Statistical Issues and Analysis Methods in Genetic Epidemiology

Three main areas of genetic epidemiology:

1. **Segregation analysis:**
   Does the trait have a genetic component ?
   If so, what (dominant, recessive, polygenic ...)
2. **Linkage analysis:**
   Where are the disease genes ? (approximate location)
3. **Association analysis:**
   Where are the disease genes ?
   (more precise location)

Segregation Analysis (very brief)

Aim: To decompose covariance of trait among pairs of relatives into components corresponding to major genes, polygenic, common environment, random effects.
To estimate the *heritability* of the trait (proportion of covariance due to genetic effects)

- **Twin studies** Compare correlation of trait among MZ (identical) twins to that among DZ (non-identical) twins.
- **Adoption studies** Compare correlation of trait between adopted proband and blood relatives (usually siblings) raised apart to that between proband and adoptive siblings.
  Useful for comparison of genetic and common-environment effects.
- **Pedigree studies** Pedigrees ascertained through trait values of proband(s).
  - Express likelihood of observed pedigree data in terms of genetic model parameters (gene frequencies, penetrances, polygenic heritability etc.). Compare likelihoods to obtain most likely model (+ parameter estimates).
  - Model correlations via path analysis.
  - Important to allow for effects of *ascertainment*.

Linkage Analysis

Interested in testing whether marker locus is *linked* to disease locus (i.e. if recombination fraction $\theta < 0.5$).

- Test θ directly (*parametric* methods)
- Use marker allele sharing among individuals with similar trait values as a surrogate for θ (*non-parametric* or *model-free* analyses).
- Different methods exist for *binary* traits (e.g. affected/unaffected) and *quantitative* traits.

Parametric methods for Binary traits

Express joint likelihood of observed marker and disease data in terms of
p, the frequency of the disease allele
f_{dd}, f_{Dd}, f_{DD}, the penetrances of the disease genotypes
θ, the recombination fraction between disease and marker loci

Denote the marker data by $\underline{M}$, the (unobserved) disease genotypes by $\underline{G}$ and the disease status of the individuals by $\underline{D}$.

$$L(\theta) = \sum_{\underline{G}} \Pr(\underline{M}|\underline{G}, \theta) \Pr(\underline{G}|p, \underline{f}) \Pr(\underline{D}|\underline{G})$$

$$\text{LOD score} = \log_{10}\left(\frac{L(\theta)}{L(\theta = 1/2)}\right)$$

If there is only one marker locus, maximise with respect to θ.

Question: What values of p and $\underline{f}$??

$$\text{MOD score} = \max_{\theta, p, \underline{f}} \log_{10}\left(\frac{L(\theta, p, \underline{f})}{L(\theta = 1/2, p, \underline{f})}\right)$$

-Computationally intensive !

More usual to analyse under a number of fixed values of p and $\underline{f}$.

Extensions of parametric analysis

Heterogeneity

- Not all pedigrees necessarily segregate the disease gene under study.
- In these pedigrees, θ will effectively be $1/2$.
- Including these pedigrees in the analysis will reduce power (and bias the estimate of θ upwards).
- Unfortunately, can't usually distinguish these pedigrees *a priori*.
- Let α = prior probability that a pedigree is "linked".

$$L(\theta, \alpha) = \alpha L(\theta) + (1 - \alpha) L(\theta = 1/2)$$

- *H*eterogeneity *LOD* score (HLOD) =

$$\log_{10} \left(\frac{L(\theta, \alpha)}{L(\theta = 1/2, \alpha)} \right)$$

  Maximise with respect to θ, α.

Extensions of parametric analysis

Multipoint Analysis (i.e. multiple marker loci)

Assume that disease locus is at fixed point x on chromosome.

$$L(x) = \Pr(\underline{M}|\underline{g}, \theta_{1x}, \theta_{x2}, \theta_{23}) \Pr(\underline{D}|\underline{g}) \Pr(\underline{g}|p, \underline{f})$$

Move x along chromosome, pick the highest lod score.

Note that distribution of resulting maximum lod score must be estimated by *simulation*.

