```
# To test for differences is allele frequencies across groups
# read in the table of allele (columns) by group (rows) frequencies
# and again perform the chi-square test

allele=matrix(c(126,74,150,50),2,2,byrow=T)
chisq.test(allele)

                X-squared = 6.1828, df = 1, p-value = 0.0129

chisq.test(allele,correct=F)

                X-squared = 6.7321, df = 1, p-value = 0.00947
```

The results show that cases and controls do differ significantly with respect to the allele frequencies.

The finding that cases and controls differ significantly with respect to the allele frequencies, but do not differ significantly with respect to the genotype frequencies, may seem strange. Partially, this discrepancy is attributable to the difference in degrees of freedom, that is, one versus two. The allele frequency test, in effect, assumes a multiplicative model. Thus, if the trend of the effect is consistent across genotype groups, then the allele test may prove more powerful. Potentially, cases and controls may differ mainly with respect to the number of homozygotes A_1A_1 and A_2A_2, rather than the number of heterozygotes A_1A_2.

Limitations of population and case-control association designs

As with any statistical test, association analysis is prone to 'false positives' or 'spurious association', that is, significant association without any genetic effect. Spurious association can originate from chance or from various confounding variables. Population stratification is one of the most heavily cited potential confounding factors for association. In essence, this stratification requires two conditions: different background allele frequencies and different trait means or prevalences as a function of the population membership (Ewens and Spielman, 1995; Pritchard and Rosenberg, 1999). As pointed out by Cardon and Bell (2001), Cardon and Palmer (2003), and Colhoun *et al.* (2003), other problems include small sample size, multiple testing, positive publication bias, and greatly varying methods of measuring the often complex phenotypes. These authors argue that the case-control design is statistically powerful (the power to detect weak genetic effects is usually higher than in the family-based designs supported by Risch (Risch, 2000; Risch and Teng, 1998)), very practical (cases and controls are often relatively easy to ascertain), and informative, provided that the cases and controls are meticulously matched. Given that any study design fails when executed poorly, the case-control design should not be discarded out of hand. A thorough discussion of the many known confounding effects (including population stratification) for association is presented in Chapter 21.

Solution to stratification

Population stratification can be dealt with in two fashions: through assessing the background genetic differences, or by employing an alternative study design. Statistical methods were recently developed that control for population stratification in the population-based design. For example, Devlin and Roeder (1999) developed a method called genomic control, which applies a correction to the χ^2 based on background genetic differences. The genomic control, however, may be conservative (Devlin *et al.*, 2004; Marchini *et al.*, 2004). Also, different freely available software packages now include the possibility to test and/or correct for the presence of population stratification (e.g., *Structure* (Pritchard and Rosenberg, 1999), *L-pop* (Purcell and Sham, 2004), *PLINK* (Purcell *et al.*, 2007)). The most powerful correction for stratification when the effect of the locus is consistence across strata relies on clustering within populations. The test of the genetic effect of the locus is calculated within these population clusters. The genetic effects are then equated across populations and jointly tested (e.g., Cochran Mantel-Haenszel test).

The problems stemming from hidden population stratification are minimized if a family-based design is used. Study design options include: comparing cases and controls drawn from the same family, and thus from the same genetic stratum, and the use of genetic principles to ensure an autocontrol (see TDT, below). These related controls can be selected from the same generation as (i.e., sibs), or a different generation to (i.e., parents) the cases, depending on the specific test, though admixture between the parents may mean that parents and offspring are not from the same genetic stratum (see Chapter 21).

15.4 Family-based Association Studies

For binary traits such as affection status, a variety of family-based tests for association have been developed, such as the (haplotype-based) haplotype relative risk test (HRR, Falk and Rubinstein, 1987; HHRR, Terwilliger *et al.*, 1992) and the transmission disequilibrium test (TDT, Spielman *et al.*, 1993). Recent studies (Boomsma

et al., 2000; Van den Oord, 1999) have shown that family-based tests based on continuous (quantitative) test scores are more powerful than those based on dichotomized test scores. Below, we briefly discuss two tests for the analysis of dichotomous traits (HHRR and TDT), but focus mainly on tests that make use of the variance decomposition framework to analyze continuous test scores. Retaining continuous phenotypic scores precludes the crisp distinction between cases and controls.

Dichotomous traits

The HHRR and the TDT tests are based on the autocontrol mechanism implied by the transmission of only one of the two parental chromosomes. A common sample structure for such tests is an affected offspring trio sample (both parents and one affected offspring all genotyped). The genotypic data of the affected child (i.e., the case) are compared with 'control' genotypes that are constructed from the two parental alleles that were *not* transmitted to the affected child. Both the HHRR and the TDT are based on this comparison of transmitted versus not-transmitted alleles.

Consider a diallelic locus with alleles A_1 and A_2. We can count the number of parents that transmit alleles A_1 and A_2 to the affected offspring as shown in *Table 15.3*.

The haplotype-based haplotype relative risk (HHRR) test makes use of the marginal counts, and for a diallelic marker is defined as

$$HHRR = \sum_i \frac{(t_{iT} - t_{iNT})^2}{(t_{iT} + t_{iNT})} = \frac{(t_{1T} - t_{1NT})^2}{(t_{1T} + t_{1NT})} + \frac{(t_{2T} - t_{2NT})^2}{(t_{2T} + t_{2NT})}.$$

where t_{iT} and t_{iNT} refer to the transmitted and nontransmitted alleles from parent to offspring. These ts are summed over all parent-offspring transmissions and nontransmissions for each allele i.

