Original Paper

Human Heredity

Hum Hered 2009;68:106–116
DOI: 10.1159/000212503

Received: January 8, 2008
Accepted after revision: December 10, 2008
Published online: April 9, 2009

Comparison of Methods for Combining Case-Control and Family-Based Association Studies

Beate Glaser[a] Peter Holmans[b]

[a]Department of Social Medicine, University of Bristol, Oakfield House, Oakfield Grove, Bristol, and [b]Biostatistics and Bioinformatics Unit, and Department of Psychological Medicine, Cardiff University, School of Medicine, Heath Park, Cardiff, UK

Key Words
Power · Association · Case-control · Case-parent trios · Affected sibpairs · Combined test · Simulation

Abstract
Objectives: Combining the analysis of family-based samples with unrelated individuals can enhance the power of genetic association studies. Various combined analysis techniques have been recently developed; as yet, there have been no comparisons of their power, or robustness to confounding factors. We investigated empirically the power of up to six combined methods using simulated samples of trios and unrelated cases/controls (TDTCC), trios and unrelated controls (TDTC), and affected sibpairs with parents and unrelated cases/controls (ASPFCC). ***Methods:*** We simulated multiplicative, dominant and recessive models with varying risk parameters in single samples. Additionally, we studied false-positive rates and investigated, if possible, the coverage of the true genetic effect (TDTCC). ***Results/Conclusions:*** Under the TDTCC design, we identified four approaches with equivalent power and false-positive rates. Combined statistics were more powerful than single-sample statistics or a pooled χ^2-statistic when risk parameters were similar in single samples. Adding parental information to the CC part of the joint likelihood increased the power of generalised logistic regression under the TDTC but not the TDTCC scenario. Formal testing of differences between risk parameters in subsamples was the most sensitive approach to avoid confounding in combined analysis. Non-parametric analysis based on Monte-Carlo testing showed the highest power for ASPFCC samples.

Introduction

An increase in sample size is for most genetic association studies on common disease a crucial step towards enhancing their inherent statistical power [1]. It is therefore a natural consideration to exploit available but fixed sample resources and combine data from different genetic sampling designs (e.g. linkage, outbred or family-based-association samples) for joint association analysis [2–13] or meta-analysis [14].

Specific ascertainment schemes exist for affected sibpair (ASP) linkage studies, which are commonly used to detect linkage to dichotomous traits [15], and genetic as-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/09/0682–0106$26.00/0

Accessible online at:
www.karger.com/hhe

Peter Holmans
Biostatistics and Bioinformatics Unit
Cardiff University, Henry Wellcome Building
Heath Park, Cardiff, CF14 4XN (UK)
Tel. +44 29 2074 8154, Fax +44 29 2074 6554, E-Mail holmanspa@cf.ac.uk

sociation studies, which test for a relationship between marker alleles/genotypes and trait in a sample of unrelated cases and unrelated controls (CC), or through investigation of families consisting of cases and their parents (trios). Both ASP linkage studies and association studies of population samples aim to estimate disease risks, in relatives or in the general population respectively [1] and were traditionally coupled in that whole genome ASP linkage analysis was followed up by smaller scale association studies. Techniques for testing association in combined samples consisting of family data and unrelated individuals have been developed and include a pooled score statistic [3], a combined relative risk (RR) estimate derived from generalised logistic regression [4], or multinomial regression [5], a weighted combination of odds ratio (OR) estimates [2], a combined likelihood ratio test (LRT) relaxing assumptions of Hardy-Weinberg equilibrium (HWE), random mating and multiplicative genetic effects [6], a likelihood framework combining sibships with unrelated individuals [7], a principal component based analysis (PCA) for family and unrelated samples [8], non-parametrically derived null distributions for test statistics based on Monte Carlo (MC) simulations [9, 10] and an yet unpublished combined global LRT, which is based on full likelihood models for nuclear families [11] [12]. Also, a methodology for combining CC with family samples for microsatellite data has been proposed [13].

Despite the large number of available methods for joint association analysis no comparisons of their power have been performed yet. We therefore propose to investigate empirically the power of combined analysis techniques for SNP data under three different study designs: combinations of case-parent trios and unrelated cases/controls (TDTCC), combinations of trios and unrelated controls (TDTC), and combinations of affected sibpairs with parents and unrelated case/controls (ASPFCC). The latter scenario refers to a sample consisting of families with two affected offspring with genotyped parents, and unrelated cases and controls.

For the purpose of our simulation study, we exclude from the aforementioned combined approaches all methods, which are sensitive to population stratification due to the form of the selected likelihood [7, 11, 12] and focus specifically on techniques for biallelic markers, which do not require additional marker locus information (e.g. required by GC-like approaches [8]) when adjusting for stratification.

This study will consider situations where allele frequencies differ between the case-control and family-based samples ('between-sample' stratification), as well as situations where genetic risk effects differ between the samples. For the ASPFCC scenario, we will investigate how well the various analysis methods correct for linkage, since this is known to cause inflated Type I error rates if not corrected.

Methods

Descriptions of Single and Combined Methods

For simplicity, the combined approaches will henceforth be referred to as Schaid and Rowland method [3], Nagelkerke method [4], Dudbridge method [5], Kazeem & Farrall method [2], Epstein method [6] and Allen-Brady method [9, 10].

- Single Sample Statistics

For CC samples an allelic χ^2-test was performed and for trio samples a transmission disequilibrium test (TDT) [16] carried out. For samples of two affected sibpairs and their families, χ^2-statistics were calculated using the Unphased software (v.3.0.7) [5]. Under the TDTCC and the ASPFCC scenario, single χ^2-statistics were corrected for two independent tests.

