BMC Genetics

METHODOLOGY ARTICLE Open Access

A general semi-parametric approach to the analysis of genetic association studies in population-based designs

Sharon Lutz[1,3,4*], Wai-Ki Yip[3,4], John Hokanson[2], Nan Laird[3,4] and Christoph Lange[3,4,5,6]

Abstract

Background: For genetic association studies in designs of unrelated individuals, current statistical methodology typically models the phenotype of interest as a function of the genotype and assumes a known statistical model for the phenotype. In the analysis of complex phenotypes, especially in the presence of ascertainment conditions, the specification of such model assumptions is not straight-forward and is error-prone, potentially causing misleading results.

Results: In this paper, we propose an alternative approach that treats the genotype as the random variable and conditions upon the phenotype. Thereby, the validity of the approach does not depend on the correctness of assumptions about the phenotypic model. Misspecification of the phenotypic model may lead to reduced statistical power. Theoretical derivations and simulation studies demonstrate both the validity and the advantages of the approach over existing methodology. In the COPDGene study (a GWAS for Chronic Obstructive Pulmonary Disease (COPD)), we apply the approach to a secondary, quantitative phenotype, the Fagerstrom nicotine dependence score, that is correlated with COPD affection status. The software package that implements this method is available.

Conclusions: The flexibility of this approach enables the straight-forward application to quantitative phenotypes and binary traits in ascertained and unascertained samples. In addition to its robustness features, our method provides the platform for the construction of complex statistical models for longitudinal data, multivariate data, multi-marker tests, rare-variant analysis, and others.

Keywords: Genetic associations studies, Secondary phenotypes, Case-control, Ascertainment, Semi-parametric

Background

In genetic association studies, individuals are often recruited based on case-control ascertainment conditions of the primary phenotype [1]. For the analysis of secondary phenotypes, this recruitment-scheme can become problematic. If the secondary phenotype is correlated with the primary phenotype in a case-control study, the distribution of the secondary phenotype can be fundamentally different from the general population. For example, in a genetic association study of COPD in which all cases have COPD and control subjects have normal pulmonary function, the distribution of quantitative lung phenotypes can deviate substantially from their distribution in the general population. For samples that are ascertained in this fashion, standard statistical methods may lead to misleading results or may lack statistical power to identify true genotype phenotype associations. There are several methods to accurately estimate the odds ratio of genetic variants for binary secondary phenotypes associated with case-control status [2-10], but these methods cannot easily accommodate continuous secondary phenotypes. For the special case that the secondary phenotype is normally distributed or binary, Lin & Zeng (2009) proposed an adjusted score test that incorporates genetic associations with affection status into the test statistic [11].

*Correspondence: sharon.lutz@ucdenver.edu
[1] Department of Biostatistics, University of Colorado Anschutz Medical Campus, Aurora, USA
[3] Department of Biostatistics, Harvard School of Public Health, Boston, USA
Full list of author information is available at the end of the article

We present a more general approach that does not require any distribution assumptions for the secondary phenotype. We refer to the approach as the non-parametric population-based association test (NPBAT). The approach has a form similar to the Family Based Association Test (FBAT), a non-parametric test statistic that is frequently used in the family based setting [12-15]. The flexibility of our approach allows us to construct a genetic association test for standard and complex phenotypes that is non-parametric with respect to the phenotype. The class of tests is very general. It includes most standard association tests and can be applied to multivariate traits and phenotypes, multiple genetic markers, and case- control studies where phenotypic information is available for the cases but correlated with the case-control status [16-18].

The general concept of the proposed association-testing framework is to condition on the phenotype of interest and treat only the genetic data as random [12,13,15]. By assuming that the phenotype data is deterministic, the validity of the approach does not depend on the correctness of the phenotypic assumptions. Nevertheless, the power of the approach can be increased by incorporating a plausible model for the phenotype into the test statistic. Based on theoretical considerations and on simulation studies, we show that the new approach is robust against misspecification of phenotype assumptions. At the same time, this approach achieves the same power level as standard genetic association tests for population-based designs when the phenotype of interest has a normal distribution or is dichotomous. For studies where a quantitative trait is correlated with case-control status, our simulation studies examine the power and significance levels for the proposed approach, which does not require any adjustment for the ascertainment conditions.

We illustrate the practical advantages of NPBAT by an application to the COPDGene study. The COPDGene study is a case-control study of the genetics of COPD in current or former smokers with at least 10 pack-years of smoking history [19]. We test the genetic association of single nucleotide polymorphisms (SNPs) in the CHRNA 3/5 region and the Fagerstrom Nicotine Dependence score (FNDS). FNDS is a validated instrument of nicotine dependence in current smokers and was measured in the current smokers, but not former smokers in the COPDGene study. NPBAT, which uses the genotype data in both current and former smokers, is compared to the published genetic association of SNPs in the CHRNA 3/5 region and FNDS that was performed in current smokers only [20].