Table 15.3 Transmissions and nontransmissions of a diallelic marker

	Nontransmitted		
Transmitted	A_1	A_2	Total
A_1	t_{11}	t_{12}	$t_{1T} = t_{11} + t_{12}$
A_2	t_{21}	t_{22}	$t_{2T} = t_{21} + t_{22}$
Total	$t_{1NT} = t_{11} + t_{21}$	$t_{2NT} = t_{12} + t_{22}$	

The cell count t_{ij} represents the number of parents who transmitted allele i and not allele j. The marginal value t_{iT} represents the number of parents who transmitted allele i (irrespective of what allele they did not transmit), and t_{iNT} represent the number of parents who did *not* transmit allele i (irrespective of what allele they did transmit).

The resulting statistic is asymptotically equivalent to a χ^2 for the contingency table (above) where the transmitted alleles and nontransmitted alleles are counted as case and control alleles, respectively. When the genotyped marker is the QTL itself, as is assumed throughout this chapter, the HHRR test simply tests whether alleles 1 and 2 are transmitted in equal proportions to the affected offspring. A drawback of the HHRR test is that within a parent, the transmission of one allele and the *non*transmission of the other allele are not independent events. That is, the alleles contributed by each parent should actually be regarded as paired observations. The HHRR test does not account for this dependency.

The paired nature of the observations *within* parents is accommodated in the TDT test. The TDT test is based on the McNemar χ^2 test, which is a test on marginal homogeneity useful for the difference between paired proportions (e.g., transmission and nontransmission of an allele from parent to offspring). From *Table 15.3*, this marginal homogeneity implies that $t_{1T} = t_{1NT}$ and $t_{2T} = t_{2NT}$, that is, that the chances for transmission of allele 1 or 2 are equal. Writing out these equations, gives $t_{11} + t_{12} = t_{11} + t_{21}$ and $t_{22} + t_{21} = t_{22} + t_{12}$, that is, $t_{12} = t_{21}$. The McNemar test statistic is subsequently defined as $(t_{12} - t_{21})^2/(t_{12} + t_{21})$, and asymptotically χ^2 distributed with 1 df. As a variant of the McNemar test, the TDT test is defined as:

$$TDT = \sum_{i>j} \frac{(t_{ij} - t_{ji})^2}{(t_{ij} + t_{ji})} = \frac{(t_{21} - t_{12})^2}{(t_{21} + t_{12})}.$$

The TDT test thus determines whether one allele is more often transmitted to the affected offspring than the other. As the TDT test is only concerned with t_{21} and t_{12} (the proportions of parents who transmitted allele 1 but not allele 2, or *vice versa*), only heterozygous parents are informative for the TDT test, that is, parents whose transmitted and not-transmitted alleles differ. The TDT test statistic is asymptotically distributed as a X^2 with 1 df. Although the paired nature of the *within* parent observations is accommodated in the TDT test, bias is still possible. Potentially, the transmission of one allele may not be as likely as another for reasons other than effect on the disease of interest. If for example, the locus is important for viability of the fetus, then the transmission proportions may not be equal.

Example 3: The HHRR test and the TDT test

The TDT and HHRR tests are rather popular, and were recently used to study association within families with respect to, for example,

Table 15.4 Allele transmission data reported by Fan *et al.* (2003)

		Transmitted	Nontransmitted
HHRR analysis	Allele 1	280	271
	Allele 2	52	61
TDT analysis	Allele 1	58	49
	Allele 2	49	58

schizophrenia (Fan *et al.*, 2003), and mental retardation (Dutta *et al.*, 2005). In studies using the family-based design to analyze dichotomous traits, the tables including the allele transmission frequencies often resemble *Table 15.4*. These data are taken from a paper by Fan *et al.* (2003), who studied association between schizophrenia and the T1945C polymorphism in the proline dehydrogenase gene in a Chinese sample.

The frequencies reported for the HHRR analyses should be interpreted as the marginal frequencies t_{1T}, t_{1NT}, t_{2T} and t_{2NT}. The HHRR statistic is thus calculated as:

$$HHRR = \frac{(280-271)^2}{(280+271)} + \frac{(52-61)^2}{(52+61)} = 0.86.$$

The TDT statistic is based on the heterozygote parents only, and is calculated as:

$$TDT = \frac{(58-49)^2}{(58+49)} = 0.76.$$

Considering that both statistics follow a χ^2 distribution with df = 1, the accompanying *p*-values equal 0.35 and 0.38, respectively. The authors therefore concluded that the data did not suggest the presence of association between schizophrenia and the polymorphism under study.

A practical drawback of designs that use parental data is that for certain diseases, parents may not be available (which is particularly relevant for the investigation of late-onset diseases). Also, as only heterozygote parents are informative for the TDT test, the effective sample size is often a lot smaller that the total sample size. More recent extensions of these tests can handle the conditional dependence of multiple sibs. We refer to Zhao (2000) for a discussion of such recent extensions of the TDT test, and to Sham (1998) for a discussion of the HHRR and the TDT for (multiallelic) markers, and for a likelihood-ratio based version of the TDT test.