- Pooled χ^2-Statistic

Given the equivalence of test statistics from CC and trio samples in the presence of HWE, multiplicative disease penetrances and the absence of recombination, a simple non-weighted pooled χ^2-statistic was obtained.

$$\chi^2_{Pooled} = \chi^2_{CC} + \chi^2_{TDT} \sim \chi^2\,(2) \text{ under } H_0$$

- Kazeem & Farrall Method [2]

A weighted combined log OR estimate (Ψ) based on single sample estimates from CC and trios data can be derived as

$$\psi = \frac{\sum_{i=1}^{2} \omega_i \varphi_i}{\sum_{i=1}^{2} \omega_i} \qquad Var(\psi) = \frac{1}{\sum_{i=1}^{2} \omega_i}$$

with $\varphi_1 = \ln(OR_{CC})$, $\varphi_2 = \ln(OR_{TDT})$, $\omega_1 = 1/\sigma^2_{ORCC}$, $\omega_2 = 1/\sigma^2_{ORTDT}$. The significance of the combined OR estimate can be tested using the test statistic

$$\chi^2_{\omega} = \frac{(\psi - \psi_0)}{Var(\psi)} \sim \chi^2\,(1) \text{ under } H_0$$

For the current study, the method was implemented in Perl scripts (Active Perl v.5.8.8.819).

- Schaid and Rowland Method [3]

A pooled score statistic was developed for the combination of family, sibling and CC data. In the presence of two strata consisting of parental (pseudo-controls) (1) and unrelated controls (2), the following likelihood can be formulated

$$L = L_1 \times L_2$$
$$L_1 = \Pi \Pr(Y \mid G_{cases}; G_{pseudocontrols})$$
$$L_2 = \Pi \Pr(Y \mid G_{foundercases}; G_{controls})$$

where, for each stratum, Y represents the affected status and G refers to the observed genotype. The first part of the likelihood corresponds to a conditional logistic regression of the disease status on genotypes of cases and pseudo-controls and is equivalent to

$$L_1 = \Pi \Pr(G_{\text{cases}} \mid G_{\text{parents}}; p, \gamma_1, \gamma_2)$$

where p is the disease MAF and γ_1, γ_2 are the RRs for heterozygous and homozygous risk genotypes, respectively. The second part of the likelihood represents the standard CC logistic regression. For both parts of the likelihood, scores and their variances can be derived, where U is the score parameter and V the variance of the score parameter; β refers to the regression parameter in L.

$$U_1 = \frac{\partial \ln L_1}{\partial \beta_1} = \sum \left(x_{a1} - e_{a1}\right) \qquad V_1 = \sum Cov_0\left(X_{a1} X'_{a1}\right)$$

$$U_2 = \frac{\partial \ln L_2}{\partial \beta_2} = \left(x_{a2} - e_{a2}\right) \qquad V_2 = Cov_0\left(X_{a2} X'_{a2}\right)$$

For the family-based part of the likelihood, x_{a1} is the vector of observed counts of allele x in the case genotype within a trio. e_{a1} represents the vector of expected allele counts for each trio under the H_0. V_1 is the covariance matrix of X_{a1}, the vector of the possible x-coded offspring genotypes within a trio; and scores and variances are summed across all trios. Similarly, for the CC part of the likelihood, x_{a2} represents the vector of allele counts in cases in the complete CC sample, e_{a2} its expected vector, V_{a2} the covariance matrix of X_{a2}, which is the vector of observed genotypes.

A pooled score statistic for the two strata can be derived as

$$S_{Pooled} = (U_1 + U_2)'(V_1 + V_2)^{-1}(U_1 + U_2) \sim \chi^2\,(1) \text{ under } H_0$$

The pooled score statistic combining parental and unrelated CC information was implemented in R code (v.2.5.0). We excluded the pooled score statistic [3] from investigation under the ASPFCC scenario as it is known that it does not account for correlated transmissions of parental alleles to multiple affected offspring in the presence of linkage between the disease and marker loci [17].

- Nagelkerke Method [4]

A combined allelic RR estimate for a sample of unrelated cases, unrelated controls and trios, or a sample of unrelated control and trios data can be obtained using Poisson regression. Assuming a multiplicative disease model and the presence of HWE, the allelic RR corresponds to a heterozygous genotypic RR γ_1 [18]. A combined γ_1 estimate for CC and TDT studies, or unrelated controls and TDT studies can be derived from the following likelihood function

$$\begin{aligned} &Likelihood = L_1 \times L_2 \times L_3 \times L_4 \\ &L_1 = \Pi \Pr(G_{cases} \mid G_{parents}; p, \gamma_1, \gamma_2) \\ &L_2 = \Pi \Pr(G_{parents} \mid p, \gamma_1, \gamma_2) \\ &L_3 = \Pi \Pr(G_{control} \mid p) \\ &L_4 = \Pi \Pr(G_{unrelated\ cases} \mid p, \gamma_1, \gamma_2) \end{aligned}$$

The first factor of the likelihood models the distribution of case genotypes conditional on the parental genotype distribution. The remaining factors describe the distribution of parental and unrelated CC genotypes respectively, all from the same population (Nagelkerke P approach). In the presence of population stratification, trio-parents have to be excluded from the unrelated cases-parent-control part of the likelihood (Nagelkerke NP approach). Under the TDTCC scenario, the combined likelihood of the Nagelkerke (P) approach corresponds to $L_1L_2L_3L_4$ and the likelihood of the Nagelkerke (NP) approach to $L_1L_3L_4$. Under the TDTC scenario these likelihoods reduce to $L_1L_2L_3$ (P) and L_1L_3 (NP) respectively. All simulations were performed using the SAS system (SAS 9.1 TS1M3; SAS Institute Inc., Cary, N.C., USA).

- Dudbridge Method [5]

A combined allelic RR estimate for CC and trio samples using multinomial regression can be obtained from a retrospective likelihood, which is partitioned into two factors, both modeling the RR separately:

$$\begin{aligned} &Likelihood = L_1 \times L_2 \\ &L_1 = \Pi \Pr(G_{cases} \mid G_{parents}; p, \gamma_1, \gamma_2) \\ &L_2 = \Pi \Pr(G_{parents} \mid Y) \end{aligned}$$

For trio designs, the first factor represents the TDT part of the likelihood i.e. the conditional likelihood of children given parents, which is analogous to a conditional multinomial logistic regression of cases and pseudo-controls [5]. The second part of the likelihood refers to the genotype probabilities of the parents given the affected status of the cases, which allows inference of missing genotypes. This part of the likelihood depends on the assumption of HWE and random mating since the parental genotype frequencies have to be modeled. All inferences are conducted on the estimate of the risk parameter γ_1 in the first part of the likelihood [5].