Methods

In a genetic association study, n unrelated study subjects have been recruited based on a predefined ascertainment condition. Let X_i denote the genotype of the individual i. The specific value of X_i will depend upon the genetic model under consideration. For instance, for an additive model, $X_i = 0, 1, 2$ for 0,1,2 disease alleles, respectively. X_i may also be a vector in order to test several alleles simultaneously. Let T_i denote the numerical trait information for individual i. For example, T_i could equal one for affected individuals and T_i could equal zero for unaffected individuals. Different coding functions are applied depending on the phenotype of interest. For binary and continuous traits, we will discuss efficient coding schemes below. First, we define a general class of test statistics as

$$S = \sum_{i=1}^{n} (X_i - E_x)\, T_i \tag{1}$$

Note that $E(S) = 0$ under the null hypothesis of no association between the genotype X and the phenotype Y. Constructing a conditional score test in which the genotype X_i is the dependent variable and we condition upon the numerical trait information T_i, the NPBAT statistic has the following form:

$$Stat_{NPBAT} = \frac{S - E[S]}{\sqrt{var(S)}} = \frac{\sum\limits_{i=1}^{n} (X_i - E_x)\, T_i}{\sqrt{\left(\sum\limits_{i=1}^{n} T_i^2\right)\left(\frac{\sum_{i=1}^{n}(X_i - E_x)^2}{n-1}\right)}} \tag{2}$$

where E_x denotes the expectation of the marker score/ genotype X under the null-hypothesis of no genetic association between the phenotype. The marker locus. E_x can be estimated based on the sample mean of the genotypes. The asymptotic distribution of the NPBAT statistic under the null-hypothesis depends on the estimation of E_x and on the specification of the trait information T_i, and is derived in the Appendix.

There are various ways to code the phenotype of interest and define the coding function T_i. For the analysis of affection status, one could specify the coding function to be $T_i = 1$ or $T_i = 0$, depending on the disease status of the proband. However, as we show in the Appendix A, a more efficient way is to set $T_i = 1 - \frac{\#cases}{n}$ for the cases, and $T_i = 0 - \frac{\#cases}{n}$ for the controls. Then the NPBAT statistic is approximately the same as the Cochran-Armitage Trend test.

If the phenotype Y_i is in fact normally distributed and $T_i = Y_i - \hat{Y}_i$ where $\hat{Y}_i$ denotes the fitted values of regressing the phenotype Y on any covariates, then the NPBAT statistic is approximately the same as a t-statistic from a linear regression. In general, if the phenotype Y_i is a continuous phenotype, we recommend $T_i = Y_i - \mu_y$ where μ_y is the phenotypic mean in the general population.

While it is appealing that the NPBAT statistic is comparable to standard methods in these simple scenarios, the real appeal of the NPBAT statistic is when there is only phenotype information available for some subjects but there is genetic information available for all subjects. For example, in case control studies, an additional quantitative phenotype may be available for the cases but not the controls. When testing for a genetic association with this additional quantitative phenotype, the NPBAT statistic uses the genotype of both the cases and the controls with the optimal coded phenotype $T_i = Y_i - Y_{\text{offset}}$ where Y_{offset} is a constant. The choice of this constant is described in detail in the simulations sub-section and the asymptotic distribution of the NPBAT statistic is derived in the Appendix. Using this optimal offset choice, the NPBAT statistic has a substantial increase in power over other methods such as the NPBAT statistic when an offset choice of $T_i = Y_i - \bar{Y}$ or the improved score test, which is uniformly more powerful than score tests based on the generalized linear model such as the Cochran-Armitage trend test, the allelic χ^2 test and the genotypic χ^2 test [21].

Adjustments for population admixture

The NPBAT statistic can be adjusted for population admixture by using standard methods such as principal components analysis or genomic control [22,23]. For example, to account for population admixture, one can treat the principal components as additional covariate representing population information, and incorporate them into the test statistic in equation (2) by taking $T_i = Y_i - \hat{Y}_i$ where $\hat{Y}_i$ denotes the fitted values of regressing the phenotype Y on the top principal components that explain the greatest amount of variability in the data. Note the above approach requires that the phenotype Y is dichotomous or roughly normally distributed.