Use of sibling data

One of the great advantages of sibling analysis is the control for population stratification, analogous to the use of parents in the TDT and HHRR tests. Siblings are usually about the same age, so the confounding effects of age are limited. Additionally, the analysis of sibling data is more practical in the study of late-onset conditions when parental data are not available. From a qualitative phenotype perspective, siblings discordant for affection status provide information about association. Essentially, the frequency of a risk allele ought to be higher in affected than unaffected offspring. The pedigree-disequilibrium test (PDT) incorporates this discordant sibling information as well as the TDT information described above (Martin *et al.*, 2000). The PDT defines a statistic *D* for the *i*th family such that

$$D_i = \frac{1}{n_T + n_S}\left(\sum_{j=1}^{n_T} X_{Tj} + \sum_{j=1}^{n_S} X_{Sj}\right)$$

where n_T and n_S are the number of transmissions from a heterozygote parent to an affected offspring, and the number of discordant sibling pairs respectively. X_{Tj} is the number of transmissions of a given allele minus the number of nontransmissions of the same allele, and X_{Sj} is the number of a given allele in the affected offspring minus the number of alleles in the unaffected offspring. The expectation (mean) of *D* under no association is 0. The significance of *D* is calculated by defining a test statistic T, which asymptotically approximates a *Z* score:

$$T = \frac{\sum_{i=1}^{N} D_i}{\sqrt{Var\left(\sum_{i=1}^{N} D_i\right)}}$$

and

$$Var\left(\sum_{i=1}^{N} D_i\right) = \sum_{i=1}^{N} D_i^2$$

PDT is a flexible framework for incorporating information on all family members and is freely available at http://wwwchg.mc.duke.edu.

Quantitative traits

Some recent examples of the use of sibling data in association studies for quantitative phenotypes are *inter alia* studies of the association between intelligence and the *CHRM2* gene (Gosso *et al.*, 2006), the association between depression and the serotonin system gene *TPH1* (Nash *et al.*, 2005), and the association between alcoholism and the ADH2 genotype (Neale *et al.*, 1999).

Basic extensions of the TDT test to include continuous or quantitative measures were developed (Allison, 1997; Rabinowitz, 1997a). Fulker *et al.* (1999) proposed a test of association for quantitative traits that is based on data collected in pairs of siblings. The approach was further extended by Abecasis *et al.* (2000a) to include parental data and multiple siblings. This quantitative test makes use of maximum-likelihood-based estimation and testing. In its full extent, the approach allows for the simultaneous modeling of variances and means, that is, the simultaneous modeling of linkage and association. Here, we focus on the model for the means only. We refer to Chapter 10 for a detailed discussion of variance components linkage analysis, and to Fulker *et al.* (1999) and Abecasis *et al.* (2000a) for discussions of the simultaneous modeling of linkage and association. Below, the rationale of sib-pair analysis is discussed, and the extension to larger pedigrees is reviewed briefly.

As shown by, for example, Fulker and Cherny (1996), the variance-covariance structure between two sibs, and their respective mean trait scores, can be modeled by maximizing the following likelihood function:

$$L = \prod_{i=1}^{M} (2\pi)^{-k_i/2} \left|\Sigma_i\right|^{-1/2} e^{-1/2[(\mathbf{y}_i - \boldsymbol{\mu}_i)'\Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)]}$$

with respect to the expected sib-pair variance-covariance matrix $\boldsymbol{\Sigma}_i$ and the vector with expected means $\boldsymbol{\mu}_i$ for family i. In the equation above, k is the number of variables observed within each family i, $\mathbf{y}_i$ is the vector of observed scores in family i, and M is the number of families. In this likelihood expression, means and variances can be made functions of parameters based on a theoretical model.

Fulker *et al.* (1999) extended this sibling model by assuming the biometrical model for diallelic markers, and considering the nine possible combinations of sib-pair genotypes. The vector of sib-pair scores can be modeled in terms of pair means and pair differences. Specifically, the vector of expected means $\boldsymbol{\mu}_i$ can be modeled as a function of an overall mean m or intercept, the family i-specific sib-pair mean m_i, (i.e., the *between* effect), and the family-specific sib-pair difference δ_i (i.e., the *within* effect). So, for a pair of siblings, $\mu_{1i} = m + m_i + \delta_i/2$, and $\mu_{2i} = m + m_i - \delta_i/2$. These family-specific means and differences are a function of the genotypes of the siblings, and thus of the genotypic value a. See *Table 15.5* for all possible combinations of a diallelic locus for a sib pair, with information on the expected scores for the sib-pair mean and difference.