Strengths and Weaknesses of Parametric Analysis

Strengths

- Most powerful form of analysis *if* the disease model is specified correctly (or almost correctly).
- Deals well with between-pedigree *heterogeneity* because it can be modelled directly (unlike model-free methods).
- Provides a direct estimate of θ (but see below).

Weaknesses

- If model specified incorrectly, power loss can be great (although Type I error probability not increased), and estimate of θ biased (upwards).
- For complex traits, model rarely known.
- Therefore, necessary to use a number of models - multiple testing.
  Estimates of θ will be unreliable.

Parametric Analysis for Model-free People

(MMLS-C method of Greenberg & Hodge)

- 2 disease models, one dominant, one recessive.
  Dominant: $p(D)$=0.005, f_{dd}=0, f_{Dd}=0.5, f_{DD}=0.5
  Recessive: $p(D)$=0.1 f_{dd}=0, f_{Dd}=0, f_{DD}=0.5
- For a complex (low-penetrance) trait, use *affecteds-only* analysis (i.e. code unaffecteds as unknown).
- Test statistic = highest LOD score over the two models.

Points to note

- Large single-locus effects unlikely to fit the observed inheritance patterns of the marker data, even in affecteds.
- For one marker locus, the effect can be weakened if necessary by increasing θ.
- For a multipoint analysis, this won't work (will bump into the next marker), so use the HLOD.
- This method is useful only for *testing* linkage - *estimates* of θ and/or α will be highly unreliable.
- Important to use disease models with large effects. If the effect is too large, it can be weakened by varying θ or α. If the effect is too small, the linkage information is lost forever...
- Performance of MMLS-C relative to model-free methods still being evaluated. MMLS-C will probably be good for large pedigrees.

Model-free (non-parametric) methods

- Based on the observed *sharing* of marker alleles among affected individuals in pedigrees.
- Increased sharing (compared to expected values) indicates linkage of marker locus to disease locus.
- No need to specify disease models (so no multiple testing)
- Less powerful than model-based methods when the model is known
- No estimates of model parameters (e.g. θ) although multipoint analysis will give an estimate of disease-locus location.
- **Note:** Many "model-free" methods have been shown to be asymptotically (or exactly) equivalent to analysing under certain parametric models
  - so some "model-free" methods aren't that model-free !!

Affected sib-pair methods

Sib pairs most common pairs of affected relatives. especially for complex traits (recurrence rates drop quickly as relationships get more distant) - large sample sizes.

Two main kinds of analysis:

1. **"Allele-counting" methods**
2. **Likelihood-ratio methods**

Assume (for now) IBD known with certainty.

"ALLELE - COUNTING" methods

1. **Mean Test**.
   - Denote no. of alleles shared IBD by i'th pair by Z_i.
   - In absence of linkage,
     $\Pr(Z_i = 0) = \Pr(Z_i = 2) = 1/4$, $\Pr(Z_i = 1) = 1/2$
   - $E(Z_i) = 0.25 \times 2 + 0.5 \times 1 + 0.25 \times 0 = 1$
     $E(Z_i^2) = 0.25 \times 4 + 0.5 \times 1 = 3/2$. So $Var(Z_i) = 1/4$.
   - $$\frac{\Sigma_{i=1}^{N} Z_i - N}{\sqrt{N/4}} \sim N(0,1)$$
2. **2-allele test**
   - Denote $Z_i = 1$ if i'th pair shares 2 IBD, $Z_i = 0$ otherwise.
   - $E(Z_i) = 1/4$. $Var(Z_i) = 3/16$.
   - $$\frac{\Sigma_{i=1}^{N} Z_i - N/4}{\sqrt{3N/16}} \sim N(0,1)$$
3. **Chi-squared Test**
   Count up number of pairs sharing 0,1,2 ibd and compare to expected values $(N/4, N/2, N/4)$ vi usual Pearson chi-square statistic:
   $$\sum_{i=0}^{2} \frac{(O_i - E_i)^2}{E_i} \sim \chi_2^2$$

Likelihood-ratio tests (Risch 1990)

