Park et al. BMC Medical Genetics (2015) 16:62
DOI 10.1186/s12881-015-0198-6

BMC Medical Genetics

RESEARCH ARTICLE Open Access

Adjusting heterogeneous ascertainment bias for genetic association analysis with extended families

Suyeon Park[1,2,3], Sungyoung Lee[4], Young Lee[1,2], Christine Herold[5,6], Basavaraj Hooli[7], Kristina Mullin[7], Taesung Park[8], Changsoon Park[1], Lars Bertram[7,9,10], Christoph Lange[6,11,12,13], Rudolph Tanzi[7*] and Sungho Won[14,15,16*]

Abstract

Background: In family-based association analysis, each family is typically ascertained from a single proband, which renders the effects of ascertainment bias heterogeneous among family members. This is contrary to case–control studies, and may introduce sample or ascertainment bias. Statistical efficiency is affected by ascertainment bias, and careful adjustment can lead to substantial improvements in statistical power. However, genetic association analysis has often been conducted using family-based designs, without addressing the fact that each proband in a family has had a great influence on the probability for each family member to be affected.

Method: We propose a powerful and efficient statistic for genetic association analysis that considered the heterogeneity of ascertainment bias among family members, under the assumption that both prevalence and heritability of disease are available. With extensive simulation studies, we showed that the proposed method performed better than the existing methods, particularly for diseases with large heritability.

Results: We applied the proposed method to the genome-wide association analysis of Alzheimer's disease. Four significant associations with the proposed method were found.

Conclusion: Our significant findings illustrated the practical importance of this new analysis method.

Keywords: Family-based association analysis, Ascertainment, Liability model

Background

Genome-wide association studies (GWASs) have been used to identify many genes involved in human diseases, and during the last decade, many disease-susceptibility variants have been identified. However, despite these successes, we have found that variants discovered from GWASs often explain only a small proportion of the heritability of diseases [1, 2]. For example, SNPs significantly associated with human height explain only about 5 % of phenotypic variance, despite studies of tens of thousands of people [3]. Many reasons, such as rare causal variants and gene/gene interactions, have been attributed to this so-called "missing heritability". However, the low power induced by the multiple-testing problem is still an intractable issue in GWASs, and further investigations of the most efficient strategies for genetic association analysis are necessary.

Careful selection of samples based on phenotypes can lead to improved power for the discovery of risk variants [4–11]. One such example is the extreme discordant sib-pair design in linkage analysis, which may result in a substantial increase in statistical power when compared to other sib-pair designs [11, 12]. Similarly, ascertaining the extremes of quantitative phenotypes from large population cohorts has also been shown to increase the power to identify associated variants [13–15]. In such a design, the effect of ascertainment conditions are homogeneous between individuals, and existing methods, such as the Cochran-Armitage(CA) trend test [16], can be an efficient choice. However, in association analysis using extended

* Correspondence: rtanzi@mgh.harvard.edu; won1@snu.ac.kr
[7]Genetics and Aging Research Unit, MassGeneral Institute for Neurodegenerative Diseases, Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Massachusetts, USA
[14]Department of Public Health Science, Seoul National University, Seoul, Korea
Full list of author information is available at the end of the article

families, the effects of ascertainment bias are often heterogeneous among family members, and depending on their relationships with probands, different magnitudes of ascertainment bias may be generated. In particular, the probability of each individual being affected when his or her relatives are affected is similar to the prevalence, if the heritability is small, which indicates that the heterogeneous effect of the ascertainment bias depends on the magnitude of heritability. However, the heterogeneous effects of ascertainment conditions and the influence of heritability on it have not yet been investigated, and should therefore be taken into account for association analysis.

Recently, the CA trend test was extended for association analysis of dichotomous phenotypes with family-based samples [17, 18]. These statistics compares the genotype frequencies between affected and unaffected individuals, and the genetic association with family-based samples is tested by building a genotype correlation matrix with either kinship coefficients or an empirical correlation matrix estimated from large-scale genetic data. This approach has been extended to include family members with known phenotypes and missing genotypes or *vice versa*. By the nature of these statistics, it performs well for ascertained family-based samples and it can be an efficient choice, even for a case–control design, if the relatives' phenotype information is available. However, their statistical efficiency is affected by the heterogeneous effect of the ascertainment bias on family members, and for extended families, its effects on statistical efficiency can be substantial.

In this report, we consider the heterogeneous effects of the ascertainment bias on family members for dichotomous phenotypes. By the nature of the proposed methods, individuals with missing genotypes and non-missing phenotypes can be utilized, and incorporation of the estimated kinship matrix to the proposed statistic provided robustness against the population substructure. The proposed method consists of two steps; the probability for each family member to be affected was calculated using a latent continuous liability [19], and then this probability is incorporated into a quasi-likelihood score test. With an extensive simulation, we showed that the proposed method performed better than the existing methods, particularly for a disease with large heritability. Application of our method to Alzheimer's disease (AD) demonstrated its practical use in the detection of genetic associations in ascertained family-based samples.

Methods

Notations and statistic

We assumed that there were n families and n_i family members in each family. We considered the situation where the family of size n_i was ascertained because it contained a particular set of p_i members, and we let $q_i = n_i - p_i$. We called the members of the set of p_i family members "probands", and the remaining q_i individuals "non-probands". To provide a clearer motivation on this concept, we randomly selected two families, family 1 and 2, from our AD data (see Fig. 1). In family 1 (Fig. 1-(a)), individual 9 was diagnosed as AD and individuals 3–8 were selected as her relatives for genetic analysis. In family 2 (Fig. 1-(b)), individual 3 was diagnosed as AD, and individuals 4–6 were selected. Therefore $p_1 = p_2 = 1$, $q_1 = 6$ and $q_2 = 3$ in this example. In real data analysis, p_i is often 1 and $q_i = n_i - 1$. We assumed that N individuals were available and thus $N = \sum_i n_i$. The genotypes were coded as 0, 1, or 2, according to the number of disease alleles. x_{ij}^P and $x_{i'j'}^N$ were defined as the genotypes of proband j and non-proband j' in family i and family i', respectively. Phenotypes were coded as 0 for an unaffected individual and 1 for an affected individual. If we let the prevalence of the disease be p, a missing phenotype was coded as p. We denoted the phenotypes of a proband and non-proband by y_{ij}^P and $y_{i'j'}^N$, respectively, and the vectors for genotypes and phenotypes in family i were defined by

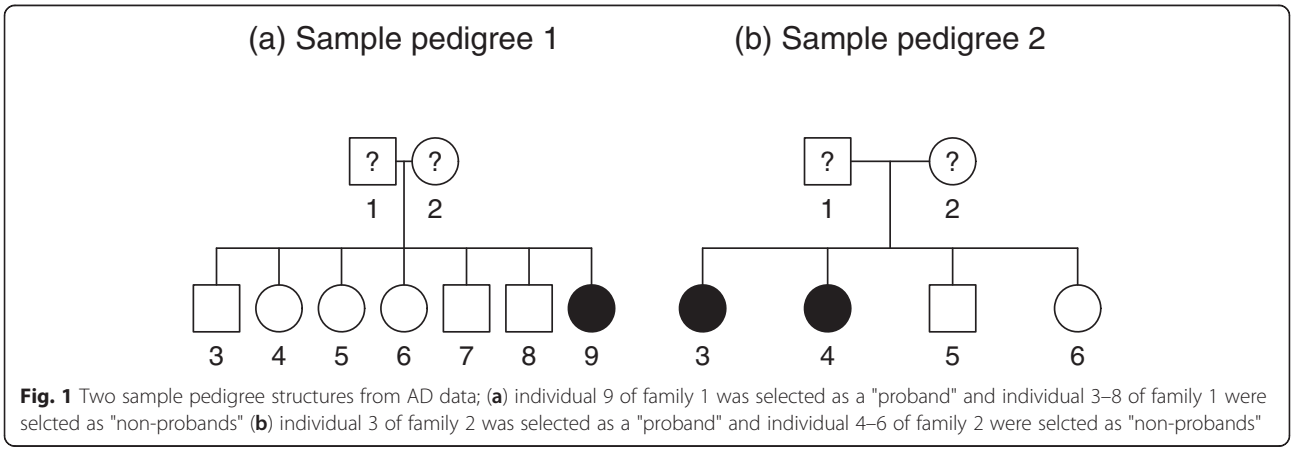

Fig. 1 Two sample pedigree structures from AD data; (**a**) individual 9 of family 1 was selected as a "proband" and individual 3–8 of family 1 were selcted as "non-probands" (**b**) individual 3 of family 2 was selected as a "proband" and individual 4–6 of family 2 were selcted as "non-probands"

