

Power of Genetic Association Studies in the Presence of Linkage Disequilibrium and Allelic Heterogeneity

Sheila A. Fisher^a Cathryn M. Lewis^{a, b}

^aDivision of Genetics and Molecular Medicine and ^bSGDP Centre, Institute of Psychiatry, King's College London, London, UK

Key Words

Allelic association · Case-control study · Heterogeneity · Linkage disequilibrium · Mutations · Sample size · Transmission disequilibrium test · Allelic heterogeneity

Abstract

Objectives: The calculation of the power and sample size required for association studies is essential, particularly for follow-up of genome-wide association studies, where much genotyping is required to replicate the original finding and identify the true disease susceptibility mutation. **Methods:** In this paper, we derive equations for estimation of sample sizes for the transmission disequilibrium test (TDT) and for case-control studies, in the presence of allelic heterogeneity and indirect association – where the genotyped tagging SNP is in linkage disequilibrium (LD) with the true mutation. Using data from *NOD2* and *PTPN22*, we show that the true sample sizes required to detect association may be incorrect when calculated under the assumption of a single mutation and complete LD with the genotyped marker. **Results:** The true sample sizes may be lower when allelic heterogeneity acts in a recessive model across mutations, or increased when mutations lie on different alleles of a common tagging SNP. **Conclusion:** Calculating power and sample size under a range of realistic models of LD and allelic heterogeneity is essential to ensure that association studies have sufficient power to detect mutations. Copyright © 2008 S. Karger AG, Basel

Introduction

Accurate estimation of sample sizes required in a genetic association study is essential before commencing genotyping, to ensure that the study is sufficiently powered to detect the subtle genetic effects that contribute to most complex diseases. Several methods have been developed to assess the power of genome-wide association studies using contrasting study designs [1–6]. However, most power calculations assume the simplest genetic model of a single diallelic disease mutation which is genotyped to test for association. Complex disease studies have revealed this to be a simplistic assumption. The extensive genetic variation and complex linkage disequilibrium (LD) across even a small genomic region gives rise to several alternative scenarios.

The simplest scenario is when a single functional variant is tested directly for association. (We will use the term disease susceptibility mutation (DSM) for such a variant, regardless of its frequency.) However, an alternative marker in LD with the true mutation may be tested (rather than the disease mutation itself). This may arise because the susceptibility mutation is not detected by initial polymorphism screening, or is not tested due to genotyping constraints. In this case, the genotyped marker will have reduced power to detect association, depending on the extent of LD between the tested marker and the true mutation. The situation is further com-

plicated by several disease susceptibility mutations existing within the same genetic region – this is common in Mendelian diseases, and examples of this also occur in complex diseases including *NOD2* (*CARD15*), a susceptibility gene for Crohn’s disease [7–9], *PTPN22*, a susceptibility gene for rheumatoid arthritis [10], and *DISC1* mutations in bipolar spectrum disorders and psychotic phenotypes [11]. For detecting association in a complex disease, if one DSM is tested, other mutations may lie on a haplotype with the observed wildtype allele and dilute the effect of the tested marker. Alternatively, the tested marker may not contribute directly to disease susceptibility but is in varying degrees of LD with each of the DSMs.

In this paper we derive the equations required to estimate power and sample sizes required for association studies based on case-control and transmission disequilibrium test (TDT) designs, exploring different scenarios of one or two DSMs in the gene, and the genotyped marker being a DSM or a marker in LD with the DSMs. This work follows on from the methods of Camp [2] without the assumption of independence between parental genotypes in TDT trios. Sample sizes required for sufficient power to detect association are calculated for the TDT and compared with those required for a case-control study. For allelic heterogeneity, calculating LD between two DSMs and the tested marker loci becomes algebraically intractable. We therefore simplify the model, assuming that the two DSMs do not occur on the same haplotype and therefore only three haplotypes exist. This arises in *NOD2* and *PTPN22*, and data from these genes are used to demonstrate the effect of multiple mutations and LD on the required sample sizes for TDT and case-control association studies.

Methods

Here, we define four different scenarios, which incorporate direct testing of a single disease locus as well as the presence of multiple disease mutations and testing of a marker in LD with the DSM(s) (fig. 1).

For a disease with prevalence K , consider a diallelic locus D which is directly associated with disease where the disease susceptibility allele D and wild-type allele d occur with frequency p_D and p_d respectively. Mutation-specific disease penetrances are defined by f_{DD} , f_D and f_d as the probability of disease for genotypes DD , Dd and dd [2]. Then the genotype relative risk (GRR) is defined as GRR_i , the relative risk to an individual carrying i copies of the D allele compared with an individual carrying none, so that $GRR_1 = f_D/f_d$ and $GRR_2 = f_{DD}/f_d$. Three common genetic inheritance models (multiplicative, dominant and recessive) are assumed. The additive linear model is not discussed here, since for

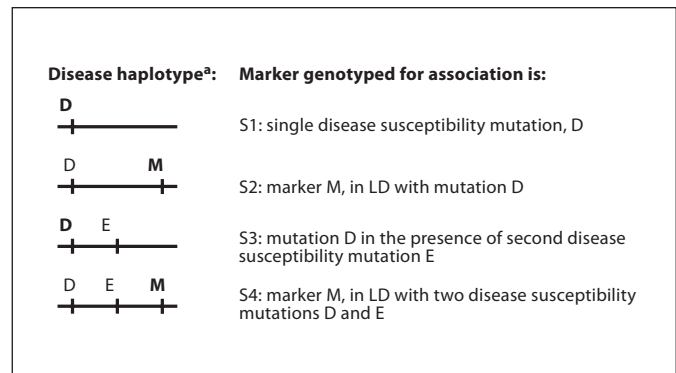


Fig. 1. Scenarios of linkage disequilibrium and genetic variation considered in sample size calculation. ^aMarker tested for association is shown in bold.

Table 1. Penetrance and genotype relative risks (GRRs) for genetic models

Genetic model	Penetrance			Genotype relative risk	
	f_d	f_D	f_{DD}	GRR_1	GRR_2
Multiplicative	α	$\alpha\gamma$	$\alpha\gamma^2$	γ	γ^2
Recessive	α	α	$\alpha\gamma$	1	γ
Dominant	α	$\alpha\gamma$	$\alpha\gamma$	γ	γ

$\alpha \neq 0, \gamma \geq 1.$

low GRRs (≤ 4), results are similar to those for the multiplicative model. Model-specific GRRs are specified for a background disease risk, α ($\alpha \neq 0$), and a genetic relative risk, γ ($\gamma \geq 1$) (table 1). We note that comparison of models with the same GRR values but different allele frequencies will result in different implied sibling relative risks [12]. Assume that N complete trios with two genotyped parents and a single affected offspring (SAO) are collected for a TDT association study. Sample sizes for the TDT can be derived under each of four alternative scenarios of linkage disequilibrium and genetic variation, as described in figure 1. The derivation of sample sizes follows a similar pattern under each scenario. This is outlined for scenario S1, assuming that a single DSM is genotyped. Corresponding equations required for the calculation of sample sizes under scenarios S2, S3 and S4 are shown in table 2. All calculations of sample size shown here assume neutral selection of alleles, no genetic drift, random mating within the population, no phenocopies and markers in Hardy Weinberg equilibrium in the population, factors which are known to affect the power of the TDT [13].