- Suppose n_0 pairs share 0 IBD, n_1 share 1, n_2 share 2.
- Express likelihood in terms of (unknown) IBD probabilities z_0, z_1, z_2:
$$L(\underline{z}) \propto z_0^{n_0} z_1^{n_1} z_2^{n_2}$$
- Maximise with respect to $\underline{z}$ and form log-likelihood ratio statistic
$$LR = 2 \ln \left(\frac{L(\hat{\underline{z}})}{L(z_{null})} \right)$$
- Unconstrained maximisation ($\hat{z}_0 + \hat{z}_1 + \hat{z}_2 = 1$):
$LR \sim \chi^2_2$.
- $\hat{z}_1 = 1/2, \hat{z}_0 \leq 1/4$. (Score test = mean test)
$LR \sim \chi^2_1$ with probability 0.5,
$LR = 0$ with probability 0.5
- $2\hat{z}_0 = \hat{z}_1, \hat{z}_0 \leq 1/4$. (Score test = 2-allele test)
$LR \sim \chi^2_1$ with probability 0.5,
$LR = 0$ with probability 0.5
- $2\hat{z}_0 \leq \hat{z}_1, \hat{z}_0 \leq 1/4$. ("Possible triangle" test)
$LR \sim \chi^2_2$ with probability 0.1,
$LR \sim \chi^2_1$ with probability 0.5,
$LR = 0$ with probability 0.4
- $z_0 = (1-p)^2$, $z_1 = 2p(1-p)$, $z_2 = p^2$.
Maximise w.r.t. $p(\geq 1/2)$.
$LR \sim \chi^2_1$ with probability 0.5,
$LR = 0$ with probability 0.5

Uncertain IBD

1. **Mean Test**

   - Calculate posterior IBD sharing probabilities of the affected pair given marker data *in whole family* and marker allele frequencies. Use to get estimate of the number of alleles shared IBD.
   - Sib pair with genotypes $X = (aa, aa)$.
     Frequency of allele $a = p_a$.
     $Pr(X \mid 0 \text{ IBD}) = p_a^4$, $\Pr(X \mid 1 \text{ IBD}) = p_a^3$, etc.

     $$\Pr(0 \text{ IBD} \mid X) = \frac{0.25p_a^4}{0.25p_a^4 + 0.5p_a^3 + 0.25p_a^2}, \text{ etc.}$$

   - In absence of linkage, expected number of alleles shared IBD = 1.
   - Also need to calculate variance of affected pair's IBD sharing, conditional on observed data in rest of family. This will generally be less than 1/2
   - For more details, see Whittemore & Halpern (1994, Biometrics 50:118-127).

2. **Likelihood-ratio test**

- Denote genotypes of affected sib pair j by X_j, genotypes of rest of family by Y_j. Then the likelihood of the data is a linear function of the IBD probabilities $\underline{z}$ for the affected sib pair:

$$L(X_j, Y_j \mid \underline{z}) = \sum_{i=0}^{2} \Pr(X_j, Y_j \mid i \text{ IBD}) z_i = \sum_{i=0}^{2} w_{ij} z_i$$

- Maximise the likelihood over all pairs

$$\Pi_{j=1}^{n} (w_{0j} z_0 + w_{1j} z_1 + w_{2j} z_2)$$

  with respect to $\underline{z}$, perform likelihood-ratio test.

- E.G. $X_j = (aa, aa)$, no other relatives.
  Then $w_{0j} = p_a^4$, $w_{1j} = p_a^3$, $w_{2j} = p_a^2$.
- If both parents genotyped ($= Y_j$), then, for all i,

$$w_{ij} = \Pr(Y_j) \Pr(X_j \mid i \text{ IBD}, Y_j)$$

  This depends on allele frequencies only through the $\Pr(Y_j)$ term, which cancels when the likelihood-ratio is formed.

- Analysis is therefore robust to mis-specification of allele frequencies and population stratification when parents are genotyped.
- Same applies for counting methods (e.g. mean test)

Multiplex sibships

If 3 or more affected sibs in a sibship, IBD status of the pairs not independent of each other.
E.G. if sib pairs (A, B) and (A, C) share 2 IBD, so must (B, C).