$$\mathbf{X}_i^P = \begin{pmatrix} x_{i1}^P \\ x_{i2}^P \\ \vdots \\ x_{ip_i}^P \end{pmatrix},\ \mathbf{X}_i^N = \begin{pmatrix} x_{i1}^N \\ x_{i2}^N \\ \vdots \\ x_{iq_i}^N \end{pmatrix},\ \mathbf{X}_i = \begin{pmatrix} \mathbf{X}_i^P \\ \mathbf{X}_i^N \end{pmatrix},$$

$$\mathbf{Y}_i^P = \begin{pmatrix} y_{i1}^P \\ y_{i2}^P \\ \vdots \\ y_{ip_i}^P \end{pmatrix},\ \mathbf{Y}_i^N = \begin{pmatrix} y_{i1}^N \\ y_{i2}^N \\ \vdots \\ y_{iq_i}^N \end{pmatrix},\ \text{and}\ \mathbf{Y}_i = \begin{pmatrix} \mathbf{Y}_i^P \\ \mathbf{Y}_i^N \end{pmatrix}$$

We also denoted the $w \times w$ identity matrix by $\mathbf{I}_w$, and the $w \times 1$ column vector $\mathbf{1}_w$ indicated a vector in which all elements were 1. Let $\pi_{ijj'}^P$ and $\pi_{ijj'}^N$ be the kinship coefficient between probands j and j' in family i, and non-proband j and j' in family i, respectively. In addition, we let $\pi_{ijj'}^{PN}$ be the kinship coefficient between proband j and non-proband j' in family i, and let d_{ij}^P and $d_{ij'}^N$ be the inbreeding coefficient for proband j and non-proband j' in family i, respectively. The inbreeding coefficient is the parameter that quantifies the departure from Hardy-Weinberg equilibrium (HWE) and ranges from 0 to 1. Several approaches [20, 21] that can estimate d_{ij} have been proposed. We let

$$\mathbf{R}_i^P = \begin{pmatrix} 1+d_{i1}^P & \cdots & 2\pi_{i1p_i}^P \\ \vdots & \ddots & \vdots \\ 2\pi_{ip_i1}^P & \ldots & 1+d_{ip_i}^P \end{pmatrix},$$

$$\mathbf{R}_i^N = \begin{pmatrix} 1+d_{i1}^N & \cdots & 2\pi_{i1q_i}^N \\ \vdots & \ddots & \vdots \\ 2\pi_{iq_i1}^N & \ldots & 1+d_{iq_i}^N \end{pmatrix},$$

$$\mathbf{R}_i^{PN} = \begin{pmatrix} 2\pi_{i11}^{PN} & \cdots & 2\pi_{i1q_i}^{PN} \\ \vdots & \ddots & \vdots \\ 2\pi_{ip_i1}^{PN} & \ldots & 2\pi_{ip_iq_i}^{PN} \end{pmatrix},$$

and $\mathbf{R}_i$ is defined by

$$\mathbf{R}_i = \begin{pmatrix} \mathbf{R}_i^P & \mathbf{R}_i^{PN} \\ (\mathbf{R}_i^{PN})^t & \mathbf{R}_i^N \end{pmatrix}$$

If we let q_A be the disease allele frequency, $E(\mathbf{X}_i)$ was $2q_A 1_{n_i}$, and q_A is estimated with the best linear unbiased estimator (BLUE). $\mathrm{var}(\mathbf{X}_i)$ is expressed by $\sigma^2\mathbf{R}_i$, and σ^2 is equal to $2q_A(1-q_A)$ under HWE.

When we analyzes the distribution of genotypes as in the FBAT approach, the statistical efficiency of the test statistic could be improved by adjustments of the phenotype with the so-called offset [22]. If we let μ_{ij}^P and $\mu_{ij'}^N$ be offsets for proband j and non-proband j' in family i and family i', respectively, the offset vector for family i is defined as

$$\mu_i^P = \begin{pmatrix} \mu_{i1}^P \\ \mu_{i2}^P \\ \vdots \\ \mu_{ip_i}^P \end{pmatrix},\ \mu_i^N = \begin{pmatrix} \mu_{i1}^N \\ \mu_{i2}^N \\ \vdots \\ \mu_{iq_i}^N \end{pmatrix},\ \text{and}\ \mu_i = \begin{pmatrix} \mu_i^P \\ \mu_i^N \end{pmatrix}$$

Setting $\mathbf{T}_i = \mathbf{Y}_i - \boldsymbol{\mu}_i$, we can define

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \end{pmatrix},\ \mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \end{pmatrix},\ \mathbf{T} = \begin{pmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \\ \vdots \end{pmatrix},\ \text{and}$$

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & 0 & \cdots \\ 0 & \mathbf{R}_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

We denoted a minor allele frequency (MAF) of a variant in unaffected individuals by q. We assumed [18] that for a constant γ,

$$E(\mathbf{X}|\mathbf{T}) = 2p\mathbf{1}_N + \gamma\mathbf{T},$$

where $0 < 2p + \gamma < 1$. Then, the score for a variant [18, 23] can be defined by

$$S = \mathbf{T}^t(\mathbf{X} - \hat{E}(\mathbf{X})) \ \text{ and } \ \hat{E}(\mathbf{X}) = \mathbf{1}_N(\mathbf{1}_N^t\mathbf{R}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t\mathbf{R}^{-1}\mathbf{X}.$$

The variance of S is

$$\mathrm{var}(S) = \sigma^2\mathbf{T}^t\mathbf{V}^{-1}(\mathbf{R} - \mathbf{1}_N(\mathbf{1}_N^t\mathbf{R}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t)\mathbf{V}^{-1}\mathbf{T},$$

and we considered the following statistic [17, 18]:

$$\frac{\mathbf{T}^t(\mathbf{I}_N - (\mathbf{1}_N^t\mathbf{R}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t\mathbf{R}^{-1})\mathbf{X}}{\sqrt{\sigma^2\mathbf{T}^t(\mathbf{R} - \mathbf{1}_N(\mathbf{1}_N^t\mathbf{R}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t)\mathbf{T}}} \sim N(0,1) \text{ if } \gamma = 0.$$

This statistic will be denoted by *WL* in the remainder of this report.

Adjusting the heterogeneous ascertainment bias

Families are often selected based on some probands, and the probability for family members to be affected depends on their relationship with the probands. Additional file 1 shows that the incorporation of conditional probability of each individual being affected to *WL* as offset lead to asymptotically smaller variance and therefore the adjustment of heterogeneous ascertainment bias is required to improve the statistical power of *WL*. This probability could be estimated with the liability model if the heritabilities, h^2, and prevalence, p, were available. We let l_{ij}^P and $l_{i'j'}^N$ be the liability of proband j and non-proband j' in family i and family i', respectively, and let $\mathbf{L}_i^P = (l_{i1}^P, \ldots, l_{ip_i}^P)$ and $\mathbf{L}_i^N = (l_{i1}^N, \ldots, l_{iq_i}^N)$. We assumed that each liability

followed the standard normal distribution, and their joint distributions were

$$\begin{pmatrix} L_i^P \\ L_i^N \end{pmatrix} \sim MVN\left(0, h^2\begin{pmatrix} \mathbf{R}_i^P & \mathbf{R}_i^{PN} \\ (\mathbf{R}_i^{PN})^t & \mathbf{R}_i^N \end{pmatrix} + (1-h^2)\mathbf{I}_{n_i}\right).$$

Benchek and Morris [24] reported that significant asymptotic biases are likely to arise when the multivariate normal (MVN) liability assumption is not met and in such a case, different assumptions should be considered. We assume that $\mathbf{M}_i^{P^*}$ and $\mathbf{V}_i^{P^*}$ are the expectation and variances of L_i^P when their disease statuses are conditioned. If all probands are affected, they becomes

$$\mathbf{M}_i{}^{P^*} \equiv E(L_i{}^P | l_{i1}{}^P > c, \ldots, l_{ip_i}{}^P > c)$$

and

$$\mathbf{V}_i{}^{P^*} \equiv \mathrm{var}(L_i{}^P | l_{i1}{}^P > c, \ldots, l_{ip_i}{}^P > c).$$