Table 15.5 Partitioning of additive effects into between and within pair components for family *i*

Genotype		E(Trait value)		Sib mean	Sib difference/2	Additive effect (B and W)	
Sib 1	Sib 2	g_{i1}	g_{i2}	$g_{i\mu}$	g_{id}	Sib 1 ($g_{i\mu} + g_{id}$)	Sib 2 ($g_{i\mu} - g_{id}$)
A1A1	A1A1	a	a	a_b	0	a_b	a_b
A1A1	A1A2	a	0	$\frac{1}{2}a_b$	$\frac{1}{2}a_w$	$\frac{1}{2}a_b + \frac{1}{2}a_w$	$\frac{1}{2}a_b - \frac{1}{2}a_w$
A1A1	A2A2	a	−a	0	a_w	a_w	$-a_w$
A1A2	A1A1	0	a	$\frac{1}{2}a_b$	$-\frac{1}{2}a_w$	$\frac{1}{2}a_b - \frac{1}{2}a_w$	$\frac{1}{2}a_b + \frac{1}{2}a_w$
A1A2	A1A2	0	0	0	0	0	0
A1A2	A2A2	0	−a	$-\frac{1}{2}a_b$	$\frac{1}{2}a_w$	$-\frac{1}{2}a_b + \frac{1}{2}a_w$	$-\frac{1}{2}a_b - \frac{1}{2}a_w$
A2A2	A1A1	−a	a	0	$-a_w$	$-a_w$	a_w
A2A2	A1A2	−a	0	$-\frac{1}{2}a_b$	$-\frac{1}{2}a_w$	$-\frac{1}{2}a_b - \frac{1}{2}a_w$	$-\frac{1}{2}a_b + \frac{1}{2}a_w$
A2A2	A2A2	−a	−a	$-a_b$	0	$-a_b$	$-a_b$

The first four columns of *Table 15.5* simply contain the possible combinations of genotypes, and the expected trait values for each genotype, in the absence of dominance. The fifth column contains the pair means, calculated as the sum of the genotypic values g_{ij} of the two siblings j in family i divided by two. The sixth column contains the difference between the genotypic values g_{ij} of the two siblings j in family i divided by two.

The additive effect (B and W) in columns 7 and 8 of *Table 15.5* is partitioned into a component attributable to the between effect, a_b (i.e., the pair mean), and a component attributable to the within effect, a_w (i.e., the subject's deviation from the pair mean). From the parameters in columns 7 and 8, it is immediately apparent which combinations of siblings are most informative for the estimation of a_b and a_w, respectively. For example, pairs in which both sibs have the same genotype are not informative for the estimation of a_w, while sib pairs with one homozygote A_1A_1 and one homozygote A_2A_2, and pairs in which both sibs are heterozygotes, do not provide information for the estimation of a_b. See *Figure 15.2* for a graphical representation of the source of the *between* and *within* information from a pair of siblings.

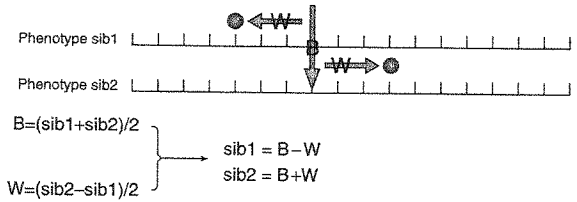

Figure 15.2 Graphical representation of the source of the *between* and *within* information from a pairs of siblings.

A range of association tests can be conducted from the sib-pair *between* and *within* information. Under this partitioned *between* and *within* model there are three possible tests of association.

- Test for association while controlling for population stratification – testing the significance of the *within* component only. Note, in this model the *between* component of means may or may not be estimated. The implications therein are discussed below.
- The combined test – a combined test of the significance of both the *between* and *within* components, following a test of equivalence of the *between* and *within* components. In the absence of population stratification, this test provides the maximum power to detect association. The between and within effects are equated for testing, so it is a 1 df test.
- Test for association without controlling for population stratification – testing the significance of the *between* component.

The combined test is similarly prone to stratification error as the *between* test because the *between* component is included in the combined test. However, a direct test for stratification at the locus can be conducted by testing the equivalence of the *between* and *within* components. If there is no effect of stratification, the between and within families effects should be equal.

Estimating the *between* component as a free parameter when testing for the *within* component effectively mean-centers each pair of siblings, such that the deviations are robust to stratification. However, the information about how deviant the sib pair is, in terms of the pairwise score, is lost. Under no stratification, doing so may reduce power to detect association. Thus, modeling only the *within* component without correcting for *between* pair differences is more powerful when no stratification is present.

The sibling case can be extended to general pedigrees, by altering the model for the phenotypic score y of subject j in family i, y_{ij}, as follows:

$$y_{ij} = m + A_{bi} + A_{wij},$$
$$= \mu + \beta_{bi} a_b + \beta_{wij} a_w$$

where m is the grand mean taken across families, A_{bi} is the between effect for family i, and A_{wij}, is the within effect for subject j in family i. A_{bi} and A_{wij} can be rewritten as functions of the parameters a_b and a_w, where β_{bi} and β_{wij} are the coefficients with which the between and within family effects a_b and a_w need to be multiplied for subject j from the ith family. For instance, given a sib pair with genotypes A_1A_1 and A_1A_2, the coefficient β_{bi} equals +½ for both sibs, and the coefficient β_{wij} equals +½ for sibling 1, and –½ for sibling 2. Or, given a sib pair with genotypes A_1A_1 and A_2A_2, the coefficient β_{bi} equals 0 for both sibs, and the coefficient β_{wij} equals +1 for sibling 1, and –1 for sibling 2, and so on.