1. **Counting methods**
   Make all possible sib pairs and treat as independent. Distribution of test statistics OK (due to *pairwise* independence of IBD status among sib pairs from the same sibship).

2. **Likelihood-ratio methods** Various options:

   - Use only one sib pair per sibship.
     - Sib pairs independent so distribution of test statistic OK.
     - Serious loss of information.
   - Pick an "index" sib and form all pairs involving this sib.
     - Same drawbacks as before.
     - May be useful if index sib is more likely to be "genetic" (e.g. early onset, more severe...)
   - Analyse all possible pairs (without downweighting)
     - If IBD is certain, distribution of test statistics OK.
     - If parents typed, distribution of test statistics reasonably OK.
     - If parents not typed, distribution depends on whether the affected sibs not involved in the pair being tested were included in the analysis (as unaffected sibs) to give information on the missing parental genotypes.
       If this is done (as is usual), simulations suggest test statistics are *conservative*, even in the absence of downweighting.

Multipoint affected sib pair analysis

- Fix a point x on chromosome ("disease locus")
- Joint likelihood of the marker data at all the loci $\underline{M} = (\underline{M}_1, \underline{M}_2, \ldots, \underline{M}_m)$:

  $$L(\underline{M} \mid j \text{ IBD at } x) = \sum_{i_1=0}^{2} \ldots \sum_{i_m=0}^{2} \Pr(\underline{M} \mid i_1, \ldots, \imath_m) \Pr(i_1, \ldots, i_m \mid j \text{ IBD at } x)$$

  where $i_1, \ldots, \imath_m$ refer to the IBD status of the sib pair at marker loci $1, \ldots, m$.

- *Assuming linkage equilibrium between the marker loci*, first term factorises into

  $$\Pi_{k=1}^{m} \Pr(\underline{M}_k \mid i_k)$$

  This assumption is made by all the commonly-used analysis packages. May be invalid if the marker loci are close together.

- Use these formulae as weights w_j in likelihood-ratio analysis or to produce posterior IBD estimates for counting analyses.

- Move x along chromosome. The location giving the highest test statistic is an estimate of the location of the disease locus (Warning: Not usually very precise !!!)

Other relative pairs

- Often available when sampling sib pairs
- May give more linkage information than sib pairs for some disease models.
- More likely to be adversely affected by incompletely informative markers - additional uncertainty about IBD vs. IBS unless intervening relatives genotyped.
- Also, power drops more quickly as recombination fraction between marker and disease locus increases (more meioses involved). Thus, a tighter marker grid would be needed than for sib pairs.
- However, if sample includes other relative pairs, these are worth including in analyses.

Non-parametric scoring statistics

(Whittemore & Halpern, 1994, Biometrics 50:109-117)

- Assume the *inheritance vector* v_i for pedigree i is known (i.e. the grandpaternal origin of each allele in the non-founders).
- $S_i = S_{pairs}(v_i) \propto$ number of pairs of alleles, one taken from each member of a pair of affected pedigree members which are shared IBD (summed over all possible pairs of affected pedigree members)
- Normalised score $Z_i(v_i) = (S_i - \mu_i)/\sigma_i$
  (μ_i, σ_i calculated by enumeration of all possible v_i)
- If IBD status incomplete, replace S_i by $\bar{S}_i$, expected value of S_i evaluated over all v_i compatible with observed data.
- $Var(\bar{S}_i) < var(S_i$ so test is conservative when IBD information incomplete.
- Test statistic for linkage $= \Sigma_i w_i \bar{Z}_i$ for some weights w_i (often chosen to be equal for all pedigrees)
- Implemented in GENEHUNTER (NPL-pairs)
- Equivalent to mean test for sibship data.
- Conservatism of NPL tests overcome by using ASM (Allele Sharing Models, Kong & Cox 1997) tests - based on likelihood models for the inheritance vectors.

Quantitative traits

- Denote trait value of i'th individual in j'th pedigree by

$$X_{ij} = \mu + g_{ij} + e_{ij}$$

  where g_{ij} depends on individual's disease genotype and e_{ij} are i.i.d $N(0, \sigma_e^2)$.