They can be calculated with the numerical algorithms [25]. If p_i is 1, both can be simply calculated. We denote the cumulative and probability density function of standard normal distribution by $\Phi(\cdot)$ and $\phi(\cdot)$. If we let c be the $(1-p)$th quantile of the standard normal distribution, $\mathbf{M}_i^{P^*}$ and $\mathbf{V}_i^{P^*}$ becomes

$$\mathbf{M}_i^{P*}\begin{cases} \phi(c)/[1-\Phi(x)] & \text{if } y_{i1}^P = 1 \\ -\phi(c)/\Phi(x) & \text{if } y_{i1}^P = 0 \end{cases}, \quad \text{and}$$

$$\mathbf{V}_i^{P*} = 1-(\mathbf{M}_i^{P*})^2 + \mathbf{M}_i^{P*}c.$$

With Pearson-Aitken formula [26, 27], we could obtain the conditional mean and variance-covariance matrix of $\mathbf{L}_i^N$ given $\mathbf{L}_i^P > 1_{p_i} \cdot c$ as follows:

$$\mathbf{M}_i^{N*} = 0 + (\mathbf{R}_i^{PN})^t(\mathbf{V}_i^P)^{-1}(\mathbf{M}_i^{P*}-0)$$

and

$$\mathbf{V}_i^{N*} = \mathbf{V}_i^N + (\mathbf{R}_i^{PN})^t((\mathbf{V}_i^P)^{-1} - (\mathbf{V}_i^P)^{-1}\mathbf{V}_i^{P*}(\mathbf{V}_i^P)^{-1})\mathbf{R}_i^{PN}.$$

We denoted the jth element in $\mathbf{M}_i^{N^*}$ by $m_j^{N^*}$ and the jth diagonal element in $\mathbf{V}_i^{N^*}$ by $v_j^{N^*}$. Then the probability of being affected for a non-proband under multivariate normality of the liabilities could be calculated as

$$\Phi\left(\frac{c-m_j^{N*}}{v_j^{N*}}\right),$$

and this will be incorporated into the proposed statistic as offset. Thus far, we have assumed that there was a well-designed set of p_i individuals who were "probands", and for this situation, we calculated the statistic as indicated and denoted $FQLS_1$. However in practice, different ascertainment condition such as sequential sampling frame [28] are often utilized, and the set of p_i individuals will not be well defined. For this situation, we calculated the probability for each individual to be affected under the assumption that all the other family members were "probands", and thus $p_i = n_i - 1$ and $q_i = 1$. The statistic calculated this way was denoted by $FQLS_2$.

Results

The simulation model

In our simulation studies, we considered two types of family structures; nuclear families with five offspring and the extended families that consist of 13 individuals along 3 generations (see Fig. 2). The latter will be called extended families in the remainder of this report. The disease allele frequency, p, was assumed to be 0.2. If we denoted the disease allele frequency by q_A, the genotype frequencies for AA, Aa, and aa became q_A^2, $2q_A(1-q_A)$, and $(1-q_A)^2$ under HWE, respectively, and founders' genotypes were generated under the corresponding multinomial distribution. The genotypes for non-founders were generated with randomly generated Mendelian transmission. The disease status was generated with the liability

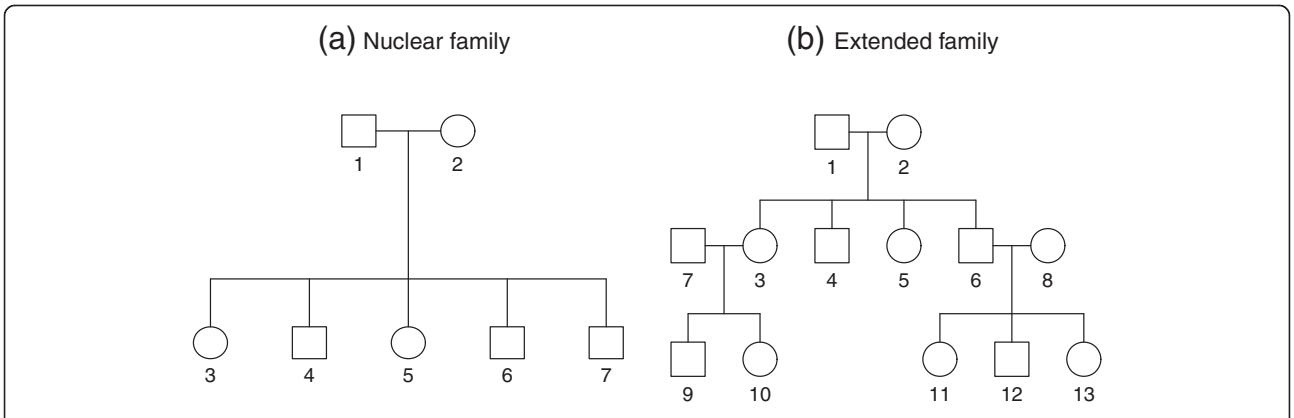

Fig. 2 Family tree. There are two different types of family structures, including **(a)** nuclear family and **(b)** extended family, which were considered in our simulation study

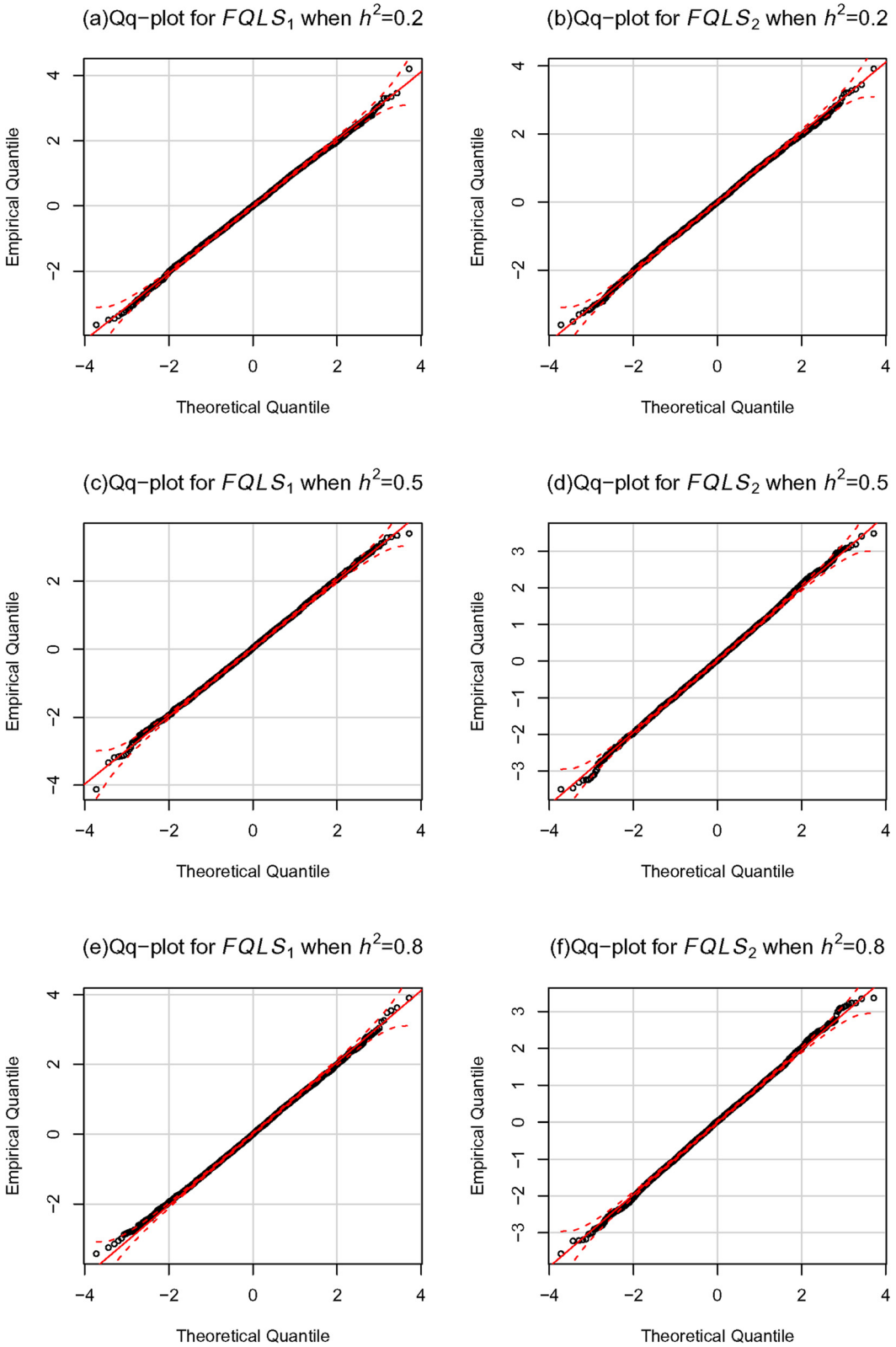

Fig. 3 QQ plots for $FQLS_1$ and $FQLS_2$ under the null hypothesis. QQplots for $FQLS_1$ and $FQLS_2$ are obtaind when h2 is 0.2((**a**), (**b**)), 0.5((**c**), (**d**)), or 0.8((**e**), (**f**)). P-values were calculated based on 5000 replicates when the number of families was 900. The genetic effect β was assumed to be 0, and the minor allele frequency was 0.2