The exponential part of the likelihood function can thus be rewritten as:

$$-\tfrac{1}{2}[(\mathbf{y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)] =$$
$$-\tfrac{1}{2}[(\mathbf{y}_i - \mathbf{m} - \mathbf{B}_{bi}\boldsymbol{a}_b - \mathbf{B}_{wi}\boldsymbol{a}_w)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \mathbf{m} - \mathbf{B}_{bi}\boldsymbol{a}_b - \mathbf{B}_{wi}\boldsymbol{a}_w)],$$

where $\mathbf{m}$ is a vector containing the grand mean for each member of the sibship (set equal across all siblings and families), $\mathbf{B}_{bi}$ is a diagonal matrix containing the coefficients with which the parameters in the vector $\boldsymbol{a}_b$ (containing the parameters a_b for each sib) need to be multiplied for family i, and $\mathbf{B}_{wi}$ is a diagonal matrix containing the coefficients with which the parameters in the vector $\boldsymbol{a}_w$ (containing the parameters $\boldsymbol{a}_w$ for each sib) need to be multiplied for family i. The three tests of association described above using the between within model with sib pairs can be applied to the more general pedigree case.

The extension of the model to include dominance is straightforward in Mx (see Posthuma *et al.*, 2004), and is summarized in *Table 15.6*. Again, columns 1 to 4 contain the possible combinations of genotypes, and the expected trait values as obtained from the biometrical model including dominance for the genotype configuration. The fifth and sixth columns contain the sib-pair means $g_{ia\mu}$ and differences $g_{ia\delta}$ regarding the additive effect only. Analogously, columns 7 and 8 contain the sib-pair means and differences with regard to the dominance effect, $g_{id\mu}$ and $g_{id\delta}$. Columns 9 and 10 contain

Table 15.6 Partitioning of additive and dominance effects into between and within pair components for family i

Genotype		E(Trait value)		Additive		Dominance		Partitioned genotypic effects	
Sib 1	Sib 2	Sib 1	Sib 2	$g_{iaμ}$	$g_{iaδ}$	$g_{idμ}$	$g_{idδ}$	Sib 1 ($g_{iaμ} + g_{idμ} + g_{iaδ} + g_{idδ}$)	Sib 2($g_{iaμ} + g_{idμ} - g_{iaδ} - g_{idδ}$)
A1A1	A1A1	a	a	a_b	0	0	0	a_b	a_b
A1A1	A1A2	a	d	½ a_b	½ a_w	½ d_b	−½ d_w	(½ a_b + ½ a_w) + (½ d_b − ½ d_w)	(½ a_b − ½ a_w) + (½ d_b+ ½ d_w)
A1A1	A2A2	a	−a	0	a_w	0	0	a_w	$-a_w$
A1A2	A1A1	d	a	½ a_b	−½ a_w	½ d_b	½ d_w	(½ a_b − ½ a_w) + (½ d_b + ½ d_w)	(½ a_b + ½ a_w) + (½ d_b− ½ d_w)
A1A2	A1A2	d	d	0	0	d_b	0	d_b	d_b
A1A2	A2A2	d	−a	−½ a_b	½ a_w	½ d_b	½ d_w	(−½ a_b + ½ a_w) + (½ d_b + ½ d_w)	(−½ a_b− ½ a_w) + (½ d_b− ½ d_w)
A2A2	A1A1	−a	a	0	−a_w	0	0	$-a_w$	a_w
A2A2	A1A2	−a	d	−½ a_b	−½ a_w	½ d_b	−½ d_w	(−½ a_b− ½ a_w) + (½ d_b − ½ d_w)	(−½ a_b+ ½ a_w) + (½ d_b+ ½ d_w)
A2A2	A2A2	−a	−a	-a_b	0	0	0	-a_b	-a_b

the genotypic effects for each sib, partitioned in additive between and within effects, and dominance between and within effects.

Abecasis *et al.* (2000a) generalized the sib-pair model to include multiple siblings by redefining the between effect A_{bi} as

$$A_{bi} = \frac{\sum_{j=1}^{n} g_{ij}}{n_i}, \tag{15.2}$$

that is, the sum of the genotypic values of all siblings of family i, divided by the number of siblings in family i. When parental genotypes are available, Abecasis *et al.* (2000a) suggested defining the between effect A_{bi} completely in terms of the genotypic values of the parents:

$$A_{bi} = \frac{\sum_{j} g_{iF} + g_{iM}}{2}, \tag{15.3}$$

that is, the sum of the genotypic values of the mother and father of family i, divided by 2. In both cases, the within effect A_{wij} is defined as $g_{ij} - A_{bi}$, that is, the genotypic value of subject j in family i minus the between effect of family i. Formulated thus, A_{bi} is the expectation of the genotypic value of each subject j in family i conditional on the data from family i, and A_{wij} is the deviation from this expectation for offspring j.

Following the argument that larger pedigrees include more information on population structure than nuclear families, Abecasis *et al.* (2000b) extended the above scoring methods to include other relatives besides siblings and parents. A general algorithm for defining A_{bij} and A_{wij} was proposed that considers all available information. Note that for these larger pedigrees, the value A_{bij} has a subscript for family (i) as well as individual (j), because the between effect can now vary between subjects coming from the same family, depending on their status (founder, sibling, etc.) in the pedigree. The algorithm includes the following scoring rules:

(i) If an individual j from family i is a genotyped founder, $A_{bij} = g_{ij}$, and $A_{wij} = g_{ij} - A_{bij} = 0$. Defined as such, the parental genotypic information contributes to the estimation of the between family effect a_b, that is, the effect that includes genuine and possibly spurious association, depending on the presence of population stratification.