- Suppose that $g_{ij} = a, d, -a$ for genotypes DD, Dd, dd respectively.

- Then the *additive* and *dominance* variances of the trait at this locusare given by

$$\sigma_a^2 = 2pq(a - d(p - q))^2 \qquad \sigma_d^2 = 4p^2q^2d^2$$

  where p and $q(= 1 - p)$ are the frequencies of D and d respectively.

- Could use this formulation to model quantitative traits directly and perform a model-based parametric analysis.

- More usual to perform a model-free analysis. Two of the most common will be discussed here.

  1. Haseman-Elston
  2. Variance components

Haseman-Elston method

(Haseman & Elston 1972, Behav Genet 2:3-19)

- Let $Y_j = (X_{1j} - X_{2j})^2$, the squared sib-pair trait difference.
- Let $\pi_j = 0, 0.5, 1$ be the proportion of alleles shared IBD at the trait locus.
- Haseman & Elston showed that

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

  where $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$, and approximation exact if σ_d^2=0.
- Under no linkage, $\beta = 0$. Under linkage, $\beta < 0$. So, regress Y on β and perform a one-sided test.
- If IBD is uncertain, can replace π_j with its MLE, $\hat{\pi}_j$. $E(Y_j \mid \hat{\pi}_j) \approx \alpha + \beta \pi_j$, with $\beta = -2(1 - 2\theta)^2 \sigma_g^2$.
- Can also use cross-product $(X_{1j} - \mu)(X_{2j} - \mu)$ instead of Y_j (Haseman-Elston revisited, Elston et al. 2000) or weighted combination of squared difference and squared mean-corrected sum (Xu et al. 2000, Forrest 2001, Sham & Purcell 2001)
- Weighted procedures generally more powerful than old or "revisited" H-E.
- Only defined for sibship data

Variance-Component methods

(Amos 1994, Blangero & Almasy 1997)

- Express the trait value of the j'th individual in a pedigree as

$$X_j = \mu + \sum_{i=1}^{n} q_i + g_j + e_j$$

  where q_i is the (additive) effect of the i'th QTL and g_j is the residual polygenic effect (assumed to be due to an unspecified number of genes acting additively.

- The covariance matrix Ω for the pedigree can be represented

$$\Omega = \sum_{i=1}^{n} \hat{\Pi}_i \sigma_{qi}^2 + 2\Phi\sigma_g^2 + I\sigma_e^2$$

  where $\hat{\Pi}_i$ is the matrix whose elements $\hat{\pi}_{ijk}$ are the proportion of alleles shared IBD by individuals j and k at QTL i, Φ is the matrix of kinship coefficients, I is the identity matrix, σ_{qi}^2 is the additive variance due to QTL i, and σ_g^2 is the total residual polygenic inheritance.

- Assuming multivariate normality, a likelihood can be formed (here for one QTL):

$$\ln L(\underline{\mu}, \sigma_q^2, \sigma_g^2, \sigma_e^2 \mid \underline{X}) = const - 1/2 \ln \mid \Omega \mid -1/2 \Delta^T \Omega^{-1} \Delta$$

  where $\Delta = \underline{X} - \underline{\mu}$, and maximum likelihood estimation and likelihood-ratio tests carried out.

- Implemented in the package SOLAR (Blangero & Almasy 1996)

- Powerful and flexible (arbitrary pedigree structures, multiple QTLs)

- Assumes trait normality (violation can increase Type I error)

Parametric or Model-free analysis?

Depends on

1. Underlying disease model
2. Type of pedigrees available

- **Single Major Locus:**
  Use parametric analysis (Likely to have large pedigrees)
- **Complex trait - large pedigrees**
  Model free analyses: ASM or NPL (ASM better unless estimating significance by simulation)
  However, MMLS-C is an interesting alternative.
- **Complex trait - small pedigrees (sibships)**
  Sib-pair test (Mean or likelihood-ratio)
- If trait is common ($> 10\%$, say) power of all these methods is reduced. May be better to find a correlated quantitative trait and analyse that via Haseman-Elston or variance components.

Which pedigrees to sample ?