Extension to multiple phenotypes

The NPBAT statistic can be extended to m phenotypes to test the null hypothesis that a marker locus is not linked to any disease-susceptibility locus for any of m selected phenotypes. Then the test statistic becomes

$$S = \sum_{i=1}^{n} (X_i - E_x) T_i \tag{3}$$

Note that $E(S) = 0$ as is the case for the univariate version above. But here T_i is the $m \times 1$ vector for the m phenotypes and X_i is just one marker. So S is $m \times 1$. The $m \times m$ variance matrix is the following

$$V_S = \hat{\sigma_X^2} \sum_{i=1}^{n} T_i T_i^t \tag{4}$$

where $\hat{\sigma_X^2}$ is the variance for marker X based on sample. Then the NPBAT statistic is the following

$$\chi^2_{\text{NPBAT}} = S^t V_S^{-1} S \tag{5}$$

Due to the estimation of E_x based on the sample, this statistic does not have a chi square distribution and a permutation test needs to be used to assess significance levels, which can be done by using the NPBAT software package (https://sites.google.com/site/genenpbat/).

Simulations

In genetic association case-control studies, only the cases may have additional phenotypic information available. For instance, in a case-control study where the cases have asthma (the primary phenotype), only the cases may have FEV measurements (the secondary phenotype). In this scenario, the secondary phenotype FEV will be more severe than it would be in the general population and the analysis of this secondary phenotype can be misleading due to the ascertainment of subjects based on the primary phenotype, asthma. To simulate this scenario, we generated the genotype X for 500 cases and 500 controls and a secondary phenotype Y for only the 500 cases from a truncated normal distribution with standard deviation $\sigma = 1$, mean aX under the alternative and mean 0 under the null and cutoff such that the secondary phenotype in the top 50 percent of the normal distribution. We consider an allele frequency of $p = 20\%$ and a is chosen such that the heritability h [24] equals $1\%, 2\%, 3\%, 5\%$. The solving for a, $a = \sigma\sqrt{h/2p(1-p)(1-h)}$.

We compute the NPBAT statistic with the coded phenotype $T_i = Y_i - Y_{\text{offset}}$ where Y_{offset} is a constant that ranges from -5 to 15 and E_x is the sample mean of the genotypes in the cases. We also compute the NPBAT statistic with E_x equal to the sample mean of the genotypes in the controls and E_x equal to the sample mean of the genotypes in the cases and the controls. We compare the power of these three NPBAT statistics to the Improved Score Test, which is uniformly more powerful than score tests based on the generalized linear model such as the Cochran-Armitage trend test, the allelic χ^2 test and the genotypic χ^2 test [21]. We also compare the power of the NPBAT approach to a standard linear regression.

Under the null hypothesis, the NPBAT method maintains a significance level of approximately 5% or less as seen in Figure 1 whether E_x is the sample mean of the cases or the controls or both. Figure 1 also depicts the power results of these simulations. Note that the spike or drop in all the plots occurs where $Y_{\text{offset}} \approx \bar{Y}$, the sample mean of the secondary phenotype for the cases since the secondary phenotype is not available for the controls in this scenario. The power of the NPBAT approach is maximized when E_x is based on the genotype of the controls and Y_{offset} is significantly different than the phenotypic