(ii) If founders are genotyped in family i, that is, the values A_{biF} (= g_{iF}) and A_{biM} (= g_{iM}) are available, define A_{bij} for the offspring as in Equation 15.2.

(iii) If no founders are typed but genotypes of full siblings are available, define A_{bij} on the basis of all genotyped siblings as in Equation 15.2.
(iv) If person j is not genotyped, and A_{biF} and A_{biM} are unavailable for this person, A_{bij} and A_{wij} remain undefined.

Recently, Havill *et al.* (2005) suggested an alternative parameterization for founders in the specific case that population stratification is absent, which would results in a more powerful test for association (or LD in the case that one uses SNP rather than candidate genes). Visscher and Duffy (2006) showed that, in terms of power, it is even advantageous to include ungenotyped relatives with measured phenotypes in the analyses.

The discussion of the above methods focuses on the change in the expected phenotypic score as a function of the genotype (i.e., the focus is on the means). The parameters of interest in the means model, are the genotypic value, a, and the dominance deviation, d, as defined in the biometrical model. In linkage analysis, where the aim is to establish whether people who resemble each other phenotypically are more likely to be IBD at the locus of interest, the focus is on the covariance matrix. The Fulker model described above allows for the joint estimation of linkage and association information. For example, association may be detected in the absence of linkage due to the differential power of the two approaches (i.e., association is considerably more powerful than linkage). Observing significant linkage in the absence of association is largely immaterial, given the poor resolution of linkage.

Below we present two examples of data analysis in QTDT (http://www.sph.umich.edu/csg/abecasis/QTDT/). The tests for association, linkage, and combined linkage and association as available in QTDT, are described by Abecasis *et al.* (2000a) and others (see online documentation), and include extensions of the above-described between/within design to more general pedigrees. QTDT can analyze discrete as well as continuous traits, covariates, and multiple alleles, but as yet only univariate phenotypic data. That is, one cannot study linkage and/or association in several phenotypes simultaneously. In principle, the Mx program (Neale *et al.*, 2003) can be used to study association in a multivariate context. We refer to Posthuma *et al.* (2004) for an example of the combined study of linkage and association in Mx in a univariate design. The scripts presented in that paper can relatively easily be extended to a multivariate context.

The files used in Examples 4 and 5 below can be found on the webpage under the names *Chap15ex5.ped, Chap15ex5.dat, Chap15ex6.ped, Chap15ex6.dat.* (Note that you may need to put those in the folder in which you installed QTDT.)

Example 4: QTDT analysis of quantitative traits measured in parent-offspring trios

The quantitative TDT test in the program QTDT uses two input files (in this example, *Chap15ex5.ped*, and *Chap15ex5.dat*), the format of which is described in Appendix 1. Briefly, the pedigree (*.ped*) file includes the family structure, genotypes and phenotypes of every family member in the dataset. The data (*.dat*) file describes the structure of the pedigree file, in terms of the markers and phenotypes. Before embarking on any genetic analysis, the use of Pedstats to check data quality is highly recommended (see Appendix 1). This dataset is comprised of 200 families, each of which has three individuals: both parents and an offspring.

Pedstats gives an overview of the descriptives of the traits and the covariates, and an overview of the available genotypes of the markers in the file. If data of multiple siblings are present, sib-correlations are calculated.

```
QUANTITATIVE TRAIT STATISTICS

          [All Phenotypes] Min    Max    Mean    Var    SibCorr
Trait_1   600 100.0%      -3.245 3.583  -0.042  1.014      -
Total     600 100.0%

          [Founders Only]  Min    Max    Mean    Var    SibCorr
Trait_1   400 100.0%      -3.245 3.583  -0.043  1.084      -
Total     400 100.0%

COVARIATE STATISTICS

          [All Phenotypes] Min    Max    Mean    Var    SibCorr
Covariate 600 100.0%      -1.923 1.509  -0.001  0.260      -
Total     600 100.0%

          [Founders Only]  Min    Max    Mean    Var    SibCorr
Covariate 400 100.0%      -1.923 1.358  0.014   0.269      -
Total     400 100.0%

MARKER GENOTYPE STATISTICS

          [Genotypes]      [Founders]      Hetero
SNP_1     600 100.0%       400 100.0%      50.0%
SNP_2     600 100.0%       400 100.0%      52.7%
SNP_3     600 100.0%       400 100.0%      45.3%
Total     1800 100.0%      1200 100.0%     49.3%
```

To analyze these data, type the command:

```
qtdt -d Chap15ex5.dat -p Chap15ex5.ped
```

By default, QTDT uses the quantitative TDT test described by Abecasis *et al.* (2000a), but one can request the tests described by Allison (1997) and Rabinowitz (1997) by including *–aa* or *–ar*, respectively, at the end of the statement.