For CC designs [5], two missing parents are assumed and the likelihood reduces to

$$L_1 = \Pi \Pr(G_{unrelated\ cases} \mid p, \gamma_1, \gamma_2)$$

The estimated risk parameters correspond to the multiplicative model by default, although genotype tests can be carried out optionally. Combined TDTCC analysis can be performed by adding the loglikelihood from both CC and trio samples. A sample indicator variable (confounder option) allows for different MAFs in the two samples, while the RR parameters in the two samples stay the same. This approach can be extended to obtain risk estimates for combined CC and affected sibpair family samples (ASPFCC scenario). Under this scenario, the method accounts for linkage through a conditioning step, which is based on the equivalence class of the inheritance vector [5]. For the combined TDTC analysis however, the approach by Dudbridge will only gain more power than the family-based design alone in the presence of missing genotype data in the parents or uncertain haplotypes, and was thus not further studied. Combined analyses were carried out with the Unphased software (v.3.0.7) (confounder option).

- Epstein Method [6]

This technique is a Conditional-on-Parental-genotypes (CPG)-based approach [19], which allows a combined analysis of unrelated cases, unrelated controls and trios, as well as an analysis of unrelated controls and trios only (extending the work by Nagelkerke and colleagues [4]). The method does not assume HWE or random mating as the likelihood is stratified according to parental mating types μ:

$$Likelihood = L_1 \times L_2 \times L_3 \times L_4$$
$$L_1 = \Pi \Pr(G_{cases} \mid G_{parents}; \gamma_1, \gamma_2, \mu)$$
$$L_2 = \Pi \Pr(G_{parents} \mid \gamma_1, \gamma_2, \mu)$$
$$L_3 = \Pi \Pr(G_{control} \mid \mu)$$
$$L_4 = \Pi \Pr(G_{unrelated\ cases} \mid \gamma_1, \gamma_2, \mu)$$

Formal LRTs are carried out to determine whether data from different sources may be combined. Differences between mating type frequencies between family data and data from unrelated individuals cannot be compared directly (e.g. as mating type frequencies are unknown for the unrelated samples). Instead, the equality of risk parameters from trios (RR_{triads}) and risk parameters derived from comparisons between parents with unrelated controls is tested ($RR_{parents}$). If this test is not rejected, these risk parameters will be set equal [6]. In a subsequent test, the risk parameters from trios (RR_{triads}) are compared for equality with the RR parameters from unrelated CC comparisons ($RR_{case/control}$). If differences in RR parameters in either of the tests are found, combined analysis is not performed. For TDTC designs, only the first test is required, for TDTCC designs both tests have to be carried out.

All power simulations were carried out with the software Scout (v.beta.0.9.5). To allow for comparisons with other methods, the simulated disease models under the TDTCC and TDTC scenario were analysed assuming a multiplicative analysis model. However, the method provides also options for dominant and recessive disease analysis models. To account for the number of rejected joint analyses using the LRT-based homogeneity testing approach above, we also report the 'adjusted' power. Given that the number of non-rejected homogeneity tests exceeded the arbitrary threshold of 100 tests, we calculated adjusted power among those tests as the number of rejected hypotheses of 'no joint association' divided by the total number of joint association tests.

- Allen-Brady Method [9, 10]

Associations in large and complex pedigrees can be assessed non-parametrically by an empirical null genotype distribution conditional on the selected pedigree structure [9]. This distribution can be generated using Mendelian genedrop after random assignment of alleles to pedigree founders in proportion to MAF estimates from the sample [9]. Empirical significance levels for a wide range of test statistics including a combined allelic OR estimate can be obtained by comparing the statistic of interest based on the original phenotype data with the one obtained from the empirical null distribution. Within our study, the null genotype configuration for an allelic OR under a TDTCC scenario and an ASPFCC scenario was investigated using 10,000 simulations. Analyses cannot be performed for TDTC data. All simulated data were investigated using combined meta-OR estimates based on Cochran-Mantel-Haenszel techniques, which can account for potential allelic differences between samples [10]. The OR estimate in single samples however is based on a contingency table, for both family and CC data. Thus, the OR estimate is biased for related samples although accurate significance levels are obtained by empirically deriving the null distribution. Consequently, the derived Meta-OR point-estimates were not assessed for coverage under the TDTCC scenario, as these estimates are also biased and not corrected for relatedness among individuals. All simulations were carried out with the Genie package (v2.6.3) [9, 10] on a 'High Performance Computing' system (http://www.acrc.bris.ac.uk/acrc/hpc.htm). We selected the 'founders' option to ensure that the specified allele frequencies will be correctly represented in the simulated samples when the alleles are assigned through mendelian genedrop.

Simulation of TDTCC and TDTC Scenarios

A single genetic locus was simulated under a multiplicative, dominant and recessive model assuming different MAFs (0.1 or 0.5), varying genetic effect sizes (1 or 2) and no missing genotypes in the individual data sets. For the TDTCC scenario, 500 parent-offspring trios and 500 CC samples were generated; for the TDTC scenario, unrelated cases were omitted. We modeled the genetic effect choosing $\gamma_1 = \gamma_2 = 2$ for the dominant model, $\gamma_1 = \sqrt{2}$ and $\gamma_2 = 2$ for the multiplicative model and $\gamma_1 = 1$ and $\gamma_2 = 2$ for the recessive model [20]. Genotype distributions for CC designs were simulated as outlined by Schaid and Sommer [19] assuming a population disease prevalence $K_p = 0.01$ (see online suppl. table 1, www.karger.com/doi/10.1159/000212503). Parent-trio genotype frequencies were simulated in triads according to different mating types and offspring genotypes, conditional on the offspring being affected [4] ('Likelihood method when HWE holds', [19]) (see online suppl. table 2). The simulated genotypes were randomly chosen from a multinomial uniform distribution based on the genotype or mating type frequencies. The marker MAFs were restricted to be equal in each single sample. For each model and parameter set we created 5000 simulated CC and trio data sets; except for the Allen-Brady approach, where, due to computational limitations, only 1000 different data sets were generated. These data were analysed with the single and combined methods described above. For the TDTCC scenario, Type I error rates were studied on neutral models ($\gamma_1 = \gamma_2 = 1$) with different MAFs in single samples assuming either equal ratios of unrelated cases, controls and trios (500 unrelated cases, 500 controls, 500 trios) or unequal ratios (50 unrelated cases, 500 controls, 100 trios). For the TDTC scenario, the unrelated cases were excluded. Choosing an α-level of 5%, the expected 95% confidence intervals (CI) are (0.036, 0.063) for 1000 simulations and (0.044, 0.056) for 5000 simulations[21].