Table 2. Components required for calculation of sample size for a tested marker which is (S1) a single disease susceptibility mutation (DSM); (S2) in LD with a single DSM; (S3) one of two DSMs; (S4) in LD with two DSMs

	P(parent heterozygous SAO)	P(parent transmits D parent Dd, SAO)	Var(B _{ij} H ₁)
S1	$h_D = \frac{p_D p_d}{K_D} \left\{ p_D f_{DD} + p_d f_D + p_D f_D + p_d f_d \right\}$ <p>where</p> $K_D = p_D^2 f_{DD} + 2p_D p_d f_D + p_d^2 f_d$	$\tau_D = \frac{p_D f_{DD} + p_d f_d}{(p_D f_{DD} + p_d f_D + p_D f_D + p_d f_d)}$	$\sigma_D^2 = 1 - 2h_D(2\tau_D - 1)^2 + \frac{(p_D p_d)^2}{h_D K_D} (f_{DD} - 2f_D + f_d)$
S2	$h_M = \frac{1}{K_D} \left\{ (p_D f_{DD} + p_d f_D) 2p_D^2 P_{M D} P_{m D} \right. \\ \left. + (p_D f_{DD} + p_d f_D + p_D f_D + p_d f_d) p_D p_d (P_{M D} P_{m D} \right. \\ \left. + P_{M d} P_{m D}) + (p_D f_D + p_d f_d) 2p_d^2 P_{M d} P_{m D} \right\}$ <p>where</p> $K_D = p_D^2 f_{DD} + 2p_D p_d f_D + p_d^2 f_d$	$\tau_M = \frac{p_m}{K_D h_M} \left\{ (p_D f_{DD} + p_d f_D) P_{M D} P_D \right. \\ \left. + (p_D f_D + p_d f_d) P_{M d} P_d \right\}$	$\sigma_M^2 = 1 - 2h_M(2\tau_M - 1)^2 + C \left\{ \frac{(p_M p_m)^2}{h_M K_D} \right\}$ <p>where</p> $C = p_D^2 f_{DD} \left\{ \frac{P_{M D}^2}{P_M^2} + \frac{P_{m D}^2}{P_m^2} - \frac{2P_{M D} P_{m D}}{P_M P_m} \right\} \\ + 2p_D p_d f_D \left\{ \frac{P_{M D} P_{m d}}{P_M^2} + \frac{P_{m D} P_{m d}}{P_m^2} - \frac{P_{M D} P_{m d}}{P_M P_m} - \frac{P_{M d} P_{m D}}{P_M P_m} \right\} \\ + p_D^2 f_d \left\{ \frac{P_{M d}^2}{P_M^2} + \frac{P_{m d}^2}{P_m^2} - \frac{2P_{M d} P_{m d}}{P_M P_m} \right\}$
S3	$h_{DE} = \frac{p_D}{K_{DE}} \left\{ p_D (1-p_D) f_{DD} + p_E f_{DE} \right. \\ \left. + p_E^2 f_{EE} + 2p_E p_d f_E + p_d f_D + p_d^2 f_d \right\}$ <p>where</p> $K_{DE} = p_D^2 f_{DD} + 2p_D p_E f_{DE} + p_E^2 f_{EE} \\ + 2p_d p_D f_D + 2p_d p_E f_E + p_d^2 f_d$	$\tau_{DE} = \frac{p_D}{(p_D f_{DD} + p_E f_{DE} + p_d f_d)(1-p_D)} \\ \left\{ p_D (1-p_D) f_{DD} + p_E f_{DE} + p_E^2 f_{EE} + \right. \\ \left. 2p_E p_d f_E + p_d f_D + p_d^2 f_d \right\}$	$\sigma_{DE}^2 = 1 - 2h_{DE}(2\tau_{DE} - 1)^2 + C \left\{ \frac{p_D^2 (1-p_D)^2}{h_{DE} K_{DE}} \right\}$ <p>where</p> $C = f_{DD} + \frac{p_E^2 f_{EE}}{(1-p_D)^2} + \frac{2p_d p_E f_E}{(1-p_D)^2} \\ + \frac{p_d^2 f_d}{(1-p_D)^2} - \frac{2p_E f_{DE}}{(1-p_D)} - \frac{2p_d f_D}{(1-p_D)}$

Table 2 (continued)

	P(parent heterozygous SAO)	P(parent transmits D parent Dd, SAO)	Var(B_{ij} H_1)
S4	$h_{M,DE} = \frac{1}{K_{DE}} \left\{ \begin{aligned} &F_1(DD, DE, Dd) [F_2(D, D) + F_2(D, E) + F_2(D, d)] \\ &+ F_1(DE, EE, dE) [F_2(D, E) + F_2(E, E) + F_2(d, E)] \\ &+ F_1(Dd, dE, dd) [F_2(D, d) + F_2(d, E) + F_2(d, d)] \end{aligned} \right\}$ <p>where, for genotypes W, Y, Z, and alleles A, B,</p> $F_1(W, Y, Z) = p_D f_W + p_E f_Y + p_d f_Z$ $F_2(A, B) = p_A p_B (p_{M A} p_{M B} + p_{M B} p_{M A})$	$\tau_{M,DE} = \frac{P_m}{K_{DE} h_{M,DE}} \left\{ \begin{aligned} &F_1(DD, DE, Dd) p_D p_{M D} \\ &+ F_1(DE, EE, dE) p_E p_{M E} \\ &+ F_1(Dd, dE, dd) p_d p_{M d} \end{aligned} \right\},$ <p>and</p> $F_1(W, Y, Z) \text{ as defined for } h_{M,DE}$	$\sigma_{M,DE}^2 = 1 - 2h_{M,DE} (2\tau_{M,DE} - 1)^2 + C \left(\frac{P_m P_m}{h_{M,DE} K_D} \right)^2$ <p>where</p> $C = F_1(DD, DE, Dd) [F_3(D, D) + F_3(D, E) + F_3(D, d)] \\ + F_1(DE, EE, dE) [F_3(D, E) + F_3(E, E) + F_3(d, E)] \\ + F_1(Dd, dE, dd) [F_3(D, d) + F_3(d, E) + F_3(d, d)]$ <p>and</p> $F_1(W, Y, Z), F_2(A, B) \text{ as defined for } h_{M,DE},$ $F_3(A, B) = \frac{P_A P_B}{P_M P_m} \left\{ \begin{aligned} &(p_{M A} P_m - p_{M B} P_m) \times (p_{M B} P_m - p_{M A} P_m) \end{aligned} \right\}$

SI: Genotyped Marker Is Single Disease Susceptibility Mutation

First, we calculate the probability that a parent of an affected offspring is heterozygous (Dd) [2]:

$$h_D = P(\text{parent } Dd | SAO) = \frac{\frac{1}{2}(p_D f_{DD} + p_d f_D + p_D f_d + p_d f_d) 2 p_D p_d}{P(SAO)},$$

where the probability that a child is affected, P(SAO), is given by the population prevalence of disease:

$$K_D = P(SAO) = p_D^2 f_{DD} + 2 p_D p_d f_D + p_d^2 f_d$$

The probability that a heterozygous parent transmits allele D to the SAO can be similarly derived [2]:

$$\tau_D = P(\text{tr } D | \text{parent } Dd, SAO) = \frac{p_D f_{DD} + p_d f_D}{(p_D f_{DD} + f_D + p_d f_d)}$$

For parents of a SAO, a random variable (proposed by Risch and Merikangas [1]) is defined as follows:

$$B_{ij} = \begin{cases} +h_D^{-1/2} & \text{if parent } j \text{ in family } i \text{ is heterozygous } Dd \text{ and} \\ & \text{transmits allele } D \\ -h_D^{-1/2} & \text{if parent } j \text{ in family } i \text{ is heterozygous } Dd \text{ and} \\ & \text{transmits allele } d \\ 0 & \text{if parent } j \text{ in family } i \text{ is homozygous} \end{cases}$$

This parameterisation uses genotype relative risks which are unknown and cannot therefore be used in the test statistic to perform the TDT. In contrast, the parameterisation of Camp [2] is a function of allele frequencies only, which are known and can be used in the TDT statistic. Our current method estimates the power of the test more accurately, particularly for low frequency alleles and for large GRRs where the difference in sample size between the two methods is typically two-fold [3].