threshold model. Once continuous liabilities that consisted of polygenic effects and random errors were generated, they were transformed to being affected if they were larger than the threshold; and otherwise, they were considered to be unaffected. The threshold was chosen to preserve the prevalence, and prevalence was assumed to be 0.2. Continuous liability was determined by combining the phenotypic mean, polygenic effect, main genetic effect, and random error. The main genetic effect for each individual was the product of β and the number of disease alleles. If we denoted the relative proportion of the phenotypic variance attributable to the main disease gene by h_a^2, and h^2 was a heritability for continuous liability, β was calculated by

$$\beta = \sqrt{\frac{h_a^2}{2q_A(1-q_A)(1-h^2)}}.$$

For the evaluation of type-1 errors and power, h_a^2 was assumed to be 0 and 0.005, respectively. Phenotypic correlations between family-members were explained by the polygenic effects. Parental polygenic effects were generated from $N(0,\ h^2)$, and h^2 was assumed to be 0.2, 0.5, or 0.8. For non-founders, the average of maternal and paternal polygenic effects was combined with the values independently sampled from $N(0,\ 0.5\,h^2)$ for the polygenic effects of offspring. Random errors were generated from $N(0,\ \sigma_e^2 = 1-h^2)$. For each replicate, sampling was repeated until a given number of ascertained families was generated. Type-1 error estimates were calculated with 5000 replicates, and empirical power estimates were calculated with 1000 replicates.

Table 1 Empirical type-1 error estimates. The empirical type-1 error rates and their 95 % confidence intervals were estimated with 5000 replicates at the 0.01 and 0.05 significance level for $h^2 = 0.2$, 0.5, and 0.8. The number of families was assumed to be 900, and the disease allele frequency was 0.2

	h^2	Statistics	Type-1 error estimates	95 % confidence interval	
				Lower	Upper
0.01	0.2	$FQLS_1$	0.011	0.008	0.014
		$FQLS_2$	0.011	0.008	0.013
	0.5	$FQLS_1$	0.010	0.007	0.013
		$FQLS_2$	0.009	0.007	0.012
	0.8	$FQLS_1$	0.009	0.006	0.011
		$FQLS_2$	0.009	0.007	0.012
0.05	0.2	$FQLS_1$	0.049	0.043	0.055
		$FQLS_2$	0.049	0.043	0.055
	0.5	$FQLS_1$	0.050	0.044	0.056
		$FQLS_2$	0.053	0.047	0.059
	0.8	$FQLS_1$	0.047	0.042	0.053
		$FQLS_2$	0.053	0.047	0.059

Evaluation of the proposed methods with simulated data

The empirical type-1 errors for $FQLS_1$ and $FQLS_2$ were evaluated from 5000 replicates under the situation of no association ($h_a^2 = 0$), and 900 nuclear families with five offspring in Fig. 2 were generated for each replicate. Fig. 3 shows the quantile quantile (QQ) plots from 5000 replicates, and the nominal significance levels for both methods were preserved for various significance levels. We also estimated the empirical type-1 error rates at the 0.01 and 0.05 significance levels; the empirical type-1 error estimates of $FQLS_1$ and $FQLS_2$ preserved these nominal significance levels (Table 1). These results verified that the use of the approximation to the standard normal distribution resulted in an accurate assessment of significance for the proposed methods.

Table 2 Empirical power estimates for scenario 1 when h^2 is 0.2. The empirical power estimates for scenario 1 were calculated with 1000 replicates at the both 0.01 and 0.001 significance levels. The disease allele frequency was assumed to be 0.2, and the prevalence was assumed to 0.2. The relative phenotypic variance attributable to the main disease gene was assumed to be 0.005

	$n_{proband}$	Statistic	N					
			100	300	600	900	1200	1400
0.01	1	*WL*	0.027	0.062	0.129	0.219	0.293	0.369
		$FQLS_1$	**0.029**	**0.064**	0.122	0.220	0.304	0.372
		$FQLS_2$	**0.029**	0.059	**0.129**	**0.230**	**0.306**	**0.385**
	2	*WL*	0.033	**0.076**	0.165	0.295	0.415	0.456
		$FQLS_1$	**0.036**	0.073	0.166	0.309	0.418	0.461
		$FQLS_2$	0.035	**0.076**	**0.178**	**0.312**	**0.422**	**0.465**
	3	*WL*	**0.040**	0.103	**0.257**	0.398	0.519	0.609
		$FQLS_1$	0.039	**0.112**	**0.257**	**0.404**	0.526	**0.623**
		$FQLS_2$	0.036	0.110	0.253	0.407	**0.527**	0.621
	4	*WL*	0.041	0.127	0.297	0.497	0.626	0.720
		$FQLS_1$	0.044	**0.129**	**0.307**	**0.502**	0.626	0.719
		$FQLS_2$	**0.046**	0.126	**0.307**	0.493	**0.637**	**0.726**
0.001	1	*WL*	**0.005**	**0.012**	**0.038**	0.075	0.113	0.153
		$FQLS_1$	**0.005**	**0.012**	**0.038**	**0.076**	**0.115**	0.157
		$FQLS_2$	0.003	0.010	**0.038**	**0.076**	0.106	**0.163**
	2	*WL*	0.006	**0.019**	0.060	0.098	0.193	0.224
		$FQLS_1$	**0.007**	0.018	0.059	0.099	**0.199**	0.217
		$FQLS_2$	0.006	**0.019**	**0.064**	**0.100**	0.197	**0.219**
	3	*WL*	0.004	0.018	0.086	0.164	0.267	0.337
		$FQLS_1$	0.007	0.020	**0.091**	0.162	**0.275**	0.333
		$FQLS_2$	**0.008**	**0.023**	0.087	**0.165**	**0.275**	**0.348**
	4	*WL*	**0.010**	0.029	0.116	0.231	0.309	0.451
		$FQLS_1$	0.009	0.029	0.116	0.228	**0.370**	0.449
		$FQLS_2$	0.009	**0.031**	**0.118**	**0.233**	0.363	**0.459**

The bold text indicates the highest empirical estimate of the power for each situation

The empirical powers at the various significance levels were measured based on 1000 replicates at the 0.01 and 0.001 significance levels. The relative proportion, h_a^2, of phenotypic variance attributable to the main disease gene, $2p_A(1-p_A)\beta^2$, was assumed to be 0.005, and nuclear and extended families in Fig. 2 were considered for the power comparison. In the first simulation setting, the numbers of nuclear families were assumed to be 100, 300, 600, 900, 1200, and 1400, and half of the families were ascertained if the number of affected family members was larger than or equal to $n_{proband}$, and the other half of the families were ascertained if the number of unaffected family members was larger than or equal to $n_{proband}$. Therefore, if 100 nuclear families were generated, half of nuclear families should have more than or equal to $n_{proband}$ affected family members, and the other half should have at least $n_{proband}$ unaffected family members. We assumed that the heritabilities were 0.2, 0.5, and 0.8, and results are shown in Tables 2, 3, 4, respectively. In the second simulation setting, the numbers of extended families were assumed to be 100, 300, 600, and 900, and all families were ascertained if the number of affected amily members was larger than or equal to $n_{proband}$. Empirical power estimates for scenario 2 were calculated when $h^2 = 0.2$, 0.5, and 0.8, and the data are shown in Tables 5, 6, 7, respectively. Our results showed that either $FQLS_1$ or $FQLS_2$ was usually the most efficient statistic, and the least efficiency was provided from *WL*. In particular, the power gap between the proposed methods and *WL* was largest if h^2 was 0.8, which indicates that power improvement may be proportional to the heritability.