Simulation of the ASPFCC Scenario

We generated dominant and recessive models under linkage conditions with the disease penetrances φ_{dd}, φ_{dD} and φ_{DD} being constrained to either $\varphi_{dD} = \varphi_{DD} > \varphi_{dd}$ or $\varphi_{DD} > \varphi_{dd} = \varphi_{dD}$ respectively. Phenocopy rates were varied between 0 and 0.5 and the sibling recurrence risk ratio λs fixed at 1.5. The LD between the marker and the disease locus was simulated as r^2 of 0, 0.1 or 0.2 respectively. The spectrum of possible marker MAFs is therefore constrained by the specified LD between marker and disease locus, and the modeled disease allele frequency, which is determined by the phenocopy rate. Simulations were carried out for MAFs between 0.1 and 0.2 under the dominant model and between 0.4 and 0.5 under the recessive model respectively; the marker MAFs were restricted to be equal in the subsamples. Model parameters for all simulated ASPFCC samples are detailed in the online suppl. table 3. Each simulated data set contained 100 affected sibpairs and their parents and 100 CC pairs. For each model specification, 5000 simulated data sets were created (1000 for the Allen-Brady method) and analysed with single and combined analysis techniques.

Results

TDTCC Scenario

We empirically studied the power and Type I error rate of the combined analysis techniques developed by Schaid and Rowland, Nagelkerke (NP and P), Epstein, Kazeem and Farrall, Dudbridge, and Allen-Brady. Combined statistics were compared with each other as well as with single sample statistics and a pooled single sample χ^2-estimate.

- Type I Error Rate

All methods showed consistent Type I error rates of approximately 5% under the studied null and strata models (see table 1), except for the method by Epstein. This method had zero Type I error rate for null and strata models with unequal MAFs in single samples, and approximately 5% Type I error rate under null and strata models with equal MAFs.

- Power Simulations Assuming Similar Effect Sizes in Single Samples

Simulation results for the multiplicative disease model are given in table 2. The power tables for dominant and recessive disease models can be found as online suppl. tables 4 and 5. For all disease models combined sample statistics gave better power than single sample statistics and pooled χ^2-statistics. The combined methods by Schaid and Rowland, Nagelkerke (NP), Kazeem and Farrall, Dudbridge and Allen-Brady had the highest power. The Epstein method had less power than other combined techniques under all disease models, and often also less power than the pooled and single χ^2-statistics, due to incorrectly rejected tests (increased Type II error). The adjusted power for the Epstein method was however comparable to the power of other combined statistics. Including parental trio genotypes in the case-control part of the likelihood using Nagelkerke's (P) method did not improve power compared with the Nagelkerke's (NP) method.

- Simulations Assuming Varying Effect Sizes in Single Samples

Among all methods, the single sample statistic of the associated sample and the pooled χ^2-statistics had the highest power (see table 2, supplementary tables 4 and 5). The power of the Allen-Brady method depended on which sample carried the simulated genetic effect. The method showed higher or equal power, compared to other combined statistics, when the effect was in the CC part of the sample, and lower power when the effect was in the trios part, regardless of the simulated disease model and the simulated MAF. The Epstein method rejected most sample designs with unequal genetic effects in individual samples. The rejection of combined analysis due to unequal parameters between the samples was most sensitive under the dominant model (see online suppl. table 4 MAF: CC 0.1, Trios 0.1). Least rejected designs were multiplicative and recessive models with low MAFs in both single samples (see table 2 and online suppl. table 5 respectively; MAF: CC 0.1, Trios 0.1).

Table 1. Type I error under the TDTCC scenario

Method	MAF			
	CC 0.1 Trios 0.1	Trios 0.1 CC 0.5	Trios 0.5 CC 0.1	Trios 0.5 CC 0.5
Neutral model				
CC-χ^2	0.050	0.047	0.051	0.047
TDT-χ^2	0.053	0.050	0.049	0.042
Pooled-χ^2	0.050	0.047	0.051	0.043
KF	0.053	0.047	0.048	0.047
SR	0.050	0.052	0.046	0.047
N(P)	0.051	NA_1	NA_1	0.048
N(NP)	0.048	0.049	0.049	0.056
D	0.054	0.050	0.052	0.051
E	0.049	<0.001	<0.001	0.047
E*	0.056	NA_2	NA_2	0.051
A	0.055	0.057	0.054	0.050
Strata model				
CC-χ^2	0.046	0.051	0.049	0.052
TDT-χ^2	0.046	0.049	0.048	0.052
Pooled-χ^2	0.046	0.047	0.046	0.052
KF	0.044	0.044	0.047	0.050
SR	0.048	0.049	0.050	0.048
N(P)	0.048	NA_1	NA_1	0.054
N(NP)	0.048	0.047	0.046	0.049
D	0.053	0.045	0.050	0.052
E	0.054	<0.001	<0.001	0.044
E*	0.060	NA_2	NA_2	0.049
A	0.044	0.040	0.043	0.048

MAF = Marker minor allele frequency; neutral = neutral model (equal sample ratios); strata = stratification model (50 founder cases, 500 controls, 100 trios); CC = case-control sample; pooled χ^2 = from single samples; KF = Kazeem & Farrall; SR = Schaid & Rowland; NK = Nagelkerke with (P) and without parental information (NP); D = Dudbridge; E = Epstein; E* = adjusted power of E; NA_1 = not designed for stratified samples; NA_2 = number of joint association tests for $E \leq 100$; A = Allen-Brady (1000 simulations); expected 95% CI for $\alpha = 0.05$: [0.044, 0.056] for 5000 simulations, [0.036, 0.063] for 1000 simulations.