In the output (below), the requested model is conveyed. In this particular case, the fit of the Null model including *Mu* (the grand mean), and *B* (the between-family effect) is compared to the Full model, which in addition includes *W* (the within-family effect). Subsequently, the results of the tests for association are shown for each trait against each allele of each marker, with *p*-values shown only if smaller than 0.10. The results suggest an association between trait 1 and markers 1 and 2, but not 3. Such correlated results are common when considering variation in a region because of linkage disequilibrium (see Chapter 19).

```
The following models will be evaluated...

NULL MODEL
Means = Mu + Covariate + B

FULL MODEL
Means = Mu + Covariate + B + W

Testing trait:                          Trait_1

Testing marker:                         SNP_1

Allele   df(0)  Rsq(0)  df(T)  Rsq(T)     F       p
1          597    0.26    596    0.27    6.94  0.0086    (151/600 probands)
2          597    0.26    596    0.27    6.94  0.0086    (151/600 probands)

Testing marker:                         SNP_2

Allele   df(0)  Rsq(0)  df(T)  Rsq(T)     F       p
1          597    0.32    596    0.34   22.14  3e-006    (157/600 probands)
2          597    0.32    596    0.34   22.14  3e-006    (157/600 probands)

Testing marker:                         SNP_3

Allele   df(0)  Rsq(0)  df(T)  Rsq(T)     F       p
1          597    0.26    596    0.26    0.05            (129/600 probands)
2          597    0.26    596    0.26    0.05            (129/600 probands)
```


Note that the covariate(s) declared in the *.dat* file are included in the analyses by default. The covariate can be excluded by:

```
qtdt -d Chap15ex5.dat -p Chap15ex5.ped -c-
```

If you do this, you'll find that only the association between the trait and the second marker is significant.

In this example, the covariate obscures the relation between the phenotype and the markers. When the variance due to the covariate is accounted for in the model, one will get a clearer view of the actual relation between the phenotype under study and the markers: inclusion of a covariate in the analyses can improve the marker signal. The data for the present example were simulated such that the covariate has the same effect for parents and offspring, but the covariate could also affect parents and offspring differently (e.g., age effects), and as such affect the association effects differently in different generations. However, the use of covariates does not always improve association evidence. For example, consider including a covariate that shares the effect of the locus on the trait of interest. In that case, addition of the covariate can remove the association information at the locus.

Example 6: QTDT analysis of quantitative traits measured in sib pairs

Consider the file *Chap15ex6.ped*, which contains data of 300 sibling pairs, but no parental genotypic or phenotypic information; however, the parents need to be included as founders to establish the genetic relationship between the siblings. The *Chap15ex6.dat* file shows that, apart from the five default variables, the file contains information on three markers, one trait and one covariate. Running Pedstats generates the sibling correlation for the trait and covariate, which are 0.324 and 0.122, respectively.

The following command specifies a model for association and linkage:

```
qtdt -d Chap15ex6.dat -p Chap15ex6.ped -weg -af
```

The *–w* option specifies the variance decomposition under the Null model, with the '*e*' and '*g*' denoting that unique environmental (e) and polygenic (g) effects are modeled. The *–af* option means that the alternative, or Full, model is modeled according to the Fulker model. Under the Fulker model, the variance decomposition is similar to under the Null model, but while the means decomposition includes *Mu* (the grand mean), *Covariate* (effect of the covariate on the mean) and *B* (the between-family effect) under the Null

model, it also includes *W* (the within-family effect) under the Fulker (Full) model. The above statement thus tests for *robust association*, that is, the significance of the within-family component W, which is free from population stratification. In QTDT, the parameterization of the means and variances for the null model are:

- Means = Mu + Covariate + B
- Variances = Ve + Vg

and for the full model:

- Means = Mu + Covariate + B + W
- Variances = Ve + Vg

As can be seen in the remaining output, the within-family component is significant for marker 2 and 3, that is, within families there is evidence for association between these markers and the trait under study.

Testing trait: **Trait_1**

Testing marker: SNP_1

Allele	df(0)	-LnLk(0)	df(T)	-LnLk(T)	ChiSq	p	
1	595	823.83	594	823.79	0.07		(208/600 probands)
2	595	823.83	594	823.79	0.07		(208/600 probands)

Testing marker: SNP_2

Allele	df(0)	-LnLk(0)	df(T)	-LnLk(T)	ChiSq	p	
1	595	797.59	594	779.81	35.57	2e-009	(228/600 probands)
2	595	797.59	594	779.81	35.57	2e-009	(228/600 probands)

Testing marker: SNP_3

Allele	df(0)	-LnLk(0)	df(T)	-LnLk(T)	ChiSq	p	
1	595	821.84	594	812.10	19.49	1e-005	(222/600 probands)
2	595	821.84	594	812.10	19.49	1e-005	(222/600 probands)

Rather than only testing for association within families, the more powerful *combined* test for association, that is the equated *between* and *within* components, can be applied. However, the test for total association is only reliable if population stratification is not present. Thus, the first step is to test for population stratification.


```
qtdt -d Chap15ex6.dat -p Chap15ex6.ped -weg -ap.
```

The means and variances for the null model are:

- Means = Mu + Covariate + X
- Variances = Ve + Vg

and for the full model:

- Means = Mu + Covariate + X + W
- Variances = Ve + Vg

where **X** stands for 'B = W', that is, the effect that is explained by both the within- and the between-family components, and **W** is the effect that is *not* explained by X. That is, if W is significant, population stratification is present (i.e., B ↑ W). In the output we see that W is only marginally significant for marker 3 ($p = 0.0702$).