Table 3 Empirical power estimates for scenario 1 when h^2 is 0.5. The empirical power estimates for scenario 1 were calculated with 1000 replicates at the both 0.1 and 0.001 significance levels. The disease allele frequency was assumed to be 0.2, and the prevalence was assumed to 0.2. The relative phenotypic variance attributable to the main disease gene was assumed to be 0.005

	$n_{proband}$	Statistic	*N*					
			100	300	600	900	1200	1400
0.01	1	*WL*	0.052	0.142	0.369	0.518	0.682	0.765
		$FQLS_1$	0.053	0.145	0.352	0.523	0.681	0.766
		$FQLS_2$	**0.056**	**0.151**	**0.389**	**0.543**	**0.702**	**0.796**
	2	*WL*	0.053	0.174	0.396	0.616	0.761	0.829
		$FQLS_1$	**0.054**	0.183	0.400	0.619	0.780	0.834
		$FQLS_2$	0.053	**0.202**	**0.422**	**0.658**	**0.799**	**0.859**
	3	*WL*	0.068	0.118	0.470	0.692	0.808	0.818
		$FQLS_1$	**0.073**	0.216	0.489	0.705	**0.826**	**0.905**
		$FQLS_2$	0.067	**0.232**	**0.498**	**0.729**	**0.826**	0.904
	4	*WL*	0.066	0.222	0.528	0.755	0.860	0.938
		$FQLS_1$	0.068	**0.250**	**0.560**	**0.774**	**0.882**	**0.939**
		$FQLS_2$	**0.072**	0.244	0.544	0.766	0.870	0.935
0.001	1	*WL*	0.012	0.034	0.143	0.247	0.376	0.505
		$FQLS_1$	**0.015**	**0.046**	0.142	0.245	0.387	0.496
		$FQLS_2$	0.013	**0.046**	**0.169**	**0.305**	**0.442**	**0.567**
	2	*WL*	0.008	0.060	0.151	0.342	0.494	0.612
		$FQLS_1$	**0.014**	**0.055**	0.163	0.357	0.512	0.640
		$FQLS_2$	0.012	**0.055**	**0.184**	**0.378**	**0.538**	**0.648**
	3	*WL*	0.005	0.033	0.223	0.404	0.579	0.595
		$FQLS_1$	0.008	0.076	0.235	0.432	0.604	**0.717**
		$FQLS_2$	**0.010**	**0.078**	**0.236**	**0.438**	**0.610**	0.699
	4	*WL*	0.010	0.079	0.274	0.484	0.655	0.763
		$FQLS_1$	0.008	**0.088**	**0.296**	**0.490**	**0.677**	**0.783**
		$FQLS_2$	**0.014**	0.084	0.280	0.474	0.671	0.769

The bold text indicates the highest empirical estimate of the power for each situation

Table 4 Empirical power estimates for scenario 1 when h^2 is 0.8. The empirical power estimates for scenario 1 were calculated with 1000 replicates at the both 0.01 and 0.001 significance levels. The disease allele frequency was assumed to be 0.2, and the prevalence was assumed to 0.2. The relative phenotypic variance attributable to the main disease gene was assumed to be 0.005

	$n_{proband}$	Statistic	*N*					
			100	300	600	900	1200	1400
0.01	1	*WL*	0.068	0.233	0.470	0.699	0.819	0.903
		$FQLS_1$	0.071	0.238	0.471	0.717	0.841	0.896
		$FQLS_2$	**0.078**	**0.298**	**0.559**	**0.817**	**0.899**	**0.943**
	2	*WL*	0.071	0.222	0.521	0.708	0.881	0.937
		$FQLS_1$	0.080	0.304	0.568	0.788	0.907	0.942
		$FQLS_2$	**0.085**	**0.311**	**0.605**	**0.830**	**0.930**	**0.948**
	3	*WL*	0.075	0.253	0.555	0.786	0.911	0.931
		$FQLS_1$	**0.079**	0.298	0.592	0.813	0.921	0.972
		$FQLS_2$	0.078	**0.311**	**0.629**	**0.823**	**0.941**	**0.982**
	4	*WL*	0.081	0.307	0.585	0.800	0.917	0.951
		$FQLS_1$	0.088	**0.325**	**0.622**	0.828	0.928	0.957
		$FQLS_2$	**0.092**	0.318	0.614	**0.832**	**0.929**	**0.962**
0.001	1	*WL*	0.016	0.074	0.229	0.444	0.602	0.710
		$FQLS_1$	0.017	0.072	0.221	0.436	0.643	0.693
		$FQLS_2$	**0.021**	**0.120**	**0.309**	**0.573**	**0.739**	**0.820**
	2	*WL*	0.020	0.074	0.251	0.436	0.676	0.778
		$FQLS_1$	0.017	0.103	0.313	0.540	0740	0.798
		$FQLS_2$	**0.024**	**0.116**	**0.350**	**0.589**	**0.782**	**0.845**
	3	*WL*	0.024	0.081	0.278	0.513	0.734	0.810
		$FQLS_1$	0.016	0.124	0.341	0.581	0.769	0.864
		$FQLS_2$	**0.017**	**0.129**	**0.365**	**0.604**	**0.802**	**0.881**
	4	*WL*	0.015	0.109	0.310	0.542	0.756	0.830
		$FQLS_1$	**0.021**	0.118	0.335	0.588	0.783	0.867
		$FQLS_2$	**0.021**	**0.133**	**0.345**	**0.597**	**0.790**	**0.874**

The bold text indicates the highest empirical estimate of the power for each situation

Table 5 Empirical power estimates for scenario 2 when h^2 is 0.2. The empirical power estimates for scenario 2 were calculated with 1000 replicates at the both 0.01, and 0.001 significance levels. The disease allele frequency was assumed to be 0.2, and the prevalence was assumed to 0.2. The relative phenotypic variance attributable to the main disease gene was assumed to be 0.005

	$n_{proband}$	Statistic	N			
			100	300	600	900
0.01	1	*WL*	0.072	0.149	0.304	0.409
		$FQLS_1$	0.072	0.147	0.305	0.415
		$FQLS_2$	**0.075**	**0.158**	**0.312**	**0.434**
	2	*WL*	0.042	0.137	0.300	0.448
		$FQLS_1$	0.039	0.136	**0.303**	0.455
		$FQLS_2$	**0.041**	**0.139**	0.295	**0.471**
	3	*WL*	**0.058**	0.188	0.410	0.608
		$FQLS_1$	0.054	**0.191**	**0.424**	**0.620**
		$FQLS_2$	0.055	0.190	0.423	0.615
0.001	1	*WL*	**0.025**	0.059	0.147	0.211
		$FQLS_1$	0.023	**0.062**	**0.152**	0.197
		$FQLS_2$	0.022	0.055	**0.152**	**0.212**
	2	*WL*	0.010	**0.038**	0.123	0.229
		$FQLS_1$	**0.012**	0.036	0.123	**0.232**
		$FQLS_2$	0.010	0.036	**0.127**	0.227
	3	*WL*	0.006	0.055	**0.182**	0.342
		$FQLS_1$	0.007	0.055	**0.182**	0.355
		$FQLS_2$	**0.009**	**0.057**	**0.182**	**0.356**

The bold text indicates the highest empirical estimate of the power for each situation

Table 6 Empirical power estimates for scenario 2 when h^2 is 0.5. The empirical power estimates for scenario 2 were calculated with 1000 replicates at the both 0.1 and 0.001 significance levels. The disease allele frequency was assumed to be 0.2, and the prevalence was assumed to 0.2. The relative phenotypic variance attributable to the main disease gene was assumed to be 0.005