Table 2. Power under the TDTCC scenario – Multiplicative model

MAF	Method	GRR: CC 2 Trios 1			GRR: CC 1 Trios 2			GRR: CC 2 Trios 2		
		0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001
CC 0.1	CC-$\chi^{2\#}$	0.62	0.38	0.16	0.02	0.00	0.00	0.60	0.38	0.16
Trios 0.1	TDT-$\chi^{2\#}$	0.02	0.00	0.00	0.59	0.37	0.16	0.61	0.38	0.15
Comb 0.1	pooled-χ^2	0.61	0.37	0.14	0.60	0.36	0.14	0.89	0.75	0.49
	KF	0.44	0.21	0.06	0.43	0.21	0.06	0.94	0.83	0.59
	SR	0.46	0.23	0.07	0.45	0.23	0.07	0.94	0.83	0.59
	N(P)	0.47	0.24	0.08	0.42	0.21	0.06	0.95	0.84	0.60
	N(NP)	0.44	0.23	0.07	0.43	0.22	0.07	0.95	0.83	0.59
	D	0.45	0.22	0.07	0.45	0.23	0.07	0.94	0.83	0.60
	E	0.12	0.05	0.02	0.12	0.06	0.02	0.84	0.75	0.53
	E*	0.41	0.19	0.07	0.43	0.21	0.06	0.94	0.83	0.59
	A	0.63	0.35	0.12	0.31	0.14	0.03	0.93	0.83	0.60
CC 0.5	CC-$\chi^{2\#}$	0.95	0.86	0.67	0.03	0.01	0.00	0.95	0.86	0.65
Trios 0.5	TDT-$\chi^{2\#}$	0.02	0.01	0.00	0.95	0.85	0.64	0.95	0.86	0.66
Comb 0.5	pooled-χ^2	0.95	0.84	0.62	0.94	0.83	0.60	1.00	0.99	0.97
	KF	0.78	0.57	0.29	0.77	0.54	0.27	1.00	1.00	0.98
	SR	0.77	0.55	0.28	0.78	0.56	0.28	1.00	1.00	0.99
	N(P)	0.77	0.56	0.29	0.78	0.56	0.28	1.00	1.00	0.99
	N(NP)	0.78	0.57	0.29	0.79	0.56	0.28	1.00	1.00	0.99
	D	0.79	0.58	0.29	0.77	0.55	0.28	1.00	1.00	0.99
	E	0.02	0.02	0.01	0.02	0.02	0.01	0.90	0.90	0.89
	E*	0.78	0.57	0.28	0.78	0.57	0.35	1.00	1.00	0.99
	A	0.90	0.77	0.47	0.58	0.31	0.11	1.00	1.00	0.96

MAF = Marker minor allele frequency; GRR = homozygous genotypic relative risk; 0.05/0.01/0.001 = p value thresholds; CC = case-control sample; Comb = combined sample; pooled χ^2 = from single samples; KF = Kazeem & Farrall; SR = Schaid & Rowland; NK = Nagelkerke with (P) and without parental information (NP); D = Dudbridge; E = Epstein; E* = adjusted power of E; NA_1 = not designed for stratified samples; NA_2 = number of joint association tests for E ≤100; A = Allen-Brady (1000 simulations); # corrected for multiple testing.

- Coverage

We also investigated, if an estimate was provided, the coverage of the true genetic effect size under multiplicative disease models with equal MAFs. The methods by Nagelkerke, Kazeem and Farrall, Dudbridge and Epstein, had approximately 95% coverage of the true genetic effect under a multiplicative disease model. The average OR/RR estimates across all simulations were close to the expected value of $\sqrt{2}$ (see online suppl. table 6).

In summary, the combined methods of Schaid and Rowland, Nagelkerke, Kazeem and Farrall, and Dudbridge gave similar power for all disease models and MAFs. The Epstein method gave similar power when the power calculation was based on the number of non-rejected tests.

TDTC Scenario

We investigated power and Type I error rate for the combined methods by Nagelkerke (NP and P) and Epstein as well as the single sample TDT statistic.

- Type I Error Rate

The Type I error for each of the combined methods was approximately 5% under both null and strata models assuming equal MAFs in the subsamples (see table 3). As observed earlier under the TDTCC scenario, the Epstein method had near zero Type I error in the presence of unequal MAFs. Nagelkerke's (NP) method showed consistent type I error of approximately 5% for null and strata models.

Table 3. Type I error under the TDTC scenario

Method	MAF			
	Cn 0.1 Trios 0.1	Cn 0.5 Trios 0.1	Cn 0.1 Trios 0.5	Cn 0.5 Trios 0.05
Neutral model				
TDT-χ^2	0.053	0.050	0.049	0.042
N(P)	0.054	NA_1	NA_1	0.051
N(NP)	0.047	0.049	0.047	0.054
E	0.046	0.001	<0.001	0.050
E*	0.048	NA_2	NA_2	0.053
Strata model				
TDT-χ^2	0.046	0.049	0.048	0.052
N(P)	0.047	NA_1	NA_1	0.051
N(NP)	–**	0.048	0.047	0.049
E	0.047	<0.001	<0.001	0.051
E*	0.049	NA_2	NA_2	0.053

MAF = Marker minor allele frequency; Neutral = neutral model (equal sample ratios); Strata = stratification model (500 controls, 100 trios); Cn = Controls; TDT = allelic TDT-χ^2; NK = Nagelkerke with (P) and without parental information (NP); E = Epstein ; E* = adjusted power of E; NA_1 = not designed for stratified samples; NA_2 = number of joint association tests for E $\leq$ 100; ** Model did not converge; expected 95%CI for α = 0.05: [0.044, 0.056] for 5000 simulations.

- Simulations

Combined statistics under the TDTC scenario gave higher power than the allelic TDT-χ^2 statistic, however Nagelkerke's (P) method showed the best performance (see table 4). The power of the Epstein method was lower due to erroneous detection of differences in RR parameters resulting in failure to perform combined analysis. After adjusting for the number of non-rejected tests, it gave similar power to the Nagelkerke (P) method.

ASPFCC Scenario

Power and Type I error rate were studied for the Dudbridge and the Allen-Brady approach, and single sample statistics.