From the probability distribution of B_{ij} , the expectation and variance of B_{ij} under both the null hypothesis (H_0 : no linkage or no association) and the alternative hypothesis (H_1 : linkage and association) can be derived:

$$E(B_{ij} | H_0) = 0; \quad \text{Var}(B_{ij} | H_0) = 1$$

$$E(B_{ij} | H_1) = \sqrt{h_D} (2\tau_D - 1); \quad \text{Var}(B_{ij} | H_1) = 1 - h_D (2\tau_D - 1)^2$$

Define a random variable

$$B = \sum_i \sum_j B_{ij}$$

as the sum of parental genotypes/allele transmissions over all families ($i = 1$ to N) and parents ($j = 1, 2$). Under the null hypothesis, parents are independent between and within families, and N families contribute $2N$ independent observations of B_{ij} , so that

$$E(B | H_0) = 0; \quad \text{Var}(B | H_0) = 2N$$

Under the alternative hypothesis the expectation of B is

$$E(B | H_1) = 2N \sqrt{h_D} (2\tau_D - 1)$$

In order to obtain the variance of B, note that parents within a family are not independent given that they have a SAO, and therefore

$$\text{Var}(B) = \sum_i \sum_j \text{Var}(B_{ij}) + 2 \sum_i \text{Cov}(B_{i1}, B_{i2})$$

The variance of B can be thus derived from the probability distribution of $B_{i1}B_{i2}$ with no assumption of independence between parents:

$$\text{Var}(B | H_1) = 2N \left[1 - 2h_D(2\tau_D - 1)^2 + \frac{(p_D p_d)^2}{h_D K_D} (f_{DD} - 2f_D + f_d) \right]$$

Applying the central limit theorem gives the distribution of B under the null and alternative hypotheses as:

$$B | H_0 \sim N(0, 2N) \quad \text{and} \quad B | H_1 \sim N(2N\sqrt{h_D}(2\tau_D - 1), 2N\sigma_D^2)$$

where

$$\sigma_D^2 = 1 - 2h_D(2\tau_D - 1)^2 + \frac{(p_D p_d)^2}{h_D K_D} (f_{DD} - 2f_D + f_d).$$

and σ_D^2 is the variance of a single observation of B_{ij} under H_1 .

Thus, from normal distribution theory, the number of families, N , required to obtain $(1 - \beta)\%$ power is

$$N = \frac{(z_{1-\alpha} - z_\beta \sqrt{\sigma_D^2})^2}{2h_D(2\tau_D - 1)}, \quad (1)$$

where $z_{1-\alpha}$ is the percentage point of the normal distribution corresponding to the two-sided significance level, α , and z_β is the one-sided percentage point of the normal distribution corresponding to $100(1 - \beta)\%$ power.

S2: Genotyped Marker M Is in Linkage Disequilibrium with a Single Disease Susceptibility Mutation

Now consider a diallelic marker, M which is tested for association and which does not contribute directly to disease susceptibility, but is in linkage disequilibrium with a single disease susceptibility locus, D . At marker M , alleles M and m occur with frequency p_M and p_m . Here, only parents heterozygous at marker M (the tested marker, rather than the true DSM) are informative for the TDT. In order to calculate the probability that a parent of a single affected offspring is heterozygous, a measure of linkage disequilibrium between M and D is required. This is most usefully expressed as $p_{M|D}$, the probability that a chromosome carries a copy of marker allele M , given that it carries disease allele D . Using the total probability rule, the probability of a single affected offspring can be further conditioned on all possible genotypes at the true DSM D :

$$P(\text{SAO} | \text{parent } Mm) = \frac{P(Mm)}{K_D} \sum_G \{ P(\text{SAO} | \text{parent } Mm, G) P(\text{parent } G | Mm) \},$$

where G is the set of all genotypes $\{DD, Dd, dd\}$ at disease locus D . The probability of each genotype G given heterozygosity at

marker M can be expressed in terms of the LD measure $p_{M|D}$. For example,

$$P(DD | Mm) = \frac{P(Mm, DD)}{P(Mm)} = \frac{2P(M, D)P(m, D)}{P(Mm)} = \frac{2p_{M|D}p_D p_{m|D}p_D}{2p_M p_m}$$

Deriving $P(Dd | Mm)$ and $P(dd | Mm)$ similarly, the probability of a single affected offspring given one parent is heterozygous Mm can be written

$$P(\text{SAO} | \text{parent } Mm) = \{ (p_D f_{DD} + p_d f_D) 2p_D^2 p_{M|D} p_{m|D} + (p_D f_{DD} + f_D + p_d f_0) p_D p_d (p_{M|D} p_{m|D} + p_{M|d} p_{m|D}) + p_D f_D + p_d f_0 \} 2p_D^2 p_{M|d} p_{m|D}$$

The probability of a SAO depends only on the genotype at D , so $P(\text{SAO}) = K_D$ and thus h_M can be calculated as a function of allele frequencies, linkage disequilibrium and disease penetrance parameters as previously.

After much algebra, all other expressions follow as for the single tested mutation scenario (S1) so that the number of families required to obtain $(1 - \beta)\%$ power is then given by equation (1) above, where the required components are given in table 2. In the case that M is in perfect LD with DSM D , $p_M = p_D$, $p_{M|D} = p_{m|D} = 1$ and $p_{M|d} = p_{m|D} = 0$. Substituting these equations reduces the formulae to those for scenario 1.

S3: Genotyped Marker Is One of Two Disease Susceptibility Mutations

Now consider a second disease susceptibility mutation E , lying close to marker D such that mutation E only arises on a haplotype with the wildtype allele d at D . Then we can summarise each haplotype as one of three possibilities: a single observed mutation at D (denoted D), a single unobserved mutation at E (denoted E) or no disease mutations (denoted d). An observed wildtype allele at tested marker D may therefore represent a haplotype which carries the second risk mutation E . Note that E is not restricted to a single allele but may include multiple disease susceptibility mutations, or equivalently, allelic heterogeneity at a single disease locus. Single locus-specific disease penetrances are defined by f_{DD} , f_{DE} , f_{EE} , f_D , f_E , f_d as the probability of disease for genotypes DD , DE , EE , Dd , dE , dd respectively. Then a single genotype can be defined across the two mutations so that the observed genotypes at tested marker D are DD , DX and XX (where X denotes either the wildtype allele d or the unobserved mutation E). These three observed genotypes represent the six possible genotypes with penetrances shown in table 3. Note that the disease risk conferred by E is not necessarily equal to that due to D .