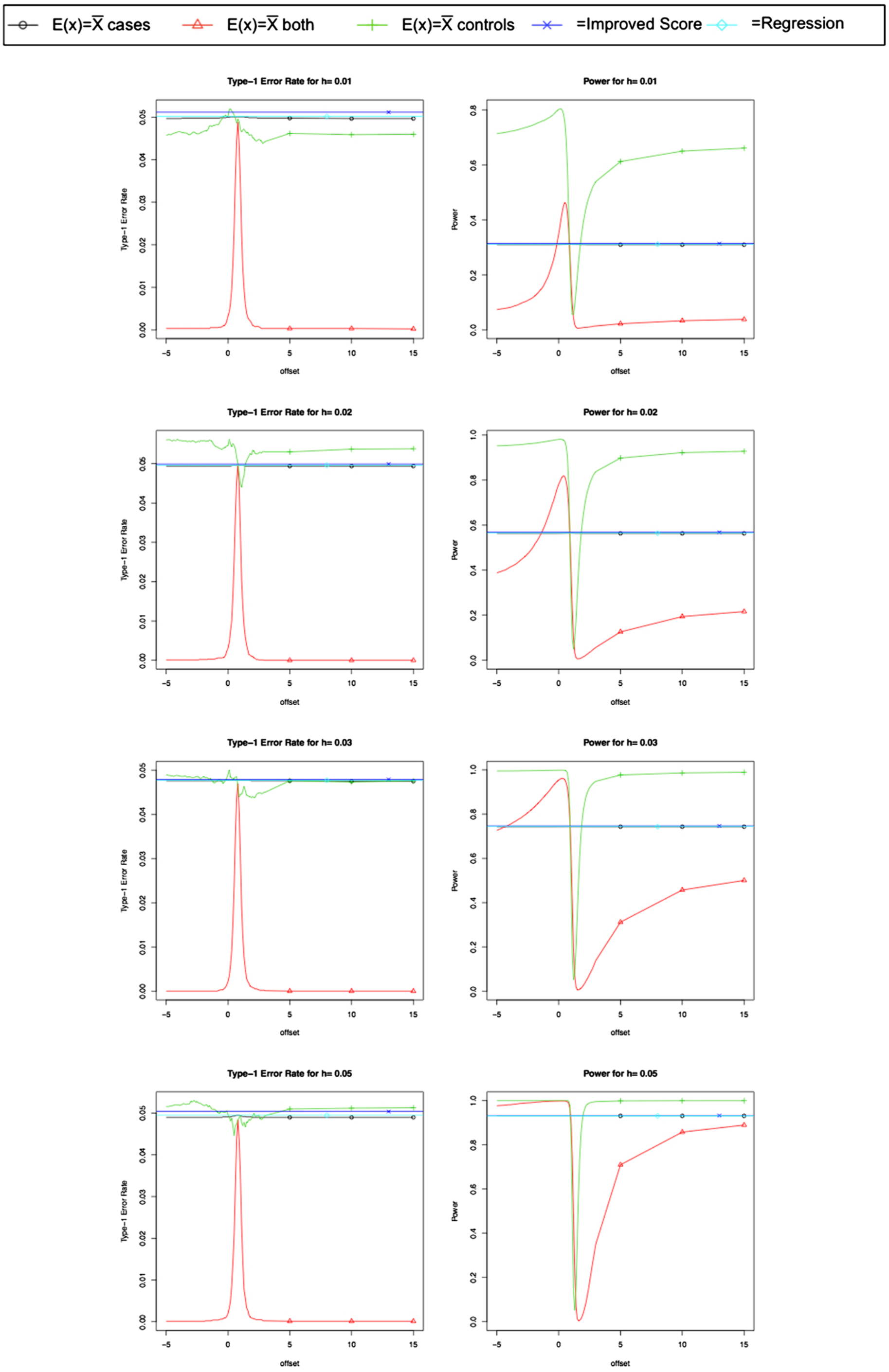

Figure 1 Power and Significance levels for NPBAT, the Improved Score Test and the Likelihood Ratio Test (LRT). This plot compares the power and type-1 error rate of the NPBAT method using E_x based on the sample mean of the cases, the controls and both the cases and controls. The power and significance levels of this method is compared to the improved score test and a standard linear regression. Note that the spike or drop in all the plots occurs where $Y_{\text{offset}} \approx \bar{Y}$, the sample mean of the secondary phenotype for the cases since the secondary phenotype is not available for the controls in this scenario. The power of the NPBAT approach is maximized when E_x is based on the genotype of the controls and Y_{offset} is significantly different than the phenotypic mean of the cases. When E_x is based on the genotype of the cases, the power of the NPBAT approach is similar to the improved score test and the regression. Note that the power of NPBAT approach when E_x is based on the genotype of both the cases and the controls is best for high values of heritability.

mean of the cases. When E_x is based on the genotype of the cases, the power of the NPBAT approach is similar to the improved score test and the regression. Note that the power of NPBAT approach when E_x is based on the genotype of both the cases and the controls is best for high values of heritability.

These simulations show that for case-control studies when analyzing secondary phenotypes correlated with case-control status, we recommend to set Y_{offset} to a constant significantly different from the phenotypic mean of the sample and E_x equal to the genotypic mean of the controls. In this situation, a robust and efficient choice for the offset Y_{offset} is the phenotypic mean in the general population. Note that the results of these simulations are analogous to the FBAT statistic in family studies where it was found that when ascertaining cases only from a quantitative distribution, one needed to choose an offset that was outside the range of the case's phenotypic values [15].

Data analysis

We applied the NPBAT method to the Genetic Epidemiology of COPD (COPDGene) Study which is a multi-center case/control study designed to identify genetic factors associated with COPD and to characterize COPD-related phenotypes [19]. The study recruited COPD cases and smoking controls who were non-Hispanic whites and African Americans ages 45 to 80 with at least 10 pack-years of smoking history. The study also collected the Fagerstrom Test for Nicotine Dependence (FTND) to assess nicotine dependence, but the FTND score was only available for cases and controls who were current smokers at study enrollment. This data analysis represents the scenario where the secondary phenotype (FTND score) is available only in current smokers but the genotypic information is available for both current and former smokers. In the first 1,000 Non-Hispanic White (NHW) individuals, the FTND score controlling for age and gender was tested for an association with SNPs in the CHRNA 3/5 region for COPD cases and controls who are current smokers and association was found for rs1051730 or rs8034191 [20]. We applied the NPBAT statistic to the first 1000 NHW using the genotype of both current (307 individuals) and former smokers (669 individuals), controlling for age and gender and obtained the results shown in Table 1 for these 2 SNPs. Note that the NPBAT statistic performed better than both the Improved Score Test and the regression controlling for age and gender.