- Type I Error Rate

Both, the Dudbridge and the Allen-Brady approach, showed consistent Type I error rate under null models with no LD ($r^2 = 0$) between marker and disease locus assuming a recessive disease model. Under some dominant null models the Dudbridge approach displayed marginally inflated type I errors of up to p = 0.061 (see table 5, e.g. dominant model, marker MAF 0.2).

Table 4. Power under TDTC scenario

MAF	Method	GRR: TDT 2								
		Multiplicative			Dominant			Recessive		
		0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001
Cn 0.1		–	–	–	–	–	–	–	–	–
Trios 0.1	TDT-$\chi^{2\#}$	0.59	0.37	0.16	0.99	0.95	0.85	0.06	0.02	0.00
Comb 0.1	N(P)	0.82	0.62	0.34	1.00	1.00	0.98	0.12	0.03	0.01
	N(NP)	0.70	0.45	0.19	1.00	0.98	0.89	0.10	0.03	0.00
	E	0.77	0.59	0.33	0.95	0.95	0.95	0.11	0.03	0.01
	E*	0.81	0.62	0.34	1.00	0.99	0.97	0.12	0.04	0.01
Cn 0.5		–	–	–	–	–	–	–	–	–
Trios 0.5	TDT-$\chi^{2\#}$	0.95	0.85	0.64	0.84	0.66	0.39	0.99	0.94	0.84
Comb 0.5	N(P)	0.99	0.97	0.87	0.97	0.88	0.66	1.00	0.99	0.96
	N(NP)	0.97	0.90	0.71	0.90	0.74	0.45	0.99	0.97	0.88
	E	0.94	0.92	0.84	0.92	0.84	0.63	0.95	0.94	0.91
	E*	0.99	0.97	0.89	0.97	0.89	0.67	1.00	0.99	0.96

MAF = Marker minor allele frequency; GRR = homozygous genotypic relative risk; 0.05/0.01/0.001 = p value thresholds; Cn = controls; Comb = combined sample; TDT = allelic TDT-χ^2; NK = Nagelkerke with (P) and without parental information (NP); E = Epstein; E* = adjusted power of E; NA_1 = not designed for stratified samples; NA_2 = number of joint association tests for E $\leq$ 100; # corrected for multiple testing.

- Simulations

Both combined approaches had more power than single sample statistics when assuming equal marker MAFs in the sample sources and LD between marker and disease locus. However, Allen-Brady's method showed consistently higher power than the Dudbridge approach. This was observed regardless of disease mode, phenocopy rate or the magnitude of marker MAF or LD.

Discussion

This study was conducted in order to empirically investigate the power of combined analysis techniques and their robustness towards confounding factors under different simulated research situations.

Under the TDTCC scenario we found that the four combined statistics by Schaid and Rowland, Nagelkerke, Kazeem and Farrall, and Dudbridge had equivalent power, coverage (if available) and Type I error rates. The similar performance of the likelihood-based approaches by Schaid and Rowland, Nagelkerke (NP) and Dudbridge is to be expected since under the modeled conditions, i.e. one disease locus, non-missing genotypes in HWE and conditioning on parental trio information, their likelihood frameworks are comparable (see Methods).

The equivalent power of the four methods may be explained as follows: Firstly, each of the four methods assesses the association in CC and trio samples independently, similar to a 'meta-analysis' statistic, allowing for sample-specific allele or genotype distributions. This is either provided by sample-specific OR estimates [2], sample-specific scores based on stratum-specific genotype frequencies [3] or an independent factorisation of the likelihood using generalised regression techniques with

Table 5. Power and type I error under the ASPFCC scenario

MAF Λ_S = 1.5	Method	LD: r^2 = 0.0		LD: r^2 = 0.1						LD: r^2 = 0.2					
		PC: 10^{-10}	PC: 0.5	PC: 10^{-10}			PC: 0.5			PC: 10^{-10}			PC: 0.5		
		type I error	type I error	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001
Dominant model															
CC 0.1	CC-$\chi^{2\#}$	0.051	0.054	0.36	0.17	0.06	0.40	0.19	0.06	0.64	0.41	0.18	0.67	0.45	0.20
Ped2A 0.1	Ped2A-$\chi^{2\#}$	0.051	0.059	0.37	0.18	0.06	0.84	0.67	0.40	0.64	0.42	0.18	0.98	0.94	0.80
Comb 0.1	D	0.044	0.059	0.77	0.53	0.25	0.97	0.89	0.70	0.96	0.85	0.62	1.00	1.00	0.97
	A	0.063	0.043	0.99	0.95	0.84	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00
CC 0.2	CC-$\chi^{2\#}$	0.047	0.052	0.43	0.21	0.07	0.44	0.23	0.08	0.71	0.48	0.23	0.75	0.53	0.26
Ped2A 0.2	Ped2A-$\chi^{2\#}$	0.064	0.054	0.41	0.22	0.07	0.92	0.78	0.52	0.73	0.51	0.24	1.00	0.98	0.92
Comb 0.2	D	0.061	0.057	0.81	0.60	0.32	0.98	0.94	0.81	0.98	0.91	0.73	1.00	1.00	0.99
	A	0.041	0.038	0.99	0.96	0.85	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
Recessive model															
CC 0.4	CC-$\chi^{2\#}$	0.051	0.046	0.44	0.25	0.08	0.34	0.17	0.05	0.76	0.56	0.30	0.62	0.41	0.18
Ped2A 0.4	Ped2A-$\chi^{2\#}$	0.055	0.061	0.28	0.13	0.04	0.69	0.47	0.23	0.52	0.31	0.12	0.95	0.86	0.66
Comb 0.4	D	0.054	0.056	0.71	0.48	0.23	0.91	0.77	0.50	0.95	0.84	0.60	1.00	0.98	0.91
	A	0.050	0.059	0.93	0.81	0.54	0.98	0.93	0.78	1.00	0.99	0.92	1.00	1.00	0.99
CC 0.5	CC-$\chi^{2\#}$	0.050	0.052	0.45	0.25	0.09	0.34	0.17	0.05	0.79	0.61	0.32	0.65	0.45	0.20
Ped2A 0.5	Ped2A-$\chi^{2\#}$	0.053	0.047	0.29	0.13	0.04	0.74	0.53	0.28	0.56	0.35	0.13	0.97	0.90	0.73
Comb 0.5	D	0.052	0.054	0.73	0.51	0.25	0.93	0.80	0.55	0.95	0.86	0.65	1.00	0.99	0.94
	A	0.050	0.051	0.94	0.81	0.56	0.99	0.96	0.81	1.00	0.99	0.93	1.00	1.00	1.00

