Single- and Multi-Locus Association Tests Incorporating Phenotype Heterogeneity

Hatef Darabi Keith Humphreys

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Key Words

Genetic association • Multi-locus test • Phenotype heterogeneity

Abstract

Taking disease subtypes into account when testing for an association between genetic factors and disease risk may help to identify specific aetiologic pathways. One way to assess a genetic association, whilst accounting for heterogeneity, is to use polytomous regression. This approach only allows heterogeneity to be considered in terms of a single categorical variable. In this article, we describe an alternative and novel test of association which incorporates multivariate measures of categorical and continuous heterogeneity. We describe both a single-SNP and a global multi-SNP test and use simulated data to demonstrate the power of the tests when genetic effects differ across disease subtypes. Applying the tests to the study of genetic variation in the oestrogen metabolic pathway and its association with breast cancer risk and prognosticators strengthened our understanding that the modulation of aromatase activity can influence the occurrence of tumours, and their grade and size, in postmenopausal women.

Copyright © 2011 S. Karger AG, Basel

Introduction

Advances in genotyping technology have stimulated the utilisation of high-density genetic markers, such as single nucleotide polymorphisms (SNPs), in the fine-mapping of genes across a wide range of disease phenotypes. This has accelerated the use of association methods for dissecting the genetic mechanism of complex traits [1, 2]. For many common diseases, an effective approach for detecting associated common variants of modest effect size is to compare the allele or genotype frequencies at markers in affected individuals with those in controls. For case-control studies, logistic regression is commonly used, modelling risk as a function of marker, often with adjustment for covariates. Implicitly homogeneity of the disease is assumed. Genetic variants under study may, however, contribute different effects to different subtypes of the disease. In such situations, association tests which account for phenotype heterogeneity will be more statistically powerful than those which do not. Heterogeneity may be attributable to either unknown/unobserved disease subtypes or to subtypes of cases that are distinguishable to the researcher. For the former scenario, Zhou and Pan [3] have proposed a test of association which is based on a binomial mixture model, which assumes that disease-associated SNPs are associated with one specific

(unknown/unobserved) subtype of disease. In this article, we concentrate on heterogeneity which is attributable to subtypes of cases that are distinguishable to the researcher. Examples used to illustrate the statistical methodology described in this article are taken from breast cancer research. Breast cancer heterogeneity has been discussed by Bertucci and Birnbaum [4]. It has been suggested, for example, that oestrogen receptor (ER)-positive and ER-negative breast cancers are biologically different diseases [5]. Moreover several of the SNPs which have recently been identified as being associated with breast cancer risk, through genome-wide association studies, have significantly different risk estimates for ER-positive and ER-negative disease [5].

When testing for association, although characteristics which may distinguish subtypes of cases may be known, the degree of heterogeneity is generally unknown. Several tests can be performed, based on, for example, logistic regression (ignoring heterogeneity), analysis stratified by subtype, test of heterogeneity (cases only), or polytomous regression (testing of association allowing for heterogeneity). Morris et al. [6] have shown that, across a wide range of scenarios (in terms of heterogeneity), polytomous regression has a small loss in power compared to the most powerful test, whilst the power of the other tests can vary greatly. Hence a test of association, allowing for heterogeneity, can be robust to different levels/forms of heterogeneity and provide a good summary analysis (negating the need to perform several tests and combine the results). Morris et al. [6] have analysed a type 2 diabetes data set and defined heterogeneity of the disease in terms of obesity in order to recognise that there are differential mechanisms underlying obese and non-obese forms of type 2 diabetes.

In this article, we propose an alternative approach to polytomous regression for testing association while accounting for heterogeneity, which is based on modelling the distributions of genotypes, conditional on disease (subtype) status. We first describe a simple test of association for single-SNP testing. We subsequently extend the test to evaluate the association of disease with several SNPs jointly. In genes with related functions, if several SNPs are weakly associated with the disease, there may be insufficient power to detect any single SNP-disease association. However, by studying the test's statistic distribution across all SNPs within the group, it may be possible to detect deviation from the null hypothesis of 'no SNPs in the group are associated with the disease' [7–9].