Testing trait: **Trait_1**

Testing marker: SNP_1

Allele	df(0)	-LnLk(0)	df(S)	-LnLk(S)	ChiSq	p	
1	595	823.97	594	823.79	0.36		(208/600 probands)
2	595	823.97	594	823.79	0.36		(208/600 probands)

Testing marker: SNP_2

Allele	df(0)	-LnLk(0)	df(S)	-LnLk(S)	ChiSq	p	
1	595	779.90	594	779.81	0.18		(228/600 probands)
2	595	779.90	594	779.81	0.18		(228/600 probands)

Testing marker: SNP_3

Allele	df(0)	-LnLk(0)	df(S)	-LnLk(S)	ChiSq	p	
1	595	813.74	594	812.10	3.28	0.0702	(222/600 probands)
2	595	813.74	594	812.10	3.28	0.0702	(222/600 probands)

Assuming that we were testing using a significance level of $\alpha = 0.01$, we conclude that there is no convincing evidence for the presence of population stratification, that is, we assume that B = W. Now we can use the more powerful *combined test of association*.

The QTDT command:

```
qtdt -d Chap15ex6.dat -p Chap15ex6.ped -weg -at
```

models the means and variances for the null model as:

- Means = Mu + Covariate
- Variances = Ve + Vg

and for the full model as:

- Means = Mu + Covariate + X
- Variances = Ve + Vg

where **X** again stands for 'B = W'.

This model estimates a single parameter for the *between* and the *within* effect on the means and tests whether it is significantly different from zero, which is a 1 df test. The output for this test shows greater evidence for association between the trait and markers 2 and 3:

```
Testing trait:                    Trait_1

Testing marker:                   SNP_1

Allele  df(0)  -LnLk(0)  df(X)  -LnLk(X)  ChiSq     p
  1      596    824.82     595    823.97    1.69          (600 probands)
  2      596    824.82     595    823.97    1.69          (600 probands)

Testing marker:                   SNP_2

Allele  df(0)  -LnLk(0)  df(X)  -LnLk(X)  ChiSq     p
  1      596    824.82     595    779.90   89.85   3e-021 (600 probands)
  2      596    824.82     595    779.90   89.85   3e-021 (600 probands)

Testing marker:                   SNP_3

Allele  df(0)  -LnLk(0)  df(X)  -LnLk(X)  ChiSq     p
  1      596    824.82     595    813.74   22.17   3e-006 (600 probands)
  2      596    824.82     595    813.74   22.17   3e-006 (600 probands)
```

All previous examples of association analysis in family-based designs were worked out for QTDT. Other programs are however available, like FBAT (Horvath and Laird, 1998; Horvath *et al.*, 2001; Laird *et al.*, 2000; Lake *et al.*, 2000). FBAT allows for testing of association with any phenotype (univariate and multivariate), in the context of different sampling structures, and missing marker-allele information (Laird *et al.*, 2000; Rabinowitz, 1997). Also worth noticing is the relatively new software tool PBAT, which can be used for the planning of family-based association studies, and which provides power calculations for a wide variety of designs (Lange *et al.*, 2004). FBAT and PBAT are freely available at the website http://www.biostat.harvard.edu/~fbat/default.html.

15.5 Conclusion

The present chapter dealt with designs and methods for the study of single-locus association analysis. Broadly, study designs for association analysis fall into one of two categories: unrelated and family-based samples. Within unrelated-individuals designs, two popular sampling techniques are random ascertainment from the population (useful for quantitative traits) and case-controls ascertainment (useful for disease association). Analysis of these study designs is easy to implement, relatively cheap, and practicable. However the effects of confounding factors such as population stratification are important considerations for analysis. Family-based designs are robust to such confounding effects, but can be less practicable, for example, when late-onset diseases are studied in designs that require the collection of parental data.

Acknowledgments

We thank Conor V. Dolan for his help with the R scripts. Preparation of this manuscript was financially supported by NWO/MaGW VIDI-016-065-318.

References

Abecasis, G.R., Cardon, L.R. and Cookson, W.O.C. (2000a) A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**: 279–292.

Abecasis, G.R., Cookson, W.O.C. and Cardon, L.R. (2000b) Pedigree tests of transmission disequilibrium. *Eur. J. Hum. Genet.* **8**: 545–551.

Allison, D.B. (1997) Transmission-disequilibrium test for quantitative traits. *Am. J. Hum. Genet.* **60**: 676–690.

Bain, L., Yang, J.D., Guo, T.W., Duan, Y., Qin, W., Sun, Y., Feng, G.Y. and He, L. (2005) Association study of the A2M and LRP1 genes with Alzheimer disease in the Han Chinese. *Biol. Psychiat.* **58**: 731–737.

Boomsma, D.I., Beem, A.L., van den Berg, M., Dolan, C.V., Koopmans, J.R., Vink, J.M., De Geus, E.J.C. and Slagboom, P.E. (2000) Netherlands twin family study of anxious depression (NEDSAD). *Twin Res.* **3**: 323–334.

Cardon, L.R. and Bell, J.I. (2001) Association study designs for complex diseases. *Nat. Rev. Genet.* **2**: 91–99.

Cardon, L.R. and Palmer, L.J. (2003) Population stratification and spurious allelic association. *Lancet* **361**: 598–604.