... depends on disease model !

- **Single major locus** Large multiply-affected pedigrees likely to be available - use them.
- **Complex trait** Choice between
  1. Small sample of large multiply-affected pedigrees
  2. Large sample of small pedigrees (sib pairs)
- If disease alleles are relatively common, with few phenocopies, but have relatively low penetrance, small pedigrees may be more informative than large ones. (Large pedigrees likely to be segregating more than one copy of the disease allele).
- If disease alleles are rare, but with high penetrance, much heterogeneity and a large proportion of phenocopies, a small sample of large pedigrees may be better than a large sample of small ones.
- Unfortunately, difficult to distinguish the two situations...
- ... that's one of the reasons why complex traits are complex !!

Testing for Association

- **Linkage:**
  Pairs of affected relatives from the *same* family sharing alleles IBD. The shared allele(s) may *differ* between families.
- **Association**
  Individuals from *different* families sharing the *same* allele (not necessarily IBD).
- Linkage may be present without association.
- Association may be due to
  1. Linkage (linkage disequilibrium)
  2. Artefacts (e.g. population stratification)
- Only interested in the first kind of association.

Case-Control Study

- Collect N unrelated cases, M unrelated controls *from the same population.*
- Perform standard Pearson chi-square test on the table of genotype counts

	aa	Aa	AA
Cases	n_{aa}	n_{Aa}	n_{AA}
Controls	m_{aa}	m_{Aa}	m_{AA}

- Genotypic relative risks estimated by the usual odds ratios (rare-disease assumption). e.g. $RR(Aa) = \frac{n_{Aa}m_{aa}}{n_{aa}m_{Aa}}$.
- Assuming that population genotype frequencies same for cases and controls, but no structure on these frequencies (e.g. Hardy-Weinberg equilibrium)
- If marker has many alleles, the number of genotypes may be large. The test statistic will then have many degrees of freedom, reducing power.
- Common to assume Hardy-Weinberg equilibrium in the general population, and a *multiplicative* model for genotypic relative risks ($RR(ab) = \beta_a\beta_b$).
- Test for association then equivalent to Pearson chi-square on a $2 \times k$ table of *allele counts.*
- **Advantages** of case-control studies:
  1. Easy to collect (even for complex traits with late onset)
  2. Relatively powerful
- **Disadvantage** of case-control studies:
  Not robust to population stratification.

Case-parent Trio designs: TDT

(Spielman et al. 1993, AJHG)

- Spielman et al. noted that the probability that a heterozygous parent transmits a particular allele to an affected offspring is equal to 1/2 unless *both* of the following are satisfied:

  1. There is *association* between marker and disease alleles.
  2. There is *linkage* between marker and disease loci.

- Therefore, test based on allele transmissions from heterozygous parents is a test of both *association* and *linkage* i.e. *linkage disequilibrium*.

- It will thus be robust to population stratification.

- The Transmission-Disequilibrium Test (TDT) counts the transmitted and non-transmitted alleles from each parent as a matched pair. E.G.

- These are summed over all parents to give a 2×2 table:

		$Trans$ 1	$Trans$ 2
Not $Trans$	1	a	b
	2	c	d

$$TDT = \frac{(b-c)^2}{b+c} \quad \text{(McNemar test)}$$

Likelihood-Ratio method

(Self et al. 1991, Biometrics)

- Denote the event that the child is affected by A.
- Denote the genotypes of the child, mother, father by G_C, G_M, G_F.

$$\Pr(G_C \mid G_M, G_F, A) = \frac{\Pr(G_C, A \mid G_M, G_F)}{\Pr(A \mid G_M, G_F)}$$

$$= \frac{\Pr(A \mid G_C)\Pr(G_C \mid G_M, G_F)}{\Sigma_{G^*}\Pr(A \mid G^*)\Pr(G^* \mid G_M, G_F)} = \frac{R(G_C)\Pr(G_C \mid G_M, G_F)}{\Sigma_{G^*} R(G^*)\Pr(G^* \mid G_M, G_F)}$$

where $R(G)$ is disease risk associated with genotype G.