Results and discussion

NPBAT is a new statistical framework for population based genetic association tests that does not require making specific assumptions about the distribution of the phenotype. By conditioning on the phenotype, NPBAT is robust against violations of phenotypic model assumptions. The practical implications of NPBAT are demonstrated when applied to the COPDGene Study. FNDS, a measure of nicotine dependence, was assessed in current smokers that represent 31% of study participants in COPDGene. We analyzed SNPs shown to be associated with FNDS [20]. NPBAT identified the same SNPs as conventional methods but with slightly greater statistical significance than a linear regression for FNDS controlling for age and gender or the improved score test. Other examples of applications of NPBAT are

1. when a sample is ascertained based on case/control status and the phenotype of interest is correlated with case status
2. in a cohort study in which prevalent cases are excluded (i.e. the classic epidemiologic cohort study) and the phenotype of interest is correlated with the disease of interest
3. a pharmacogenetics study using a randomized clinical trial when participants are ascertained based on the levels of the target of therapy

The broad application of NPBAT is to scenarios where samples are ascertained based on selection criteria that are correlated with the phenotype of interest.

Conclusions

In conclusion, the key advantage that defines the attraction of the proposed approach is its robustness against model specification of the phenotypes. This enables extensions to different types of traits and the integration of complex statistical models for the phenotype. While, at the same time, the validity of the approach is not compromised by such generalization. Though the power is sensitive to the offset choice, NPBAT is valid regardless of the offset. As with all population-based association

Table 1 This table displays the p-values for the association between the Fagerstrom Test for Nicotine Dependence (FTND) and the markers listed above for the different statistical tests: the NPBAT where $E_x = \bar{x}_c$ is the genotypic mean of the current smokers, NPBAT where $E_x = \bar{x}_f$ is the genotypic mean of the former smokers, the Improved Score Test and a linear regression

Method	NPBAT: $E_x = \bar{x}_c$	NPBAT: $E_x = \bar{x}_f$	Improved Score Test	Regression
rs1051730	0.00134	0.00138	0.00227	0.00259
rs8034191	0.00386	0.00391	0.00694	0.00744

tests, population stratification can be a problem. Adjusting for known population sub-structure using principal components of ancestral informative markers (AIMs) or using genomic controls can reduce the impact of population stratification. The NPBAT software package which implements this method is detailed in the Appendix.

Appendix

Appendix A: Offset choice when Y is binary

The following considers the offset choice for the coded trait T when Y is binary. Assume the phenotype of interest is binary and the genotype of interest follows an additive model. Let r_0, r_1, and r_2 denote the number of cases with 0, 1, and 2 disease alleles, respectively. Let R denote the total number of cases. Let S denote the total number of controls. Let n_0, n_1, and n_2 denote the number of cases and controls with 0, 1, and 2 disease alleles, respectively. Let $N = S + R$ denote the total number of cases and controls. In this scenario, the standard statistical method used is the Cochran-Armitage Trend test which can be written as follows:

$$z_{\text{Cochran}} = \frac{N\left(r_1 + 2r_2\right) - R\left(n_1 + 2n_2\right)}{\sqrt{\left(\frac{SR}{N}\right)\left(N\left(n_1 + 4n_2\right) - \left(n_1 + 2n_2\right)^2\right)}} \tag{6}$$

In this scenario, let the coded phenotype $T_i = Y_i - \mu_y$ where μ_y is the offset. The NPBAT statistic has the following form:

$$\frac{N\left(r_1 + 2r_2\right) - R\left(n_1 + 2n_2\right)}{\sqrt{\left(\left(\frac{N\mu_y^2}{R}\right) + \left(\frac{N(1-\mu_y)^2}{S}\right)\right)\left(\frac{SR(N(n_1+4n_2)-(n_1+2n_2)^2)}{N-1}\right)}} \tag{7}$$

Note that the numerators of both statistics are the same. The ratio of the test statistics can be written as follows:

$$\frac{Stat_{\text{Cochran}}}{Stat_{\text{NPBAT}}} = \sqrt{\frac{N}{N-1}}\sqrt{\left(1 + \frac{1}{\gamma}\right)\mu_y^2 + (1+\gamma)\left(1 - \mu_y\right)^2} \tag{8}$$

where $\gamma = \frac{\#cases}{\#controls}$. Given this ratio, the power of the NPBAT statistic relative to the Cochran-Armitage trend test is maximized for the offset choice $\mu_y^{optimal} = \frac{\gamma}{1+\gamma} = \frac{\#cases}{N}$. For example, if the ratio of the cases versus the controls is 1, the offset choice μ_y is $\frac{1}{2}$. This corresponds to equally weighting the cases and controls in the conditional test statistic. For large sample size N, such that $\sqrt{\frac{N}{N-1}} \approx 1$, the ratio of the test statistics is approximately one when the offset is set to $\mu_y^{optimal} = \frac{\#cases}{n}$. Consequently, for the optimal offset choice, the test statistics are approximately the same.