Here, only parents heterozygous at marker D are informative for the TDT. By noting that an observed genotype DX could in fact be either of two possible genotypes (DE or Dd), we can calculate the probability of a heterozygous DX parent both unconditional and conditional on a single affected offspring:

$$P(\text{parent } DX) = P(DE) + P(Dd), \\ P(\text{parent } DX | \text{SAO}) = P(\text{parent } DE | \text{SAO}) + P(\text{parent } Dd | \text{SAO}).$$

Expressions for the probability of a parent being heterozygous DX given that they have a SAO, h_{DE} , and the probability that a heterozygous parent transmits allele D to the SAO, τ_{DE} , follow as

Table 3. Observed/ actual genotypes and penetrances in the presence of two disease susceptibility alleles

Observed genotype	Actual genotypes	Penetrance
DD	DD	$p_D^2 f_{DD}$
DX	Dd, DE	$2p_D p_d f_{Dd} + 2p_D p_E f_{DE}$
XX	dd, dE, EE	$p_d^2 f_{dd} + 2p_d p_E f_{dE} + p_E^2 f_{EE}$

for scenario S1. By defining a random variable B_{ij} as previously, the expectation and variance of B under the null and alternative hypotheses are derived as for scenario S1, where h_{DE} and τ_{DE} are substituted for h_D and τ_D throughout. Note that in the derivation of $Var(B | H_1)$, the transmitted allele X is defined as any allele other than D. The frequency of the transmitted allele X is therefore scaled accordingly; for example, the probability that the transmitted allele is E is:

$$P(E | X) = \frac{P(E, X)}{P(X)} = \frac{P(E)}{1 - P(D)}$$

where X denotes any allele but not D.

Derivation of TDT sample sizes follows as previously, with summary equations shown in table 2. For a single mutation D, substituting $p_E = 0$ in all equations reduces the formulae to that for scenario S1.

S4: Tested Marker Is in Linkage Disequilibrium with Multiple Disease Susceptibility Mutations

Finally, consider marker M as defined above, which is tested for association and which does not contribute directly to disease susceptibility, but is in linkage disequilibrium with two disease susceptibility loci, D and E which occur as defined above. We impose no conditions on the extent of LD between D and M , or between E and M , or on allele frequencies at these sites. Then applying the methods described above, the expressions in table 2 can be derived. By substituting $p_E = 0$, all equations reduce to those for scenario S2. Similarly, substituting $p_M = p_D$, $p_{M|D} = p_{m|d} = 1$ and $p_{M|d} = p_{m|D} = 0$ reduces all equations to those for scenario S3.

Case-Control Sample Sizes

For comparison with the TDT, sample sizes for a case-control association study (number of case-control pairs) are now derived for each of the four scenarios. Three commonly applied test statistics for case-control association are considered: tests for difference in (a) allele frequency (equivalent to a multiplicative model); (b) homozygous DD genotype frequency (equivalent to a recessive model), and (c) homozygous dd genotype frequency (equivalent to a dominant model). For comparison with the TDT, the optimal test (i.e. that which gives the lowest sample size) is assumed and sample sizes for this optimal test are reported.

Under a test for difference in allele frequencies in cases and controls, the sample size required for Pearson's chi-squared test, with 1 degree of freedom is [14]:

$$N = \frac{z_{1-\beta} \sqrt{p_{D1}(1-p_{D1}) + p_{D0}(1-p_{D0})} + z_\alpha \sqrt{2p_D(1-p_D)}}{(p_{D0} - p_{D1})^2},$$

where

$$p_D = \frac{p_{D1} + p_{D0}}{2}$$

and p_{D1} , p_{D0} are the frequencies of allele D in cases and controls respectively. For a risk allele with frequency p_{D0} in controls and penetrance f_G for genotype G (table 1), the frequency of genotype G in cases is estimated by

$$P(G | case) = \frac{f_G P(G)}{K_D}$$

and $P(G)$ is estimated from controls. For association testing under a specific genetic model (dominant, recessive), risk allele frequencies p_{D1} and p_{D0} are replaced by risk genotype frequencies according to the genetic model assumed.

When a marker M in linkage disequilibrium with the disease susceptibility locus D is tested, genotype frequencies at M in cases are required. For genotype MM,

$$\begin{aligned} P(MM | case) &= \frac{P(MM, case)}{P(case)} \\ &= \frac{1}{P(case)} \sum_G P(case | MM, G) P(MM, G) \end{aligned}$$

where G is the set of genotypes at disease locus D . Substituting genotype penetrances and LD measures,

$$P(MM | case) = \frac{1}{K_D} \{ p_D^2 p_{M|D}^2 f_{DD} + 2p_D p_d p_{M|D} p_{M|d} f_{Dd} + p_d^2 p_{M|d}^2 f_{dd} \}.$$

Similarly,

$$P(Mm | case) = \frac{1}{K_D} \{ 2p_D^2 p_{M|D} p_{m|D} f_{DD} + 2p_D p_d (p_{M|D} p_{m|d} + p_{M|d} p_{m|D}) f_{Dd} + 2p_d^2 p_{M|d} p_{m|d} f_{dd} \},$$

$$P(mm | case) = \frac{1}{K_D} \{ p_D^2 p_{m|D}^2 f_{DD} + 2p_D p_d p_{m|D} p_{m|d} f_{Dd} + p_d^2 p_{m|d}^2 f_{dd} \}$$

where disease prevalence, K_D is calculated as for the TDT (table 2).

In the case of a second unobserved mutation E in addition to genotyped mutation D, six possible genotypes (dd, dD, dE, DD, EE, DE) are included in the calculation of disease prevalence, K_{DE} as for the TDT, although these actual genotypes correspond to only three observed genotypes (table 3) from which the frequency of the tested mutation D in cases is derived. The sample size can then be calculated as for the single tested mutation.

For marker M tested in the presence of multiple disease loci, equations for genotype frequencies at M in cases follow by similar methods, where K_{DE} is substituted in place of K_D .

Table 4. Sample sizes (total number of individuals) required for TDT and case-control (CC) association studies, where the single tested DSM, D occurs with frequency p_D

GRR	p_D	Multiplicative			Recessive			Dominant		
		TDT	CC	ratio ^a	TDT	CC	ratio ^a	TDT	CC	ratio ^a
4	0.01	534	450	1.2	2.1×10^6	87,236	2.4	561	464	1.2
	0.05	123	100	1.2	19,209	3,514	5.5	153	118	1.3
	0.1	72	58	1.2	2,814	898	3.1	111	80	1.4
	0.2	51	38	1.3	492	246	2	108	70	1.5
2	0.01	2,838	2,398	1.2	1.9×10^7	4.7×10^5	4.0	2,925	2,454	1.2
	0.05	612	516	1.2	1.6×10^5	18,922	8.5	714	578	1.2
	0.1	339	284	1.2	22,164	4,796	4.6	459	354	1.3
	0.2	207	172	1.2	3,390	1,266	2.7	381	268	1.4
1.5	0.01	9,414	7,964	1.2	7.5×10^7	1.6×10^6	46.9	9,660	8,120	1.2
	0.05	1,998	1,690	1.2	6.4×10^5	63,018	10.2	2,280	1,862	1.2
	0.1	1,080	912	1.2	85,542	15,926	5.4	1,407	1,108	1.3
	0.2	636	536	1.2	12,570	4,160	3.0	1,095	794	1.4

GRR = genotype relative risk; ^a Sample size ratio of TDT to CC. Sample sizes are given for single test ($\alpha = 0.05$).