- Fix one genotype (aa) as the reference, express disease risks of other genotypes *relative* to this genotype. $R(G) = R(aa)e^{\beta_G}$. Maximise resulting likelihood with respect to (log-) relative risk parameters $\underline{\beta}$. E.G.

Likelihood contribution =

$$\frac{0.25e^{\beta_{AA}}}{0.25e^{\beta_{AA}} + 0.5e^{\beta_{Aa}} + 0.25}$$

- Advantageous because gives estimates of genotypic relative risks. Also, extendable to model interactions, parent-of-origin effects etc.
- Multiplicative model for relative risk gives a test equivalent to the TDT.

Missing parental genotypes

1. **Discard families where transmission is uncertain**
   - Lose information
   - Can introduce bias when allele frequencies unequal (Sham & Curtis 1995 AJHG).
2. **Joint likelihood of child's, parents' genotypes**
   Suppose mother's genotype is missing.

   $$\Pr(G_C, G_F \mid A) = \sum_{G_M^*} \Pr(G_C \mid G_F, G_M^*, A) \Pr(G_F, G_M^* \mid A)$$

   No longer robust to stratification.
3. **Sib-TDT** (Spielman & Ewens AJHG 1998)
   - Needs genotyped unaffected sibs.
   - Tests whether a particular allele is more common among affected sibs than would be expected, given observed numbers of affected and unaffected sibs, together with the total number of copies that allele in the sibship.
4. **Matched case-control analysis**
   - Treat genotyped unaffected sibs as matched controls
   - Loses information on any genotyped parents (as does Sib-TDT)
   - Extendable to include covariates

Multiple affected sibs

If linkage is present, transmissions of parental genotypes to affected sibs not independent. So, tests of association invalid.

- **TDT:** Can use *bootstrapping* - generate replicate samples by sampling (with replacement) entire sibships, and recording the proportion of replicates for which a particular allele shows excess transmission.
- **Likelihood-based tests** (TRANSMIT, Clayton 1999)
  1. Empirical variance-covariance matrix for score vectors
  2. Bootstrapping: Sample (with replacement) score vectors for whole sibships (adjusted to have zero mean). Calculate test statistic. Compare to observed value.
- **SDT** (Horvath & Laird 1998)
  - Denote mean number of copies of allele 1 among affected (unaffected) sibs by m_A^1 (m_U^1).
  - Let b = no. of sibships where $m_A^1 > m_U^1$
    Let c = no. of sibships where $m_A^1 < m_U^1$
  - SDT = $\frac{(b-c)^2}{b+c}$
  - Requires typed unaffected sibs
  - Loses parental genotype information.

- **PDT** (Martin et al. 2000)
  - For each parent-offspring trio, denote X_T = (no. of allele 1 tranmitted) - (no. of allele 1 not transmitted).
    For each discordant sib pair, denote X_S = (no. of allele 1 in affected sib) - (no. of allele 1 in unaffected sib).
  - For pedigree i,
$$D_i = \frac{1}{n_T + N_S}(\sum_{j=1}^{n_T} X_{T_j} + \sum_{j=1}^{n_S} X_{S_j})$$
  - Test statistic
$$PDT = \frac{\Sigma_i D_i}{\sqrt{\Sigma_i D_i^2}} \sim N(0,1)$$

Association vs. Linkage

- **Association**
  Relatively high power to detect small effects, but only over short distances (Depends on population history and also varies over genome)
- **Linkage**
  Lower power to detect same effects, but effects extend over larger distances.
- **Genome Scan**
  Linkage: Marker every 10cM, $\approx$ 350 markers
  Association: Marker every 1cM, $\approx$ 3500 markers (or more).
- A genome scan using association requires *much* more genotyping (Pooling ?)
- **Power**
  Risch & Merikangas (Science 1996) calculated that an association scan would be more powerful than a linkage scan, even allowing for multiple testing via Bonferroni.
  - However, they assumed that one of the typed SNPs actually was the disease locus.
  - In fact, power of association methods depends on marker and disease allele frequencies, together with penetrances of disease genotypes and degree of linkage disequilibrium - hard to predict.
  - ... jury still out.