	$n_{proband}$	Statistic	N			
			100	300	600	900
0.01	1	*WL*	0.130	0.293	0.567	0.787
		$FQLS_1$	0.130	0.295	0.568	0.773
		$FQLS_2$	**0.140**	**0.323**	**0.603**	**0.823**
	2	*WL*	0.093	0.332	0.645	0.864
		$FQLS_1$	0.094	**0.357**	0.654	0.871
		$FQLS_2$	**0.108**	0.354	**0.694**	**0.902**
	3	*WL*	0.100	0.382	0.735	0.904
		$FQLS_1$	0.108	0.406	0.751	0.915
		$FQLS_2$	**0.130**	**0.408**	**0.772**	**0.932**
0.001	1	*WL*	0.046	0.148	0.341	0.560
		$FQLS_1$	**0.050**	0.149	0.353	0.559
		$FQLS_2$	0.047	**0.166**	**0.386**	**0.617**
	2	*WL*	0.019	0.127	0.387	0.634
		$FQLS_1$	**0.021**	0.119	0.394	0.648
		$FQLS_2$	0.017	**0.144**	**0.432**	**0.695**
	3	*WL*	0.023	0.166	0.481	0.749
		$FQLS_1$	0.026	0.183	0.511	0.772
		$FQLS_2$	**0.028**	**0.196**	**0.532**	**0.782**

The bold text indicates the highest empirical estimate of the power for each situation

If h^2 was 0.2, the proposed methods were only slightly better than *WL*. While all methods in our power comparison focused on the distribution of genotypes to calculate statistics, the proposed methods uniquely considered the heterogeneous effects of ascertainment bias among family members which were proportional to the magnitude of heritability; this explained the power improvement of the proposed methods. Furthermore the differences of empirical power estimates from *WL* and the proposed methods are larger for Tables 5, 6, 7 than Tables 2, 3, 4, which indicates that the heterogeneity of ascertainment condition may be positively related with family size and the proposed methods become more efficient for large families. Last our simulation results show that $FQLS_2$ was slightly better than $FQLS_1$, and this may be induced by the uncertainty of probands in our simulation studies. Therefore, we concluded that the incorporation of a sampling scheme to the offset could make a substantial difference, and test statistic should be carefully selected depending on type of sampling scheme.The bold text indicates the highest empirical estimate of the power for each situationThe bold text indicates the highest empirical estimate of the power for each situationThe bold text indicates the highest empirical estimate of the power for each situationThe bold text indicates the highest empirical estimate of the power for each situationThe bold text indicates the highest empirical estimate of the power for each situation

Robustness of the proposed methods against the misspecification of prevalence

The statistical powers of the proposed methods may depend on the accuracy of the prevalence and we evaluated the sensitivity of the proposed method to the misspecified prevalence with simulated data. h_a^2 and h^2 were assumed to be 0.05 and 0.8, and nuclear families (Fig. 2-(a)) were considered in this simulation. The number of nuclear families was assumed to be 900 and $n_{proband}$ was assumed to be 1, 2, or 3. Prevalence was assumed to be 0.2 for phenotype generation, and the offset for which we recommended prevalence was set to be 0.1, 0.2 or 0.3 for calculation of $FQLS_1$ and $FQLS_2$. In particular, y_{ij} for individuals with missing phenotypes are coded by the

Table 7 Empirical power estimates for scenario 2 when h^2 is 0.8. The empirical power estimates for scenario 2 were calculated with 1000 replicates at the both 0.01 and 0.001 significance levels. The disease allele frequency was assumed to be 0.2, and the prevalence was assumed to 0.2. The relative phenotypic variance attributable to the main disease gene was assumed to be 0.005

	$n_{proband}$	Statistic	N			
			100	300	600	900
0.01	1	*WL*	0.164	0.445	0.749	0.906
		$FQLS_1$	0.156	0.441	0.751	0.905
		$FQLS_2$	**0.194**	**0.508**	**0.817**	**0.944**
	2	*WL*	0.132	0.473	0.823	0.970
		$FQLS_1$	0.131	0.505	0.861	0.969
		$FQLS_2$	**0.150**	**0.564**	**0.905**	**0.978**
	3	*WL*	0.140	0.475	0.835	0.958
		$FQLS_1$	0.134	0.520	0.867	0.970
		$FQLS_2$	**0.167**	**0.556**	**0.886**	**0.981**
0.001	1	*WL*	0.059	0.230	0.519	0.759
		$FQLS_1$	0.053	0.236	0.519	0.757
		$FQLS_2$	**0.070**	**0.311**	**0.632**	**0.841**
	2	*WL*	0.039	0.239	0.561	0.858
		$FQLS_1$	0.033	0.250	0.594	0.884
		$FQLS_2$	**0.053**	**0.300**	**0.702**	**0.924**
	3	*WL*	0.033	0.215	0.629	0.865
		$FQLS_1$	**0.046**	0.247	0.671	0.900
		$FQLS_2$	0.044	**0.287**	**0.713**	**0.925**

The bold text indicates the highest empirical estimate of the power for each situation

Table 8 Empirical power estimates for three situations when h^2 is 0.8. The empirical power estimates for three situations were calculated with 1000 replicates at the 0.01 significance levels. Phenotypes were generated under the assumption that the prevalence was assumed to 0.2. Prevalence was set to be 0.1, 0.2 or 0.3 to calculate the proposed statistics. The relative phenotypic variance attributable to the main disease gene was assumed to be 0.005

$n_{missing}$	$n_{proband}$	Statistic	Prevalence to be set for statistics		
			0.1	0.2	0.3
0	1	$FQLS_1$	0.486	**0.479**	**0.464**
		$FQLS_2$	**0.572**	**0.539**	0.525
	2	$FQLS_1$	0.562	**0.557**	0.532
		$FQLS_2$	**0.606**	**0.621**	0.624
	3	$FQLS_1$	0.632	**0.642**	0.634
		$FQLS_2$	0.644	**0.647**	0.654
1	1	$FQLS_1$	0.432	**0.434**	**0.418**
		$FQLS_2$	0.468	**0.469**	0.462
	2	$FQLS_1$	0.522	**0.530**	0.519
		$FQLS_2$	0.547	**0.554**	0.559
	3	$FQLS_1$	0.580	**0.584**	0.576
		$FQLS_2$	0.575	**0.584**	0.585
2	1	$FQLS_1$	0.415	**0.412**	0.399
		$FQLS_2$	0.428	**0.420**	0.412
	2	$FQLS_1$	0.482	**0.487**	**0.468**
		$FQLS_2$	0.483	**0.495**	0.492
	3	$FQLS_1$	0.487	**0.497**	0.490
		$FQLS_2$	**0.500**	**0.515**	0.511

The bold text indicates the reference for the proposed statistics to compare with the misspecified prevalence

assumed prevalence, and sensitivity of the proposed methods can be substantial when there are individuals with missing phenotypes. Therefore, individuals were randomly selected from non-probands, and their phenotypes were assumed to be unknown for calculation of the proposed statistics. The number of family members with missing phenotypes in each family was denoted by $n_{missing}$. The empirical powers were calculated at the 0.01 significance level with 1000 replicates. Table 8 shows that the results obtained by setting prevalence to be 0.1 and 0.3 are similar to the results when the prevalence was set to be 0.2, which indicates that the power loss attributable to the misspecified prevalence is not substantial. Furthermore the empirical power estimates are positively related with $n_{proband}$ and inversely related with $n_{missing}$. If $n_{missing}$ is larger than 3, the power loss may be more substantial.The bold text indicates the reference for the proposed statistics to compare with the misspecified prevalence.

Application of the proposed method to AD

AD is an irreversible, progressive brain disorder characterized by genetic heterogeneity. However, the genetic variations that contribute to AD still remain elusive. Thus, we applied the proposed method for identification of the disease susceptibility loci for AD. The heritability and prevalence of AD are approximately 0.8 [29, 30] and 0.1, respectively; therefore, we chose heritabilities of 0.8, and a prevalence of 0.1 for the calculation of proposed methods. Samples were collected as part of the National Institute of Mental Health Genetics Initiative (NIMH). The NIMH Alzheimer's Disease Genetics Family Sample was used along with the information about the genotype platform (Affy 6.0) [20, 31], and ethical approach and participant approval were obtained through the NIMH IRB panel. Families were selected based on the disease status of a certain family member. However the proband for each family was not clear, and $FQLS_2$ was uniquely applied for the proposed method. 1376 individuals from 410 families were available, and all families were nuclear. All individuals were of self-reported European ancestry. HWE for each single nucleotide polymorphism (SNP) was tested, and MAFs were estimated. SNPs for which p-values for HWE were less than 10^{-6} or MAFs less than

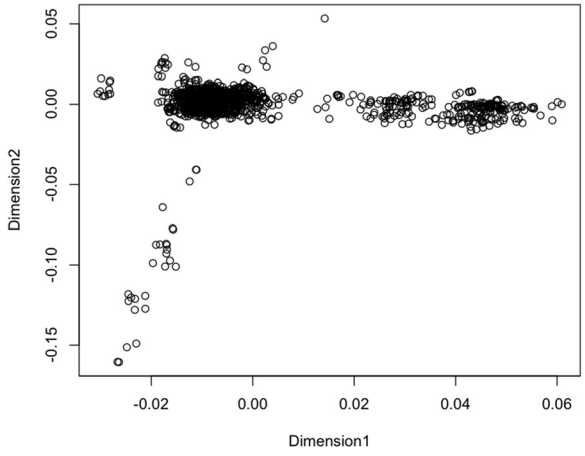

Fig. 4 Multidimensional scaling plots from samples for the GWAS for AD. Founders were selectively used, and multidimensional scaling plots were obtained with the first and second PC scores

0.05 were excluded, and therefore, 417,680 SNPs were analyzed for genetic association analysis.