For both TDT and case-control studies, unless stated otherwise, sample sizes required to provide 80% power to detect association are calculated assuming a nominal level of significance ($\alpha = 0.05$). All calculations are performed using R v2.2.2 for Windows (<http://www.r-project.org>).

Examples in Complex Disease: Crohn's Disease and Rheumatoid Arthritis

The methods described above are implemented for two genes, *NOD2* and *PTPN22*, which have been shown to contribute multiple risk mutations to Crohn's disease and rheumatoid arthritis respectively. *NOD2* genotype relative risks for Crohn's disease estimated previously in our UK study were used [15]. Population allele frequencies and LD measures were estimated from *NOD2* genotypes in 687 population controls from the 1958 British Birth Cohort (National Child Development Study: <http://www.cls.ioe.ac.uk>) and from Guy's Hospital, London [15]. LD was calculated from haplotype frequencies, estimated in Haploview [16]. All *PTPN22* allele frequencies, linkage disequilibrium and genotype relative risks were extracted from tables of previously published results [10]. Unless stated otherwise, sample sizes are derived for a single test of association ($\alpha = 0.05$) with no correction for multiple testing incorporated. All analyses are implemented in R v2.2.0 for Windows.

Results

Where a single disease mutation D exists and is genotyped, sample sizes required for 80% power to detect association were calculated for a range of genotype relative risks {4, 2, 1.5} and susceptibility allele frequencies {0.01,

0.05, 0.1, 0.2}. Sample sizes for TDT and case-control studies (expressed as the total number of individuals to be genotyped) are shown in table 4. Sample sizes required for the TDT are greater than for a case-control study, irrespective of GRR or allele frequency. For a nominal level of significance, sample sizes are generally feasible (<1,000 trios) for allele frequencies >10% for all models except recessive. For a recessive model, sample sizes increase substantially for rare allele frequencies or low genotype relative risks such that for a GRR of ≤ 2 and an allele frequency of $\leq 10\%$, several thousand TDT trios would be required for sufficient power to detect association. The equivalent scenario for a case-control study would require more than 2,500 case-control pairs. Allowing for multiple testing of ten SNPs as might typically be the case in a candidate gene study ($\alpha = 0.005$), an approximately two-fold increase in sample size (number of genotypes) would be required than for a single test; this increase is fairly constant across genetic models. The corresponding increase in sample size for 100 tests ($\alpha = 0.0005$) compared with a single test is approximately three-fold, or six-fold for a genome-wide test of association ($\alpha = 5 \times 10^{-8}$).

For a nominal level of significance, the relative increase in the number of genotypes required for the TDT is less than 1.3-fold that for a case-control study for the multiplicative genetic model, with little variation across allele frequencies and GRRs. Under a dominant model, the relative difference between TDT and case-control

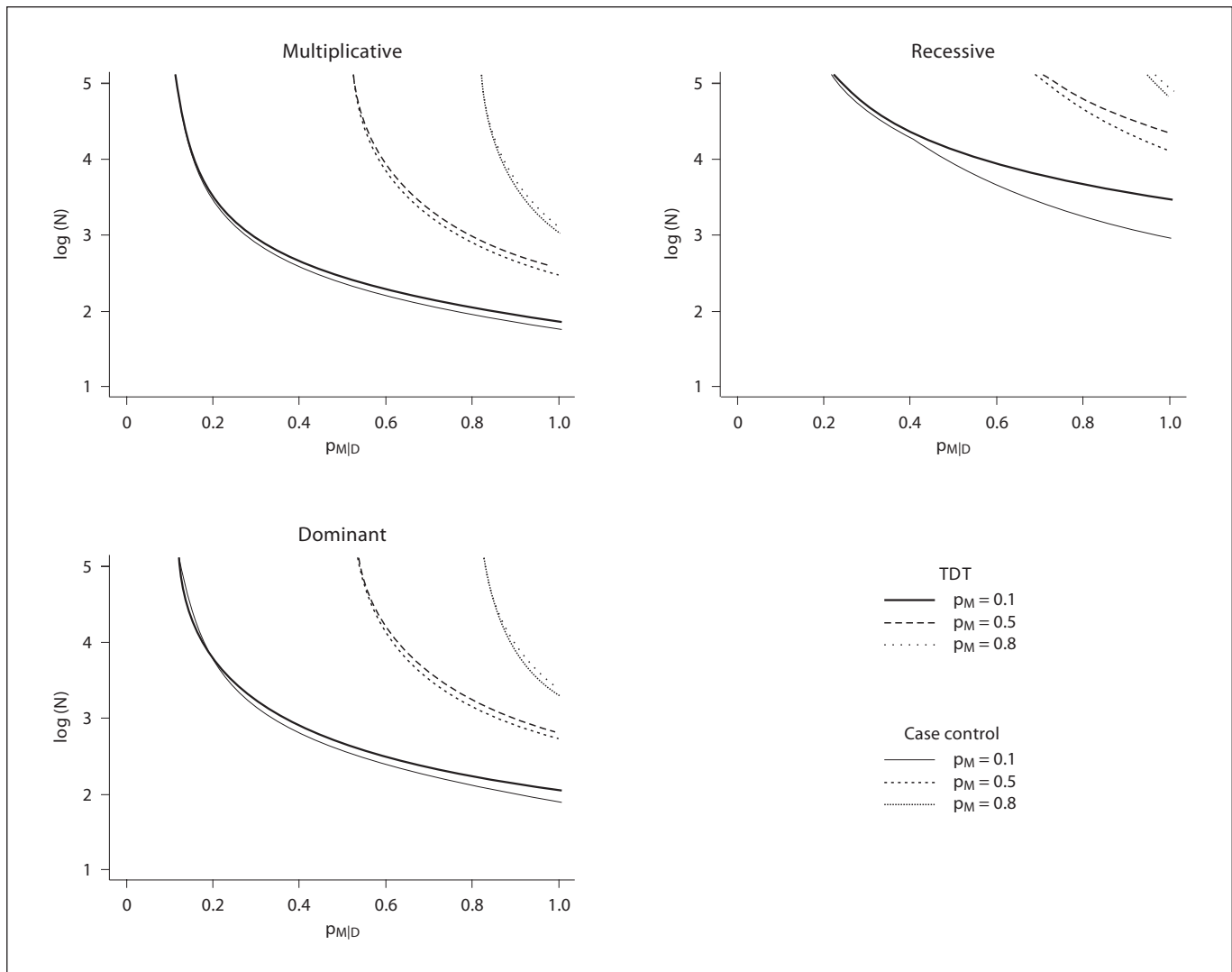


Fig. 2. Sample size (shown as $\log(N)$) required to attain 80% power to detect association under scenario S2, genotyping marker M in LD with the DSM D . Mutation D has frequency 10%, and genotype relative risk = 4.

sample sizes increases as the disease allele frequency increases. This difference is greatest for high GRRs; for example, if the disease allele frequency is 0.2 and the GRR for carriers of the mutation is 4, the TDT would require 108 genotyped individuals compared with 70 required for a case-control study. For common risk alleles (>50% frequency), the relative increase is much higher. For example, if the disease allele frequency is 0.8 and the GRR for carriers of the mutation is 4, the TDT would require 4,515 genotyped individuals compared with 872 required for a case-control study, a 5-fold increase in sample size. Under a recessive model, the sample size required for the TDT is generally much higher than a case-control study,

with the highest TDT:CC ratio for low allele frequencies and small GRRs, although for susceptibility genes of such weak effect, even the case-control sample size required is beyond the resources of most studies. The increased number of TDT trios is partly due to the test statistics used: the case-control test is maximised over multiplicative, dominant and recessive models, while the TDT test assumes a multiplicative model, and equivalent tests for other models were not applied [17], as reflects common practice with the TDT.