Appendix B: asymptotic distribution when the secondary phenotype is available for both the cases and controls

To derive the asymptotic distribution of the NPBAT statistics for various phenotypic offset choices, let σ_X^2 denote the variance of X and σ_Y^2 denote the variance of Y. Let $||a||$ denote the Euclidean norm. Let $T_{\text{offset}} = ((Y_1 - Y_{\text{offset}})...(Y_n - Y_{\text{offset}}))^t$ and let $T_\mu = (T_{\mu_1}, ..., T_{\mu_n})^t = ((Y_1 - \bar{Y})...(Y_n - \bar{Y}))^t$ where $T_{\mu_i} = (Y_i - \bar{Y})$. Let $X^t = (X_1 - \bar{X}, ..., X_n - \bar{X})$. Define $Z_i = \frac{(X_i - \bar{X})T_{\mu_i}}{||T_\mu||\hat{\sigma_x}}$. Then $\sum_{i=1}^{n} Z_i = \frac{X^t T_\mu}{||T_\mu||\hat{\sigma_x}}$. By treating X as random given Y is fixed, it can be shown that the Z_is are independent, $E(Z_i) = 0$ and $Var\left(\sum_{i=1}^{n} Z_i\right) = 1$. The Lindberg condition [25] for Z_i, which ensures asymptotic normality of $\sum Z_i$, is then given by

$$\forall \epsilon > 0 : lim_{n\to\infty}\left\{\sum_{i=1}^{n}\int_{|Z_i|\geq\epsilon} Z_i^2 dP\right\} = 0 \tag{9}$$

Since Z_i has a discrete distribution, the Lindberg condition can only be fulfilled when the integration set $\{|Z_i| \geq \epsilon\}$ is empty for $n \to \infty$. Since X is the coded genotype and Y is a biological quantity, assume $\hat{\sigma_x} \neq 0$, $\hat{\sigma_y} \neq 0$ and both are finite. Then, there exists some constant K such that $\frac{|(X_i - \bar{X})||T_{\mu_i}|}{\hat{\sigma_x}\hat{\sigma_y}} \leq K$. Hence we rewrite the Lindberg condition by

$$\forall \epsilon > 0 : \epsilon \leq |Z_i| = \frac{\left|(X_i - \bar{X})\right|\left|T_{\mu_i}\right|}{\hat{\sigma_x}||T_\mu||} \leq \frac{K}{n} \to 0 \text{ as } n \to \infty \tag{10}$$

Hence the integral in the Lindberg condition is always computed over a set that is empty for $n \to \infty$. Thus the Lindberg condition is always fulfilled when the regularity condition holds. Then the Lindberg theorem [26] implies convergence to normality. Then

$$\left(\frac{||T||}{||T_\mu||}\right) Stat_{\text{NPBAT}} = \sum_{i=1}^{n} Z_i \to^d N(0,1) \tag{11}$$

Note that the statistic is maximized and has a standard normal distribution when $Y_{\text{offset}} = E[\,Y]$.

Appendix C: asymptotic distribution when the secondary phenotype is only available for the cases

Here, we derive the asymptotic distribution of the NPBAT statistic for secondary phenotypes in case/control studies. Consider a case control study where genetic information is available for both the cases and the controls, but the phenotypic information is only available for the cases. Here n is only the number of cases and all summations are only over the number of cases since the phenotypic information is not available for the controls where as in Appendix B, n is the number of cases and controls and the summation is over both the number of cases and controls. Let $\bar{X}_{\text{cases}}$ denote the sample mean of the genotypes of the cases and σ_X^2 be the true variance of the genotypes. Let $E_x = \bar{X}_{\text{controls}}$ be the sample mean of the genotypes of the controls. Under the null hypothesis and assuming no population stratification, the sample mean of the genotypes of the cases and the sample mean of the genotypes of the controls both converge to $E[X]$ since X is not associated with Y. Let $X_{\text{text}} = (X_1 - \bar{X}_{\text{text}} ... X_n - \bar{X}_{\text{text}})^t$ where *text*=cases or controls, meaning $X_1..X_n$ is the coded genotype of the cases but $\bar{X}$ can be computed based on the cases, the controls, or both. Define

$$Z_i = \frac{\left(X_i - \bar{X}_{\text{control}}\right)\left(Y_i - Y_{\text{offset}}\right)}{\hat{\sigma_x}\sqrt{||T_\mu||^2 + 2(\bar{Y} - Y_{\text{offset}})^2}} \tag{12}$$

then

$$\begin{aligned}\sum_{i=1}^{n} Z_i &= \frac{X_{\text{control}}^t T}{\hat{\sigma_x}\sqrt{||T_\mu||^2 + 2(\bar{Y} - Y_{\text{offset}})^2}} \\ &= \frac{X_{\text{case}}^t T_\mu + n(\bar{X}_{\text{case}} - \bar{X}_{\text{control}})(\bar{Y} - Y_{\text{offset}})}{\hat{\sigma_x}\sqrt{||T_\mu||^2 + 2(\bar{Y} - Y_{\text{offset}})^2}}\end{aligned} \tag{13}$$