The choice of kinship coefficient matrix for the proposed method depends on the presence of population substructure. To confirm the presence of population substructure, multidimensional-scaling [32] analysis was performed with PLINK1.07 [20], and we constructed a multidimensional-scaling plot (Fig. 4) to provide evidence concerning the presence of population substructure. Therefore, the genomic control method [31, 33, 34] was used for explicit detection and correction of population stratification with common SNPs, and they were incorporated into both $FQLS_2$ and WL. Fig. 5 shows

Table 9 Top significant results of GWAS for AD. For the genome-wide significant SNPs from $FQLS_2$ and WL, their p-values are given

SNP	WL	$FQLS_2$	$FBAT$
SNP1	3.53×10^{-7}	1.94×10^{-8}	4.23×10^{-4}
SNP2	5.74×10^{-7}	4.45×10^{-8}	4.9×10^{-5}
SNP3	1.69×10^{-7}	8.36×10^{-9}	8.6×10^{-5}
SNP4	2.79×10^{-9}	2.86×10^{-10}	6.94×10^{-12}

QQ plots for $FQLS_2$ and WL; these plots revealed that the presence of population substructure was appropriately adjusted. Results showed four genome-wide significant results for $FQLS_2$ and one significant result for WL.

Detailed information for these significant results is provided in Table 9. We also considered FBAT statistics [35]. $FQLS_2$ identified four genome-wide significant SNPs, while WL and FBAT identified one genome-wide significant SNP. In addition, one SNP that was significant according to WL was more significant according to $FQLS_2$. The most significant result acquired using $FQLS_2$ was SNP4 ($p = 2.10\times 10^{-10}$). The other three SNPs, i.e., SNP1, SNP2 and SNP3, reached the genome-wide significance level.

Discussion

Although major advances in high-density genome scans have enabled the genetic association analysis of more than 10,000 individuals, disappointing results in the mapping of many common diseases have illustrated the need for more powerful methods for detecting disease susceptibility loci. Statistical efficiency is known to being affected by the ascertainment bias, and its careful adjustment can

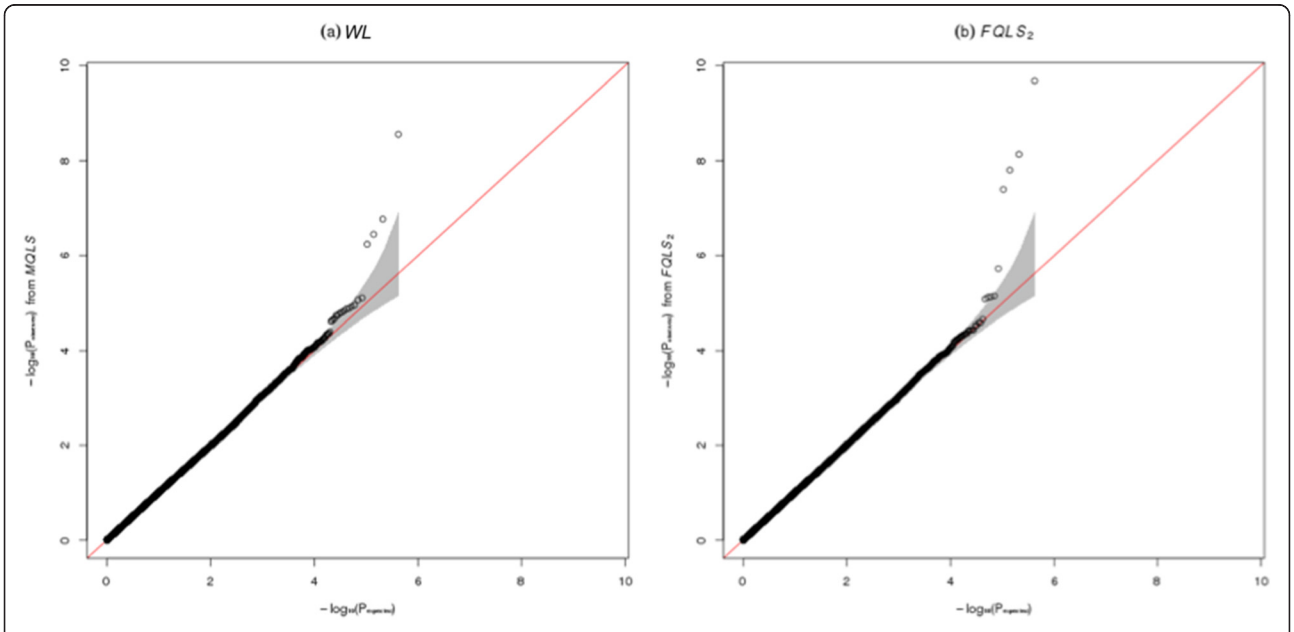

Fig. 5 QQ plots of results from GWAS for AD. QQ plots are provided with the results from **(a)** WL and **(b)** $FQLS_2$

lead to substantial improvement of statistical power [36–38]. In particular, genetic association analysis has often been conducted using family-based designs, but without addressing the fact that the probability for each family member to be affected is inversely related with the familial relationship with affected probands. In this report, we proposed new methods to adjust this heterogeneity with known prevalence and heritability. Our simulation studies showed that the proposed methods provided substantial power improvement. In particular, the mis-specified heritability and prevalence can lead to the statistical power loss for the proposed methods, but it was found to be not substantial at least in our simulation studies. $FQLS_1$ and $FQLS_2$ were suggested, and $FQLS_1$ is an efficient choice if probands for each family are clearly defined and all remaining family members are incorporated to the genetic analysis. However, these conditions are often not satisfied, and different methods such as sequential sampling frame [28] are usually utilized. Simulation studies showed that $FQLS_2$ is usually better than $FQLS_1$ if the ascertaining condition is not clearly defined and thus we recommend $FQLS_2$ unless probands are clearly defined. However we considered the limited ascertainment conditions and comprehensive simulation studies are still necessary.

Furthermore, the proposed method was conceptually simple and can be applied to the large families. Our methods require only a single calculation of offset for all markers, and the real data analysis could be completed with a single CPU in a few hours. For M markers and N individuals, the time complexity is $O(N^3 + MN^2)$ for the proposed method. The proposed method was implemented with C++, and can be downloaded from http://healthstat.snu.ac.kr/mfqls/.

Heterogeneity between samples is an important issue in large-scale genetic analysis, and the proposed method can likely be applied to various additional scenarios with some modifications. For instance, the disease status of relatives reveals the importance of genetic components for each individual, and for this reason, such information has been used, albeit only on occasion, in genetic association analysis. The effect of relatives' disease statuses is dependent on prevalence and heritability, and the probability for each individual to be affected could be calculated with the proposed method. This probability can be used to improve the statistical efficiency of genetic association analysis. In addition, the heterogeneity of the ascertainment bias is often an important issue for genome-wide meta-analysis because samples are collected from multiple medical centers [39, 40], and different sampling schemes among studies need to be adjusted to improve statistical efficiency. Therefore, we believe that proposed method can be extended to provide a statistical framework that adjusts the heterogeneity between samples.

Conclusions

We proposed *FQLS* method to adjust this heterogeneity with known prevalence and heritability and the software was implemented with C++. We identified several significant associations between AD and SNPs, and their potential functional information will provide the better understanding of the pathogenesis of AD. Although this study has some limitations, our proposed methods illustrated important features required for genetic analysis with family-based samples, and an extension of the proposed method to rare variant association analysis such as *FARVAT* [41] will be investigated in future studies.