For a tested marker in LD with a single disease mutation, sample sizes for the TDT and case-control study are shown in figure 2 for $GRR = 4$, a mutation occurring with

Table 5. Sample sizes (number of genotyped individuals) required for the TDT in the presence of multiple disease mutations, where the tested DSM, D occurs with frequency P_D and the ratio of frequencies between D and E is given by $P_E:P_D$

GRR	P_D	Multiplicative				Recessive				Dominant				
		$P_E:P_D$	0:1	1:2	1:1	2:1	0:1	1:2	1:1	2:1	0:1	1:2	1:1	2:1
4	0.01		534	549	564	597	2.1×10^6	9.6e5	5.5e5	2.5e5	561	591	624	693
	0.05		123	141	159	207	19,209	9,411	5,868	3,255	10,153	195	249	384
	0.1		72	93	123	201	2,814	1,551	1,107	837	111	174	267	612
	0.2		51	84	144	468	492	366	375	702	108	258	603	4,167
2	0.01		2,838	2,883	2,934	3,033	1.9×10^7	8.5e6	4.8e6	2.2e6	2,925	3,024	3,126	3,342
	0.05		612	669	729	870	1.6×10^5	77,022	46,566	24,141	714	843	996	1,386
	0.1		339	405	486	711	22,164	11,469	7,611	4,914	459	642	900	1,800
	0.2		207	306	468	1,296	3,390	2,175	1,884	2,547	381	777	1,647	10,092
1.5	0.01		9,414	9,543	9,672	9,942	7.5×10^7	3.4e7	1.9e7	8.8e6	9,660	9,927	10,200	10,770
	0.05		1,998	2,145	2,307	2,676	6.4×10^5	3.0e5	1.8e5	91,599	2,280	2,619	3,015	4,014
	0.1		1,080	1,254	1,464	2,043	85,542	43,458	28,236	17,340	1,407	1,878	2,526	4,740
	0.2		636	891	1,296	3,315	12,570	7,692	6,297	7,464	1,095	2,067	4,167	23,808

GRR = genotype relative risk. Sample sizes are given for single test ($\alpha = 0.05$).

10% frequency ($p_D = 0.1$) and a range of tested marker frequencies ($p_M = \{0.1, 0.5, 0.8\}$). A non-smooth curve for the case-control study indicates that the optimal test (i.e. that which gives the minimum sample size) changes from a comparison of homozygote frequencies to that of allele frequencies. When $p_M = p_D$ and $p_{M|D} = 1$, M is in absolute LD with D and sample sizes are equal to those shown in table 4 for a single tested mutation. However, as the extent of LD between marker and disease mutation decreases ($p_{M|D} < 1$), the sample size increases rapidly (fig. 2). An asymptotic limit is reached at $p_{M|D} = p_M$, where the study has no power to detect association as M and D are independent ($r^2 = D' = 0$). Sample sizes also increase substantially as the difference in frequency between the tested marker allele M and the disease mutation D increases, even in the presence of complete LD ($D' = 1$); this is true for a tested marker of greater or lower frequency than the disease mutation.

For a tested mutation D in the presence of a second (unobserved) mutation E (or group of mutations), TDT and case-control sample sizes are calculated for tested mutation frequencies of 0.01, 0.05, 0.1 and 0.2, allowing for a second mutation with a frequency ratio of 1:2, 1:1 and 2:1 compared with the tested mutation. For example, given a tested mutation frequency, p_D of 0.1, sample sizes are calculated for unobserved mutation frequencies $p_E = \{0.05, 0.1, 0.2\}$. Sample sizes for the TDT are shown in table 5.

For multiplicative and dominant models, the sample sizes required for 80% power to detect association increase as the frequency of the second mutation increases. This increase is highest for common tested alleles and a high GRR. For a single tested mutation of 10% frequency and a multiplicative genetic model with $GRR = 2$, 339 TDT genotypes (113 trios) are required. If this combined mutation frequency of 10% is distributed between two (or more) alleles, the sample sizes can increase substantially. For example, given two mutations of equal frequency ($p_D = p_E = 0.05$), testing one of these mutations would require 729 genotypes (243 trios). A tested mutation of 7.5% ($p_D = 0.075, p_E = 0.025$) would require 468 genotypes (156 trios) compared with 1,509 genotypes (503 trios) required if the rarer mutation was tested ($p_D = 0.025, p_E = 0.075$). Under a recessive model and for moderate frequencies of the second mutation, sample sizes are lower than for a single tested mutation. For example, if the frequency of the tested mutation is 5% and the GRR is 4, then using the TDT, a single tested mutation requires 19,209 genotypes (6,403 trios) compared with 3,255 genotypes (1,085 trios) when a second mutation of 10% frequency is present.

These results imply that the sample sizes required to detect association may be substantially underestimated if a single disease mutation is assumed when allelic heterogeneity exists. In the presence of additional mutations, the relative increase in the sample size required for the TDT compared with a case-control study is generally similar to that required for a single tested mutation. In-

terestingly, under a recessive model, for common tested allele frequencies and as the relative frequency of the second mutation increases, a test for difference in allele frequency between cases and controls is a more powerful test than the usual test for difference in the frequency of mutation homozygotes.

Application to Crohn's Disease Susceptibility Locus NOD2

Three rare *NOD2* mutations (L1007fs, R702W, G908R), all with frequency <5% in controls, have been identified as contributing directly and independently to Crohn's disease susceptibility [7–9]. A fourth, common, non-synonymous SNP (P268S) (frequency 26%) also showed an association with disease [7, 15]. However, each rare mutation usually occurs on a common haplotype background containing 268S and the association with P268S was due to the LD with each of the rare mutations [15].

In this example, P268S is equivalent to tested marker *M* in strong LD with each of three rare disease susceptibility mutations (L1007fs, R702W, G908R), as in scenario S4. Estimated minor allele frequencies were 1.7% (1007fs), 4.8% (702W), 0.95% (908R) and 26.0% (268S). Pooling each of the three mutations into a composite *NOD2* disease susceptibility mutation, $\Pr(268S | DSM) = 0.89$. Each of the rare mutations has been shown to confer similar risk of disease and for simplicity, we will assume equal genotype relative risks of 3.0 for individuals heterozygous for a rare mutation and 23.4 for individuals who are homozygous (carrying two copies of the same mutation) or compound heterozygous (carrying two mutations but with a different mutation present on each chromosome) [15].

TDT and case-control sample sizes were calculated assuming the tested mutation to be: (i) a single mutation (for each risk mutation individually and for the pooled *NOD2* DSM); (ii) one of two mutations, where the second mutation is defined by the pooled alleles of the remaining two risk mutations (for example, L1007fs tested in the presence of pooled 702W/908R risk alleles), and (iii) a marker (P268S) in LD with the pooled *NOD2* DSM.