It is important to note that the Z_is are independent, $E(Z_i) = 0$ and $Var\left(\sum_{i=1}^{n} Z_i\right) = 1$, which is obtained by first taking the conditional expectation treating X as random and Y as fixed. The Lindberg condition [25] for Z_i, which ensures asymptotic normality of $\sum Z_i$, is then given by

$$\forall \epsilon > 0 : lim_{n \to \infty} \left\{ \sum_{i=1}^{n} \int_{|Z_i| \geq \epsilon} Z_i^2 dP \right\} = 0 \tag{14}$$

Since Z_i has a discrete distribution, the Lindberg condition can only be fulfilled when the integration set $\{|Z_i| \geq \epsilon\}$ is empty for $n \to \infty$. Since X is the coded genotype and Y is a biological quantity, assume $\hat{\sigma_x} \neq 0$, $\hat{\sigma_y} \neq 0$ and both are finite. Then, there exists some constant K such that $\frac{|(X_i - \bar{X}_{\text{control}})||T_i|}{\hat{\sigma_x}\hat{\sigma_y}} \leq K$. Hence we rewrite the Lindberg condition by

$$\begin{aligned}\forall \epsilon > 0 : \epsilon \leq |Z_i| &= \frac{\left|(X_i - \bar{X}_{\text{control}})\right| |T_i|}{\hat{\sigma_x}\sqrt{||T_\mu||^2 + 2(\bar{Y} - Y_{\text{offset}})^2}} \\ &\leq \frac{\left|(X_i - \bar{X}_{\text{control}})\right| |T_i|}{\hat{\sigma_x}||T_\mu||} \leq \frac{K}{n} \to 0 \,\text{as}\, n \to \infty\end{aligned} \tag{15}$$

Hence the integral in the Lindberg condition is always computed over a set that is empty for $n \to \infty$. Thus the Lindberg condition is always fulfilled when the regularity condition holds. Then the Lindberg theorem [26] implies convergence to normality. Then

$$\frac{||T||}{\sqrt{||T_\mu||^2 + 2(\bar{Y} - Y_{\text{offset}})^2}} Stat_{\text{NPBAT}} = \sum_{i=1}^{n} Z_i \to^d N(0,1) \tag{16}$$

Then the NPBAT statistic is normally distributed with mean zero and variance given above. Note that the variance is always greater than or equal to one and equals one when $Y_{\text{offset}} = E[Y]$. Note that if $Y_{\text{offset}} = \bar{Y}$ and $E_x = \bar{X}_{\text{controls}}$ then NPBAT has a standard normal distribution. As seen in the Simulations section and Figure 1, when E_X is based on the the controls and the phenotype information is only available for the cases, then the power is maximized when $Y_{\text{offset}} \neq \bar{Y}$ because the variance equals the minimum when $Y_{\text{offset}} \approx E[Y]$.

Appendix D: NPBAT software

A software package implemented in C++ to compute both single phenotype and multiple phenotypes NPBAT statistics is available for download at the following website: https://sites.google.com/site/genenpbat/. In addition to NPBAT statistics, other population based statistics such as the Armitage Trend Test, Fisher Exact Test are also available. Currently, only two platforms are supported: linux64 and windows64. The NPBAT software package reads in genetic data through the PLINK style pedigree (ped), map (map) and phenotype (phe) files. The website provides detail information on how to use the software package.

Abbreviations

COPD: Chronic obstructive pulmonary disease; FBAT: Family Based Association Test; FEV: Forced expiratory volume; FNDS: Fagerstrom nicotine dependence score; GWAS: Genome Wide Association Study; LRT: Likelihood ratio test; NPBAT: Nonparametric Population Based Association Test; PC: Principal component; SNPs: Single nucleotide polymorphisms.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SL derived the asymptotic distribution, performed the simulations studies and the data analysis. SL, WKY, JH, NL, and CL were involved in drafting the manuscript or revising it critically. CL made substantial contributions to conception of the method and assisted in the simulation studies. All authors read and approved the final manuscript.

Acknowledgements

We would like to acknowledge Carla Wilson at National Jewish Health for her help with the COPDgene dataset. This work was funded by NIH/ NHLBI U01 HL089856 Edwin K. Silverman, PI. COPDGene is supported by NHLBI Grant Nos U01HL089897 and U01Hl089856.