MAF = Marker minor allele frequency; 0.05/0.01/0.001 = p value thresholds; CC = case-control sample; Comb = combined sample; Ped2A = sample of two affecteds and their parents, Ped2A = allelic Ped2A-χ^2 (estimated using D); D = Dudbridge; A = Allen-Brady (1000 simulations); PC = phenocopy rate; # p value for power studies corrected for multiple testing, uncorrected for Type I error analyses; Λ_S = sibling recurrence risk ratio; expected 95%CI for α = 0.05: [0.044, 0.056] for 5000 simulations, [0.036, 0.063] for 1000 simulations.

sample indicator covariates [4, 5]. Thus, each of these techniques will be robust to 'between sample' stratification, which also accounts for Type I error, though potential confounding through CC sample-specific stratification remains. Secondly, each of the techniques assesses the association within the trios part of the combined sample using either a TDT-based approach or an asymptotically equally powerful case-pseudocontrol approach. The Nagelkerke (NP) and the Kazeem & Farrall approach base the assessment of the association in the trios directly on the transmission of alleles from heterozygous parents. Schaid and Rowland's score-statistic in trios is derived from the joint genotype distribution within a family, which is determined by Mendelian probabilities. This is equivalent to a conditional logistic regression approach using pseudo-controls with stratification on families [3]. Similarly, the Dudbridge approach factorises the likelihood for trio designs into a TDT part as conditional likelihood of children given parents, which is analogous to a conditional multinomial logistic regression of cases and pseudo-controls (F. Dudbridge, personal communication), and a second likelihood part, which refers to the genotype probabilities of the parents given the affected status of the cases. It should be noted that the last part of this likelihood assumes HWE and random mating. We did not investigate the robustness of this method to breakdowns in these assumptions.

The power of the Allen-Brady method was only equivalent to other combined statistics e.g. by Schaid and Rowland, Nagelkerke, Kazeem and Farrall, and Dudbridge for models with identical MAFs and genetic effects in single samples. Regardless of the simulated disease model, the method of Allen-Brady showed variable power when the genetic effect differed between the subsamples: The power was higher or equal when the effect came from the CC part of the combined sample and lower when the effect was modeled for the trio part. This may be related to the fact that the OR estimate in the trio subsample is a based on a contingency table, which is underpowered for trio data (see also 'Methods'), whereas other combined approaches include well-established likelihood formulations for trio samples [19, 22, 23].

In our simulations under the TDTCC scenario, we observed furthermore no loss in power using the Nagelkerke (NP) compared to the Nagelkerke (P) method, and for similar methods it was also shown that excluding parental genotype information from the likelihood function does not decrease the power of a test [19]. Under the TDTC scenario however, the power of the Nagelkerke (P) method exceeded or equaled the power of the Nagelkerke (NP) method. This might be related to additional information on MAFs.

All methods gave approximately correct Type I error in the absence of genetic effects in either sample even under extreme differences between MAFs between samples (0.1 vs. 0.5), except that of Epstein. For combined samples with unequal MAFs, this method showed nearly no Type I error. The Epstein method specifically tests for differences in RR parameters between samples, and only performs combined analysis if neither of these tests is rejected. Thus, the method cannot be used to analyse combined TDTCC samples if either the MAF or the strength of the genetic effect differs between the subsamples. Even if genetic effects are the same, the method will loose power relative to other combined methods by falsely detecting differences in RR parameters (according to statistical theory in approximately 5% of all tests). The method is therefore recommended for testing whether the strength of association differs between samples. For non-rejected tests however, and assuming HWE and non-missing genotype information, the likelihood formulation by Epstein reduces to the one presented by Nagelkerke [6]. The most robust method under the TDTC scenario, which provides non-inflated test-statistics under conditions of between-sample stratification, was the Nagelkerke (NP) method.

For the combined analysis of affected siblings and their parents with unrelated CC samples (ASPFCC, scenario 3), the method of Allen-Brady was more powerful than the approach of Dudbridge. The application of a meta-OR statistics incorporated in MC significance testing [9, 10] could model the more complex relationships among families and their combination with CC data better than the parametric likelihood implemented in the Dudbridge approach, which also showed marginally inflated Type1 error under some dominant models.

All combined analysis techniques under the TDTCC scenario had better or equal power (adjusted power for Epstein method) than single sample statistics and the pooled χ^2-statistics, when model parameters were similar between the subsamples. The pooled χ^2-statistics only exceeded other combined statistics in models with different genetic effects in single samples (see e.g. table 2, column 1 and 2 vs. 3).

The combined methods investigated in our study account for differences in risk parameters between single sample sources using either a 'meta-statistic'-like approach [2–5, 9, 10] or test specifically for differences in risk parameters between subsamples [6]. New methods have been proposed however, which claim to be more powerful than 'meta-statistic'-like approaches [8]. How-

ever, these methods control stratification using GC-based techniques, which require the availability of multiple unlinked additional marker genotypes; a condition, which was not investigated in this study.

The primary focus of our work was the study of power as the overall evidence for association using the largest sample available. The question of whether there is significant heterogeneity in risk parameters such as the magnitude of the genetic effects was secondary. However, this issue is crucial once association has been found and requires interpretation. In such a situation, careful analysis and comparison of the risk parameters in the subsamples is important. This may involve an assessment of the sample homogeneity using the Epstein approach [6] or other available techniques such as developed by Kazeem and Farrall [2], as well as comparisons between combined and pooled χ^2-statistics.

For our power simulations under the TDTCC and TDTC scenario we chose a homozygous genotypic relative risk of 2, corresponding to an allelic RR of 1.41 under a multiplicative model. These are strong, but genetically plausible, assumptions for the genetic effect size. The most recent meta-analyses on genome-wide association studies on complex disease [e.g. 24] reported genetic risk effects (allelic odds ratios) within the range of 1.06 to 1.52. The advantages of our parameter selection are that all investigated methods will reach adequate power to compare their performance with each other in relatively small samples, thus ensuring computational feasibility. It is likely that similar conclusions regarding the relative power of the various methods would be reached with smaller relative risks and larger sample sizes.

