrected trait sum-squared and trait difference-squared $y = \frac{1}{4}(Y_S - Y_D)$
  - improvement in power in certain circumstances.
- Weighted/unified/combined H-E (Xu et al. 2000, AJHG 67:1025-1028; Forrest 2001, Hum Hered 52:47-54; Sham and Purcell 2001, AJHG 68: 1527-1532)
  - Use weighted combinations of trait sum-squared and trait difference-squared measures.

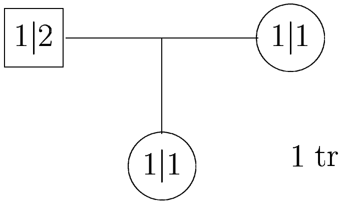
- Sham et al. (2002) AJHG 71:238-253.
  - Extension that applies to general pedigrees
  - Regression is 'other way round': appropriate for selected samples
  - Appears to combine robustness of H-E with power of variance components
  - But requires estimate of population mean, variance, heritability (or correlation between different relationship types)

Allele transmission methods

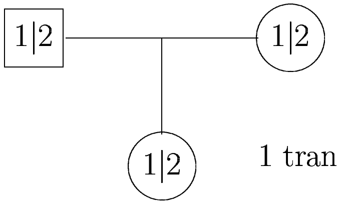
- Transmission disequilibrium test (TDT)
- Idea is to examine transmission of specific alleles from parent(s) to affected child.
  - Sample families on the basis of single affected offspring.
  - Affected offspring and both parents genotyped.
- Under null hypothesis of no linkage or no association ($\theta = 0.5$ or $\delta = 0$) parents should transmit either of their two alleles to child with equal probability.
- If not → linkage AND association ($\theta < 0.5$ and $\delta \neq 0$)
- Originally conceived as test of linkage in presence of association.
- Often used as test of association in presence of linkage (needs care).

- TDT counts transmissions of allele 1, say, from heterozygous parents to an affected child

- Only heterzygous parents used.



1 transmission



1 transmission, 1 non-transmission

- If altogether T transmissions and N non-transmissions;

$$\text{TDT} = \frac{(T-N)^2}{T+N} \sim \chi^2_1$$

- Parents considered independent: true under null of no association and under some alternatives (e.g. if the genetic association follows a multiplicative model for the effects of the alleles on penetrance)

- The data can be arranged as a 2 × 2 table:

Allele	Transmitted	
Untransmitted	1	2
1	a	b
2	c	d

- The test of association is McNemar's test:

$$\frac{(b-c)^2}{b+c} \overset{\text{Asymptotically}}{\sim} \chi^2_1$$

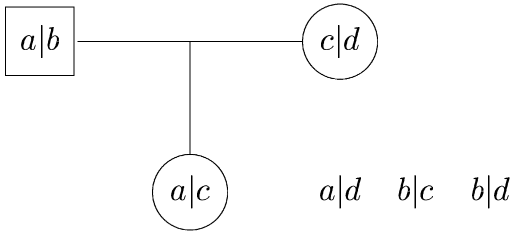
- Transmitted 'case' allele is matched to untransmitted 'control' allele

Extensions

- Multiallelic TDTs (many df) (Sham and Curtis 1995; Bickeboller and Clerget Darpoux 1995; Cleves et al 1997)
- Missing parents (Curtis and Sham 1995; Knapp 1999): RC-TDT, S-TDT, Sib-TDT
- Haplotypes
  - TDTPHASE (Dudbridge)
  - TRANSMIT (Clayton 1999)

Case/Pseudo-control methods

- Genotypes constructed for 3 ‘pseudo-controls’, consisting of other possible genotypes that offspring could have received.



- Data analysed as if real matched case/control sample.
- Why does this work? (Self et al. 1991; Schaid 1996): Consideration of *conditional* likelihood, conditional on parental genotypes and fact that off spring is affected.
- Condition on affected offspring through ascertainment scheme.
- Conditioning on parental genotypes:
  - removes spurious effects e.g. due to population stratification
  - avoids estimating nuisance parameters such as parental mating type frequencies.

- Let g_c, g_m, g_f be the genotypes of the child, mother and father, and let D denote the event that the child is affected.
- Then

$$P(g_c|g_m, g_f, D) \;=\; \frac{R(a/c)}{R(a/c)+R(a/d)+R(b/c)+R(b/d)}$$

where R denotes the disease risk for a genotype relative to some arbitrary baseline genotype e.g. relative to a/a.

- This is identical to the likelihood used in matched case/control studies for a case with genotype a/c matched to three controls with genotypes a/d, b/c, b/d.
- Analysed via conditional logistic regression with genotype indicator variables as the predictors of outcome (disease).