Additional file

Additional file 1: Web-based Supporting Materials for "Adjusting heterogeneous ascertainment bias for genetic association analysis with extended families" by Suyeon Park, Sungyoung Lee, Young Lee, Christine Herold , Basavaraj Hooli, Kristina Mullin, Lars Bertram, Taesung Park, Changsoon Park, Christoph Lange, Rudolph Tanzi , and Sungho Won.

Abbreviations

GWAS: Genome-wide association studies; CA: Cochran-Armitage; AD: Alzheimer's diseas; PD: Parkinson's disease; HWE: Hardy-Weinberg equilibrium; BLUE: Best linear unbiased estimator; QQ: Quantile quantile; MVN: Multivariate normal; MAF: Minor allele frequency; SNP: Single nucleotide polymorphism; FBAT: Family Based Association Testing software; FQLS: Family based quasi-likelihood score test.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SP participated in the design of the study, and performed the simulation studies and statistical analysis. SL and TP developed the program. YL conducted the statistical analysis. CH, BH, KM, CP, LB, CL and RT conceived of the study. SW conceived of the study, and participated in data analysis. All authors read and approved the final manuscript.

Acknowledgements

SP, SL, YL, and SW were supported by the Industrial Core Technology Development Program (10040176, Development of Various Bioinformatics Software Using Next Generation Bio-Data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea), and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2010437). CP was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-220-C00004). BH, KM, CL and RT were supported by NIMH RO1 grant (2R01MH060009) and the Cure Alzheimer's Fund.

Author details

[1]Department of Applied Statistics, Chung-Ang University, Seoul, Korea. [2]Center for Genome Science, National Institute of Health, Osong Health Technology Administration complex, Chungcheongbuk-do, Seoul, Korea. [3]Department of Biostatistics, Soonchunhyang University, College of Medicine, Seoul, Korea. [4]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea. [5]German Center for Neurodegenerative Diseases (DZNE), Sigmund-Freud-Str. 25, Bonn 53127, Germany. [6]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA. [7]Genetics and Aging Research Unit, MassGeneral Institute for Neurodegenerative Diseases, Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Massachusetts, USA. [8]Department of Statistics, Seoul National University, Seoul, Korea. [9]Department of Vertebrate Genomics, Neuropsychiatric Genetics Group, Max Planck Institute for Molecular Genetics, Berlin, Germany. [10]Department of Medicine, School of Public Health, Imperial

College London, London, UK. [11]Harvard Medical School, Boston, MA, USA. [12]Institute for Genomic Mathematics, University of Bonn, Bonn, Germany. [13]German Center for Neurodegenerative Diseases, Bonn, Germany. [14]Department of Public Health Science, Seoul National University, Seoul, Korea. [15]Institute of Health and Environment, Seoul National University, Seoul, Korea. [16]National Cancer Center, Seoul, Korea.

Received: 6 December 2014 Accepted: 6 July 2015
Published online: 19 August 2015

References

1. Maher B. Personal genomes: the case of the missing heritability. Nature. 2008;456(7218):18–21.
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.
3. Visscher PM. Sizing up human height variation. Nat Genet. 2008;40(5):489–90.
4. Clement K, Vaisse C, Manning BS, Basdevant A, Guygrand B, Ruiz J, et al. Genetic-variation in the beta(3)-adrenergic receptor and an increased capacity to gain weight in patients with morbid-obesity. New Engl J Med. 1995;333(6):352–4.
5. Gu C, Todorov AA, Rao DC. Genome screening using extremely discordant and extremely concordant pairs. Genet Epidemiol. 1997;14(6):791–6.
6. Hu SW, Zhong YF, Hao YT, Luo MQ, Zhou Y, Guo H, et al. Novel rare alleles of ABCA1 are exclusively associated with extreme high-density lipoprotein-cholesterol levels among the Han Chinese. Clin Chem Lab Med. 2009;47(10):1239–45.
7. Khor CC, Goh DLM. Strategies for identifying the genetic basis of dyslipidemia: genome-wide association studies vs. the resequencing of extremes. Curr Opin Lipidol. 2010;21(2):123–7.
8. Li BS, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83(3):311–21.
9. Liang KY, Huang CY, Beaty TH. A unified sampling approach for multipoint analysis of qualitative and quantitative traits in sib pairs. Am J Hum Genet. 2000;66(5):1631–41.
10. Price RA, Li WD, Zhao H. FTO gene SNPs associated with extreme obesity in cases, controls and extremely discordant sister pairs. BMC Med Genet. 2008;9:4.
11. Risch N, Zhang H. Extreme discordant sib pairs - the design of choice for mapping quantitative trait loci in humans. Am J Hum Genet. 1995;57(4):1159–9.
12. Risch NJ, Zhang HP. Mapping quantitative trait loci with extreme discordant sib pairs: Sampling considerations. Am J Hum Genet. 1996;58(4):836–43.
13. Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR. Power of deep, all-exon resequencing for discovery of human trait genes. Proc Natl Acad Sci U S A. 2009;106(10):3871–6.
14. Lander ES, Botstein D. Mapping Mendelian factors underlying quantitative traits using Rflp linkage maps. Genetics. 1989;121(1):185–99.
15. Van Gestel S, Houwing-Duistermaat JJ, Adolfsson R, van Duijn CM, Van Broeckhoven C. Power of selective genotyping in genetic association analyses of quantitative traits. Behav Genet. 2000;30(2):141–6.
16. Sasieni PD. From genotypes to genes: doubling the sample size. Biometrics. 1997;53(4):1253–61.
17. Won S, Lange C. A general framework for robust and efficient association analysis in family-based designs: quantitative and dichotomous phenotypes. Stat Med. 2013;32(25):4482–98.
18. Thornton T, McPeek MS. Case–control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet. 2007;81(2):321–37.
19. Falconer DS. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. Ann Hum Genet. 1967;31(1):1–20.
20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
21. Yang JA, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
22. Lange C, DeMeo DL, Laird NM. Power and design considerations for a general class of family-based association tests: Quantitative traits. Am J Hum Genet. 2002;71(6):1330–41.
23. McPeek MS, Wu X, Ober C. Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics. 2004;60(2):359–67.
24. Benchek PH, Morris NJ. How meaningful are heritability estimates of liability? Hum Genet. 2013;132(12):1351–60.
25. Leppard P, Tallis GM. Evaluation of the mean and covariance of the truncated multinormal distribution. J R Stat Soc C. 1989;38(3):543–53.
26. Aitken AC. Note on selection from a multivariate normal population. Proc Edinburgh Mathematical Society B. 1934;4:106–10.
27. Pearson K. Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. Philos Transact A Math Phys Eng Sci. 1903;200:1–66.
28. Elston RC, Sobel E. Sampling considerations in the gathering and analysis of pedigree data. Am J Hum Genet. 1979;31(1):62–9.
29. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, et al. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry. 2006;63(2):168–74.
30. Gatz M, Pedersen NL, Berg S, Johansson B, Johansson K, Mortimer JA, et al. Heritability for Alzheimer's disease: the study of dementia in Swedish twins. J Gerontol A Biol Sci Med Sci. 1997;52(2):M117–25.
31. 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(8):904–9.
32. Li Q, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. Genet Epidemiol. 2008;32(3):215–26.
33. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55(4):997–1004.
34. Reich DE, Goldstein DB. Detecting association in a case–control study while correcting for population stratification. Genet Epidemiol. 2001;20(1):4–16.
35. Laird NM, Horvath S, Xu X. Implementing a unified approach to family-based tests of association. Genet Epidemiol. 2000;19 Suppl 1:S36–42.
36. Vieland VJ, Hodge SE. Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. Am J Hum Genet. 1995;56(1):33–43.
37. Nielsen R, Hubisz MJ, Clark AG. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics. 2004;168(4):2373–82.
38. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005;15(11):1496–502.
39. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet. 2010;86(1):6–22.
40. Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM, Breen G, et al. A mega-analysis of genome-wide association studies for major depressive disorder. Mol Psychiatry. 2013;18(4):497–511.
41. Choi S, Lee S, Cichon S, Nothen MM, Lange C, Park T, Won S: FARVAT: a family-based rare variant association test. Bioinformatics. 2014;30(22):3197-205.