For the ASPFCC scenario we modeled linkage in families by choosing a sibling recurrence risk ratio λ_S of 1.5. This is quite a strong effect, but not unreasonable, since similar and stronger effects have been observed for complex traits (e.g. the APP locus in late-onset Alzheimers disease showed $\lambda s > 3$ for ASPs with mean age >95 [25], and an affected sib pair linkage scan of Type I diabetes showed $\lambda s = 3.1$ at IDDM1 and $\lambda s = 1.3$ at IDDM2 [26]). The use of a relatively strong linkage model allowed us to assess whether combined methods correct for linkage adequately as the presence of unaccounted linkage is known to increase the false-positive rate of family-based association studies [17]. Under weaker linkage models, methods would give approximately correct Type I error rates, regardless of whether they correctly allow for linkage.

Our study specifically addressed 'between-sample' stratification (i.e. different minor allele frequencies in the case-control and TDT samples). However, population stratification within samples must also be considered. Stratification within the TDT sample will have no effect, since the TDT is robust to stratification. Stratification within the case-control sample will increase Type I error rate, and this will invalidate all methods that perform analysis on case-control data. Thus, we did not consider within sample stratification in our simulations, since our purpose was to compare the performance of analysis methods in situations where they are valid. However, careful correction for population stratification is essential when analysing case-control data (regardless of the method used). If sufficient markers have been typed (e.g. as part of a genome-wide association analysis), methods such as EIGENSTRAT [27] can be used. Otherwise, homogeneity of genetic effects can be tested between the samples [2, 6] although significant heterogeneity in risk parameters does not necessarily imply stratification.

In conclusion, our study investigated the power of up to six different combined analysis techniques under three different research designs. Under the TDTCC scenario we identified four statistical approaches [2–5], which had equivalent power and Type I error rates. All combined test statistics were more powerful than single sample statistics or a pooled χ^2-statistic when MAFs and genetic effects were similar in single subsamples. Adding parental information to the joint likelihood increased the power of a generalised logistic regression under the TDTC but not the TDTCC scenario [4]. Formal testing of differences between risk parameters in single samples was the most sensitive approach to avoid confounding in combined analysis under both the TDTCC and TDTC scenario. Non-parametric analysis based on MC testing [9, 10] gave the highest power under the ASPFCC scenario.

Acknowledgements

We would like to thank F. Dudbridge, G.R. Kazeem, N. Nagelkerke, D. Schaid, K. Allen-Brady and M. Epstein for helpful advice and explanations of their methods.

References

1 Freimer N, Sabatti C: The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. Nat Genet 2004;36:1045–1051.

2 Kazeem GR, Farrall M: Integrating case-control and TDT studies. Ann Hum Genet 2005;69:329–335.

3 Schaid DJ, Rowland C: Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. Am J Hum Genet 1998;63:1492–1506.

4 Nagelkerke NJD, Hoebee B, Teunis P, Kimman TG: Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. Eur J Hum Genet 2004;12:964–970.
5 Dudbridge F: Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Hum Hered 2008;66:87–98.
6 Epstein MP, Veal CD, Trembath RC, Barker JNWN, Li C, Satten GA: Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet. 2005;76:592–608.
7 Li M, Boehnke M, Abecasis GR: Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. Am J Hum Genet 2006;78:778–792.
8 Zhu X, Li S, Cooper RS, Elston RC: A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet 2008;82:352–365.
9 Allen-Brady K, Wong J, Camp NJ: PedGenie: an analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. BMC Bioinformatics 2006;7:209.
10 Curtin K, Wong J, Allen-Brady K, Camp NJ: PedGenie: meta genetic association testing in mixed family and case-control designs. BMC Bioinformatics 2007;8:448.
11 Becker T, Knapp M: Maximum-likelihood estimation of haplotype frequencies in nuclear families. Genet Epidemiol 2004;27:21–32.
12 Becker T, Knapp M: Impact of missing genotype data on Monte-Carlo simulation based haplotype analysis. Hum Hered 2005;59:185–189.
13 Guo W, Fung WK: Combining the case-control methodology with the small size transmission/disequilibrium test for multiallelic markers. Eur J Hum Genet 2005;13:1007–1012.
14 Nicodemus KK: Catmap: case-control and TDT meta-analysis package. BMC Bioinformatics 2008;9:130.
15 Holmans P: Affected sib-pair methods for detecting linkage to dichotomous traits: review of the methodology. Hum Biol 1998;70:1025–1040.
16 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–516.
17 Martin ER, Bass MP, Hauser ER, Kaplan NL: Accounting for linkage in family-based tests of association with missing parental genotypes. Am J Hum Genet 2003;73:1016–1026.
18 Sasieni PD: From genotypes to genes: doubling the sample size. Biometrics 1997;53:1253–1261.
19 Schaid DJ, Sommer SS: Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet 1993;53:1114–1126.
20 Camp NJ: Genomewide transmission/disequilibrium testing – consideration of the genotypic relative risks at disease loci. Am J Hum Genet 1997;61:1424–1430.
21 Manly B: Randomization, Bootstrap and Monte Carlo Methods in Biology. London, Chapman & Hall, 2001.
22 Schaid DJ, Sommer SS: Comparison of statistics for candidate-gene association studies using cases and parents. Am J Hum Genet 1994;55:402–409.
23 Schaid DJ: Transmission disequilibrium, family controls, and great expectations. Am J Hum Genet 1998;63:935–941.
24 Allen NC, Bagade S, McQueen MB, Ioannidis JPA, Kavvoura FK, Khoury MJ, Tanzi RE, Bertram L: Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. Nat Genet 2008;40:827–834.
25 Olson JM, Goddard KAB, Dudek DM: The amyloid precursor protein locus and very-late-onset Alzheimer disease. Am J Hum Genet 2001;69:895–899.
26 Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SC, Jenkins SC, et al: A genome-wide search for human type 1 diabetes susceptibility genes. Nature 1994;371:130–136.
27 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–909.