Sample sizes estimated to detect each single mutation (ignoring the presence of other mutations) ranged from 307 trios (or 384 CC pairs) required for the rarest SNP (G908R) to 52 trios (62 CC pairs) required for the more common SNP (R702W). However, if the sample sizes to detect association are correctly calculated by modelling the other two pooled mutations, the required sample sizes decrease, indicative of the moderate heterozygous risk ($GRR_1 = 3.0$) and high homozygous risk ($GRR_2 = 23.4$),

Table 6. Summary of *PTPN22* haplotype structure and association results

Haplotype (frequency)	SNP (frequency)				
	SNP2 (23.9%)	R620W (8.6%)	SNP27 (41.7%)	SNP35 (20.6%)	SNP37 (17.8%)
H1 (12.0%)	A	C	T	C	A
H2 (8.6%)	A	T	T	C	A
H3 (3.3%)	A	C	T	T	A
H4 (17.8%)	G	C	T	T	C
H5–10 (58.3%)	G	C	C	T	A

Extracted from Carlton et al. (tables 1 and 3). SNP frequencies are estimated from the pooled cohort of 1,797 controls (see Carlton et al. table 1). Alleles associated with increased disease risk are shown in bold.

which tends towards a recessive mode of inheritance. The greatest relative difference in sample size (compared with the assumed absence of other mutations) is observed for G908R in the presence of 1007fs/702W, for which 274 trios (344 CC pairs) would be required to detect an association. If a surrogate marker could be identified with frequency equal to the pooled frequency of all three SNPs and in complete LD, the sample size would reduce to just 32 trios (37 CC pairs). Direct testing of P268S would require a modest sample size of 106 trios (131 CC pairs). This illustrates the power of a common tagging SNP to detect association in the presence of allelic heterogeneity, provided the LD pattern of multiple alleles is consistent, with mutations all existing on haplotypes with the same allele of the tagging SNP.

Application to Rheumatoid Arthritis Susceptibility Locus PTPN22

A non-synonymous mutation in the *PTPN22* gene, R620W, is associated with several autoimmune diseases including rheumatoid arthritis [18–20]. A detailed haplotype analysis of this gene in two large independent case-control cohorts revealed that at least one additional *PTPN22* variant contributes to RA susceptibility [10] and multiple associated *PTPN22* haplotypes have also been described in Type I diabetes [21]. Specifically, two individual haplotypes (H2 and H4 as defined by Carlton et al.) were identified which increased risk of RA, whereby 620W occurred uniquely on H2, and H4 could be defined by the rare allele at any of three SNPs (denoted SNP15+, SNP36+ and SNP37+) (table 6). Several other SNPs

showed association with RA including SNP2, SNP27 and SNP35. However, these were in strong LD with H2 and/or H4, resulting in an indirect association. Here, SNP27 is in complete LD ($D' = 1$) with both R620W and SNP37 and thus presence of either 620W or SNP37+ indicates the presence of SNP27+ also. Of haplotypes carrying the minor allele at SNP27 (T), 67% carry either 620W or SNP37+. SNP2 (or SNP35) demonstrates a more complex picture, where 620W occurs on the same haplotype with SNP2+, but SNP37+ lies exclusively on a haplotype with the wild-type allele of SNP2. Pooling both sample sets from Carlton et al. [10], genotype relative risks for H2 and H4 are $f_{H2} = 2.10$, $f_{H2,H2} = 2.30$, $f_{H4} = 1.34$, $f_{H4H4} = 1.67$ and $f_{H2,H4} = 2.64$.

If 620W were assumed to be a unique DSM in the *PTPN22* gene contributing to disease susceptibility, then power calculations would estimate 140 TDT trios or 166 case-control pairs required for 80% power to detect an association. However, in the presence of an additional DSM (SNP37+), the true sample size required increases to 194 trios or 232 CC pairs. Conversely, if SNP37 was tested for association in the presence of 620W, 1,528 trios or 1,939 CC pairs would be required, compared with 507 trios or 644 CC pairs if SNP37 was a single DSM. Although 620W occurs at lower frequency than SNP37+, the smaller sample sizes required are due to the higher genotype relative risks associated with R620W (H2).

The more common SNP27, with a frequency of 41%, defines a set of haplotypes (H1–H4) which includes both DSMs (620W, SNP37+). Testing this common SNP for association would require considerably smaller sample sizes (44 trios or 71 CC pairs) for equivalent power, demonstrating that direct testing of either of the two causal mutations is not necessarily the most powerful test. In contrast, direct testing of SNP2 would require 1,428 trios or 6,550 CC pairs as SNP37+ occurs with the wildtype allele of SNP2 and therefore dilutes the effect of 620W. As in *NOD2*, genotyping a common SNP with a consistent pattern of LD for the mutations (SNP27) decreases the sample size required, but genotyping SNP2, where the mutations at R620W and SNP37 are carried on different haplotypes, increases the required sample size considerably.

Discussion

In this study, the methods of Camp [2, 3] to calculate sample sizes for the TDT under a range of genetic models have been extended to allow for linkage disequilibrium

between a tested marker and the true disease susceptibility mutation. The methods described here overcome previous assumptions of independence of parental genotypes, and have also been extended to allow for multiple disease susceptibility mutations. We have implemented the model parameterisation suggested by Risch and Merikangas [1] which estimates the power of the TDT more accurately than that used by Camp [2, 3]. These equations enable accurate estimates of sample size to be made, to ensure that association studies are adequately powered in realistic genetic scenarios, not only in the most simplistic case of genotyping a single DSM. R programs for sample size calculation under each of the four scenarios are available from the authors upon request.

Several other approaches have been used to assess the power of the TDT [4–6, 22–24]. In particular, the power of the TDT in the presence of allelic heterogeneity was first considered by Slager et al. [22] who used a similar approach to that implemented here, although with a parameterisation similar to that of Camp [2]. The most generalised method to date is that proposed by Chen and Deng [25], whose approach incorporates both LD and allelic heterogeneity whilst allowing for families with a number of affected and unaffected sibs as well as affection status of parents. However, the computer program ‘TDT Power Calculator’ implementing their method does not allow for allelic heterogeneity, and covers only scenarios S1 and S2. Our development therefore provides users with a straightforward method to assess sample size under complex association models.

Allelic heterogeneity can substantially increase the sample sizes required to detect association, although the presence of several disease mutations may increase power when the underlying genetic inheritance model is recessive. A single disease mutation with a recessive mode of inheritance will typically require infeasible sample sizes, particularly if the mutation occurs at low frequency. Therefore, the increase in power which may occur in the presence of several disease mutations acting in a recessive manner is reassuring and provides hope that the identification of, in particular, low frequency genetic variation of moderate or weak effect may be less difficult than previously thought. Our model of allelic heterogeneity is limited to two DSMs being in complete LD, so that only three of the possible four haplotypes arise for the loci (although no such condition is imposed on LD between the tested marker and either DSM). The limitation is imposed for tractable algebraic manipulations, and fits the current data on *NOD2* and *PTPN22* well. It may also apply to other complex disease associations, particularly as

co-occurrence of two rare mutations will be infrequent, and may be ignored to give an adequate approximation for sample size estimates.

The application of the methods to *NOD2* and *PTPN22* demonstrates the complexity of the design and interpretation of results from association studies. It is clear from these results that the most powerful method of association testing may not be obtained from direct testing of single disease mutations. Alternative methods include simultaneous testing of several alleles or 'omnibus' tests [26]. The value of haplotype analysis is well-documented and can be more powerful than single marker tests [27, 28], particularly in the presence of multiple disease susceptibility alleles [29]. However, these methods are often restricted to common tagging SNPs and the pooled frequency of rarer mutations may still be considerably lower than a tested haplotype resulting in lower power to detect association. Thus weak association observed with individual SNPs or haplotypes may in fact be indicative of multiple unobserved disease mutations in LD with the tested SNP. Caution is therefore advised before discarding positional and functional candidate genes in which weak association is observed with common non-coding SNPs or haplotypes.