Genotype relative risks

- Can test and estimate risks conferred by the various genotypes using this procedure.
- Null hypothesis is usually $R(i/j) = 1$ for all genotypes i/j
- One may reduce number of parameters under alternative by making assumptions e.g. multiplicative effects of alleles i, j $R(i/j) = R_i R_j$ (true under null)
- Then

$$\begin{aligned} & \frac{R(a/c)}{R(a/c)+R(a/d)+R(b/c)+R(b/d)} \\ = & \frac{R_aR_c}{R_aR_c+R_aR_d+R_bR_c+R_bR_d} \\ = & \frac{R_a}{R_a+R_b} \times \frac{R_c}{R_c+R_d} \end{aligned}$$

- Multiplicative effect of alleles ⇒ independent contributions from each parent.
- Score test of $R_i = 1 \quad \forall i \quad \equiv$ TDT.

Several linked loci

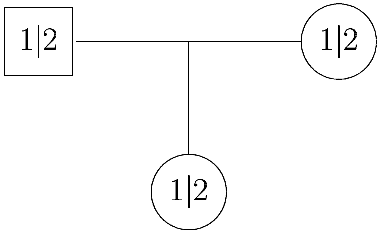
- We recently extended this method to evaluating the effects of several closely-linked loci in a region (Cordell and Clayton 2002)
  - May be more than one causal locus in region
  - Causal locus may lie on ancestral haplotype marked by several loci.
- Enter variables coding effects at each locus in stepwise manner in conditional logistic regression equation.
- Mimics standard epidemiological procedures for real case/control studies via logistic regression.
- Can test effect of second locus once first locus has been accounted for (i.e. already entered into equation).

Further extensions

- Multiple linked loci in multiple unlinked regions
- Parent-of-origin (imprinting) effects
- With more than one locus in a region we have the problem of phase uncertainty.
- E.g. individual with genotypes a/A, b/B

$$\begin{array}{c|c} a & A \\ b & B \end{array} \qquad\qquad \begin{array}{c|c} a & A \\ B & b \end{array}$$

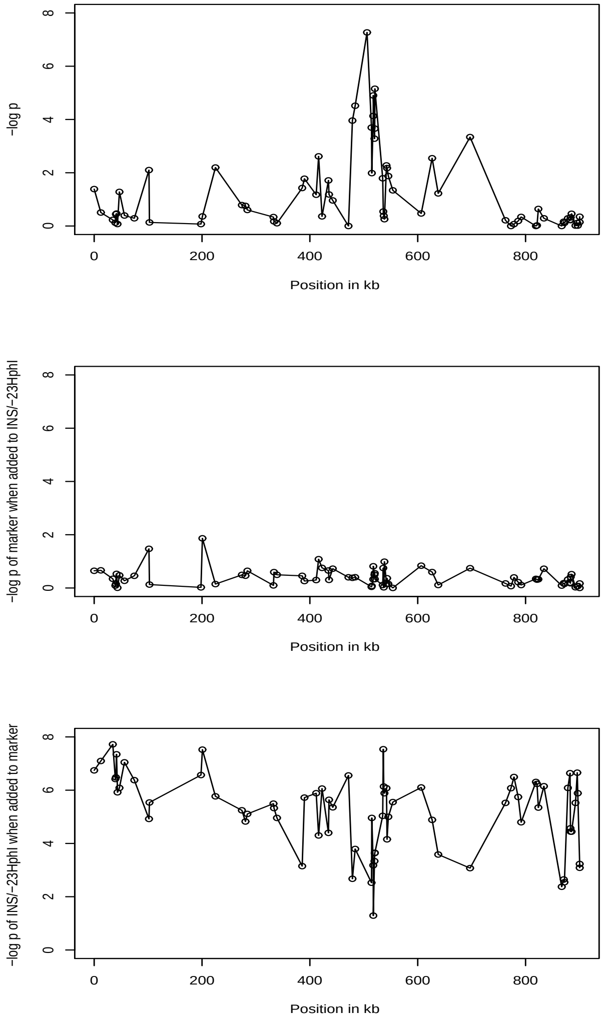
- Also issues of uncertainty in parent-of-origin



General approach

- Use modified conditioning argument to construct set of pseudo-controls for every case (affected child) in sample.
- Exact conditioning argument depends on what genotype relative risk models are to be fitted (e.g. whether risks depend on phase, parent-of-origin etc.)
- Analyse as matched case/control sample using conditional logistic regression software.

Example: INS region in Type 1 diabetes


−log p
Position in kb
−log p of marker when added to INS/−23HphI
Position in kb
−log p of INS/−23HphI when added to marker
Position in kb