Author details

[1]Department of Biostatistics, University of Colorado Anschutz Medical Campus, Aurora, USA. [2]Department of Epidemiology, University of Colorado Anschutz Medical Campus, Aurora, USA. [3]Department of Biostatistics, Harvard School of Public Health, Boston, USA. [4]Channing Laboratory, Harvard Medical

School, Boston, USA. [5]Institute for Genomic Mathematics, University of Bonn, Bonn, Germany. [6]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany.

Received: 15 October 2012 Accepted: 1 February 2013
Published: 28 February 2013

References

1. Thomas DC: *Statistical Methods in Genetic Epidemiology*. 2nd edn, Vol. 1. Oxford: Oxford University Press; 2004.
2. Wang J, Shete S: **Power and type I error results for a bias-correction approach recently shown to provide accurate odds ratios of genetic variants for the secondary phenotypes associated with primary diseases.** *Genet Epidemiol* 2011, **35:**739–743.
3. Wang J, Shete S: **Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases.** *Genet Epidemiol* 2011, **35:**190–200.
4. Richardson DB, Rzehak P, Klenk J, Weiland SK: **Analyses of case control data for additional outcomes.** *Epidemiol* 2007, **18**(4):441–445.
5. Li H, Gail MH, Berndt S, Chatterjee N: **Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies.** *Genet Epidemiol* 2010, **34:**427–433.
6. Li H, Gail MH: **Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies.** *Hum Hered* 2012, **73**(3):159–173.
7. Monsees GM, Tamimi RM, Kraft P: **Genome-wide association scans for secondary traits using case-control samples.** *Genet Epidemiol* 2009, **33:**717–728.
8. Greenland S: **Quantifying biases in causal models: classical confounding vs collider–stratification bias.** *Epidemiol* 2003, **14**(3):300–306.
9. He J, Li H, Edmondson AC, Rader DJ, Li M: **A Gaussian copula approach for the analysis of secondary phenotypes in case control genetic association studies.** *Biostatistics* 2011, **13**(3):497–508.
10. Kraft P: **Letter to the editor: analyses of genome-wide association scans for additional outcomes.** *Epidemiol* 2007, **18**(6):838.
11. Lin DY, Zeng D: **Proper Analysis of secondary phenotype data in case-control association studies.** *Genet Epidemiol* 2009, **33**(3): 256–265.
12. Laird NM, Horvath S, Xu X: **Implementing a unified approach to family based tests of association.** *Genet Epidemiol* 2000, **19:**S36.
13. Laird NM, Lange C: **Family-based methods for linkage and association analysis.** *Adv Genet* 2008, **60:**219-252.
14. Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM, PBAT: **Tools for family-based association studies.** *Am J Hum Genet* 2004, **74**(2): 367–369.
15. Lange C, DeMeo DL, Laird NM: **Power and design considerations for a general class of family-based association tests: the asymptotic distribution, the conditional power, and optimality considerations.** *Genet Epidemiol* 2002, **23**(2):165–180.
16. Lange C, Blacker D, Laird NM: **Family-based association tests for survival and times-to-onset analysis.** *Stat Med* 2004, **23**(2):179–189.
17. Lange C, Silverman EK, Xu X, Weiss ST, Laird NM: **A multivariate family-based association test using generalized estimating equations: FBAT-GEE.** *Biostatistics* 2003, **4**(2):195–206.
18. Lange C, Laird NM: **Power calculations for a general class of family-based association tests: dichotomous traits.** *Am J Hum Genet* 2002, **71**(3):575–584.
19. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch D, Silverman EK, Crapo JD: **Genetic epidemiology of COPD (COPDGene): study design and methods.** *COPD* 2010, **7**(1):32–43.
20. Kim DK, Hersh CP, Washko GR, Hokanson JE, Lynch DA, Newell JD, Murphy JR, Crapo JD, Silverman EK: **Epidemiology, radiology, and genetics of nicotine dependence in COPD.** *Respis Res* 2011, **12:**9–15.
21. Sha Q, Zhang Z, Zhang S: **An improved score test for genetic association studies.** *Genet Epidemiol* 2011, **35**(5):350–359.
22. Devlin B, Roeder K: **Genomic control for association studies.** *Biometrics* 2009, **55**(4):997–1004.
23. Price AK, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nature Genet* 2006, **38**(8):904–909.
24. Falconer DS, Makcay TFC: *Introduction to Quantitative Genetics*. London: Longman; 1997.
25. Serfling R: *Approximation Theorems of Mathematical Statistics*. New York: Wiley; 1980.
26. Billingsley P: *Probability and Measure*. New York: Wiley; 1995.

doi:10.1186/1471-2156-14-13
Cite this article as: Lutz *et al.*: A general semi-parametric approach to the analysis of genetic association studies in population-based designs. *BMC Genetics* 2013 **14**:13.