Here, allelic heterogeneity is restricted to two susceptibility mutations; however, these methods are easily extended to include additional mutations with complex LD relationships. In the presence of allelic heterogeneity, multi-allelic methods may be more powerful. For example, the extended TDT (E-TDT) [30] uses a logistic re-

gression method to identify groups of alleles showing transmission distortion.

In summary, the presence of linkage disequilibrium between a tested marker and the true disease susceptibility mutation can substantially increase the sample size requirements of both the TDT and the case-control study. In contrast, the presence of multiple mutations can actually increase the power to detect association, even if some of these disease mutations are unobserved or not directly tested for association. Genetic variation across a region studied should be carefully evaluated and consideration should be given to possible LD and allelic heterogeneity when evaluating the power of an association study.

Acknowledgments

S.A.F. is an RCUK Academic Fellow. This work was supported by the Wellcome Trust (076024 to CML and 072029 to C. Mathew). We are grateful to Professor Christopher Mathew, King's College London School of Medicine, for the provision of *NOD2* R702W and L1007fs genotypes as previously documented [15]. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the Medical Research Council (grant G0000934) and The Wellcome Trust (grant 068545/Z/02). We gratefully acknowledge the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory for allowing the use of *NOD2* G908R and P268S genotypes from their nsSNP genome scan data (data version 1.0) [31]. Those who carried out the data collection and original analysis bear no responsibility for further analysis and interpretation of these data.

References

- 1 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996;273:1516–1517.
- 2 Camp NJ: Genomewide transmission/disequilibrium testing – consideration of the genotypic relative risks at disease loci. *Am J Hum Genet* 1997;61:1424–1430.
- 3 Camp NJ: Genomewide transmission/disequilibrium testing: A correction. *Am J Hum Genet* 1999;64:1485–1487.
- 4 Knapp M: A note on power approximations for the transmission/disequilibrium test. *Am J Hum Genet* 1999;64:1177–1185.
- 5 Abel L, Muller-Myhsok B: Maximum-likelihood expression of the transmission/disequilibrium test and power considerations. *Am J Hum Genet* 1998;63:664–667.
- 6 Tu IP, Whittemore AS: Power of association and linkage tests when the disease alleles are unobserved. *Am J Hum Genet* 1999;64:641–649.
- 7 Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G: Association of *NOD2* leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001;411:599–603.
- 8 Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar JP, Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nunez G, Cho JH: A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* 2001;411:603–606.
- 9 Hampe J, Cuthbert A, Croucher PJP, Mirza MM, Mascheretti S, Fisher S, Frenzel H, King K, Hasselmeier A, MacPherson AJS, Bridger S, van Deventer S, Forbes A, Nikolaus S, Lennard-Jones JE, Foelsch UR, Krawczak M, Lewis C, Schreiber S, Mathew CG: Association between insertion mutation in *NOD2* gene and Crohn's disease in German and British populations. *Lancet* 2001;357:1925–1928.
- 10 Carlton VEH, Hu XL, Chokkalingam AP, Schrodi SJ, Brandon R, Alexander HC, Chang M, Catanese JJ, Leong DU, Ardlie KG, Kastner DL, Seldin MF, Criswell LA, Gregersen PK, Beasley E, Thomson G, Amos CI, Begovich AB: *PTPN22* genetic variation: Evidence for multiple variants associated with rheumatoid arthritis. *Am J Hum Genet* 2005;77:567–581.

- 11 Palo OM, Antila M, Silander K, Hennah W, Kilpinen H, Soronen P, Tuulio-Henriksson A, Kieseppa T, Partonen T, Lonnqvist J, Peltonen L, Paunio T: Association of distinct allelic haplotypes of DISC1 with psychotic and bipolar spectrum disorders and with underlying cognitive impairments. *Hum Mol Genetics* 2007, doi:10.1093/hmg/ddm207.
- 12 Iles MM: Effect of mode of inheritance when calculating the power of a transmission disequilibrium test study. *Hum Hered* 2002;53: 153–157.
- 13 Xiong M, Guo S-W: The power of linkage detection by the transmission disequilibrium tests. *Hum Hered* 1998;48:295–312.
- 14 Kirkwood BR. *Essentials of Medical Statistics*. Blackwell Scientific, 1988.
- 15 Cuthbert AP, Fisher SA, Mirza MM, King K, Hampe J, Croucher PJP, Mascheretti S, Sanderson J, Forbes A, Mansfield J, Schreiber S, Lewis CM, Mathew CG: The contribution of NOD2 gene mutations to the risk and site of disease in inflammatory bowel disease. *Gastroenterology* 2002;122:867–874.
- 16 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005;21: 263–265.
- 17 Schaid DJ: Likelihoods and TDT for the case-parents design. *Genet Epidemiol* 1999; 16:250–260.
- 18 Begovich AB, Carlton VEH, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, Ardlie KG, Huang QQ, Smith AM, Spoerke JM, Conn MT, Chang M, Chang SYP, Saiki RK, Catanese JJ, Leong DU, Garcia VE, McAllister LB, Jeffery DA, Lee AT, Batliwalla F, Remmers E, Criswell LA, Seldin MF, Kastner DL, Amos CI, Sninsky JJ, Gregersen PK: A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet* 2004;75:330–337.
- 19 Steer S, Lad B, Grumley JA, Kingsley GH, Fisher SA: Association of R602W in a protein tyrosine phosphatase gene with a high risk of rheumatoid arthritis in a British population: Evidence for an early onset/disease severity effect. *Arthritis Rheum* 2005;52:358–360.
- 20 Criswell LA, Pfeiffer KA, Lum RF, Gonzales B, Novitzke J, Moser KL, Begovich AB, Carlton VEH, Li W, Lee AT, Ortmann W, Behrens TW, Gregersen PK: Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes. *Am J Hum Genet* 2005;76:561–571.
- 21 Onengut-Gumuscu S, Buckner JH, Concannon P: A haplotype-based analysis of the PTPN22 locus in type 1 diabetes. *Diabetes* 2006;55:2883–2889.
- 22 Slager SL, Huang J, Vieland VJ: Effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genet Epidemiol* 2000;18:143–156.
- 23 McGinnis R: General equations for P-t, P-s, and the power of the TDT and the affected-sib-pair test. *Am J Hum Genet* 2000;67: 1340–1347.
- 24 Iles MM: On calculating the power of a TDT study – comparison of methods. *Ann Hum Genet* 2002;66:323–328.
- 25 Chen WM, Deng HW: A general and accurate approach for computing the statistical power of the transmission disequilibrium test for complex disease genes. *Genet Epidemiol* 2001;21:53–67.
- 26 Longmate JA: Complexity and power in case-control association studies. *Am J Hum Genet* 2001;68:1229–1237.
- 27 de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: Efficiency and power in genetic association studies. *Nat Genet* 2005;37:1217–1223.
- 28 Schaid DJ: Power and sample size for testing associations of haplotypes with complex traits. *Ann Hum Genet* 2006;70:116–130.
- 29 Morris RW, Kaplan NL: On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol* 2002;23:221–233.
- 30 Sham PC, Curtis D: An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* 1995;59: 323–336.
- 31 Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JMM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA: Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* 2005;37:1243–1246.