There are many ways to test for multi-marker association, ignoring phenotype heterogeneity, based on com-

binning single-marker test statistics. One way, based on Wilkinson's test, is to use a statistics which is the number of single-marker test results that are significant at the 5% level. Another example is the rank truncated product test, which has been described for SNP association analysis by Dudbridge and Koeleman [10]. Tyrer et al. [11] recently proposed an admixture maximum likelihood (AML) approach, which assesses the distribution of Cochran-Armitage test statistics for an association with a disease (yes/no) and simultaneously estimates both the proportion of associated SNPs and their (assumed common) effect size on the outcome of the disease of interest. The AML test was compared to a number of alternative global tests of association and was found to perform favourably in terms of statistical power across a range of disease-generating mechanisms [11]. Because the AML approach performs well, in the absence of heterogeneity, we have developed our (single-marker) approach to the multi-marker setting using a mixture model approach. In our approach, phenotype heterogeneity can be handled straightforward, because we model the distribution of genotypes conditional on observed disease (subtype) status. Motivation for developing our multi-marker test, incorporating heterogeneity, came from our genetic association study of breast cancer based on genes involved in oestrogen metabolism [12].

We begin by describing our novel association test incorporating heterogeneity, both in single-SNP and multi-SNP settings ('Methods'). Then in the next section, we describe simulation studies carried out to evaluate the performance of our tests, and in 'A Genetic Association Analysis of Breast Cancer Based on Genes Involved in Oestrogen Metabolism', we apply our (single-SNP and multi-SNP) tests to re-analyse the association of (SNPs in) genes involved in oestrogen metabolism with breast cancer risk using previously analysed data [12, 13].

Methods

The association between SNP genotypes and a case-control status is commonly assessed using the 1 degree of freedom Cochran-Armitage trend test, which is equivalent to performing a score test on the coefficient of a genotype covariate in a logistic regression model, when the genotype is assigned a genetic score such as $x = (0, 1, 2)$ and treated as continuous [14]. The use of this additive model on the logit scale is equivalent to assuming that allelic effects are multiplicative on the genotype relative risk scale ('multiplicative penetrance'). Logistic regression can be extended to handle phenotypes that are represented by more than two categories. Suppose that cases can be divided into $K - 1$ subgroups. Let Y denote the disease status, with $Y = 1$ representing control status and

$Y = 2, \dots, K$ representing the $K - 1$ disease subgroups. For a single covariate/genetic marker X , a polytomous regression model can be used to express the probability of belonging to class j as

$$P(Y = j|X = x) = \frac{e^{\alpha_j + \beta_j x}}{\sum_{h=1}^K e^{\alpha_h + \beta_h x}}, j = 1, \dots, K, \quad (1)$$

where, for identifiability, $\alpha_1 = \beta_1 = 0$ (fixed). Morris et al. [6] have proposed the use of a likelihood ratio test of association, based on equ. 1, to test for association, whilst allowing for heterogeneity of genetic effect across disease subgroups.

A Single-SNP Association Test Incorporating Heterogeneity

We propose an alternative approach for testing of association, allowing for heterogeneity, based on modelling the distribution of marker genotype, conditional on observed disease (subtype) status. If a disease-associated SNP is in Hardy-Weinberg equilibrium (HWE) in the general population, with minor allele frequency (MAF) p , then it is straightforward to show, for a rare disease and under multiplicative penetrance, that controls will be approximately in HWE with MAF p and cases will be approximately in HWE with MAF $pe^\mu/(1 - p + pe^\mu)$, where μ is the log odds ratio (OR) [15]. We used this result in developing our test. We assumed that HWE holds both within controls and (subgroups of) cases, which is approximately true for a (moderately) rare disease under multiplicative penetrance. We let $\mathbf{X} = \{X_1, \dots, X_N\}$, $\mathbf{Y} = \{Y_1, \dots, Y_N\}$ and $\mathbf{W} = \{W_1, \dots, W_N\}$ be random variables denoting genotype, disease status and disease subtype, respectively, for a total of N individuals. Let $X_i = x_i \in \{0, 1, 2\} \forall i$ denote the genotype score for individual i , $Y_i = y_i$ denote the case-control status of individual i ($y_i = 0$ if i is a control, $y_i = 1$ if i is a case), and $W_i = w_i$ denote the individual disease subtype. W_i is defined only for cases (we arbitrarily assigned a value of 0 to W_i when subject i is a control). Considering only two subtypes for now, we allow w_i to take values 0 or 1. We base inference on the conditional distribution of genotypes given control/case (subtype) status. For data on individuals $i = 1, \dots, N$, we base inference on the likelihood

$$P(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y}, \mathbf{W} = \mathbf{w}) = \prod_{i=1}^N P(X_i = x_i | Y_i = y_i, W_i = w_i) \quad (2)$$

$$= \prod_{i=1}^N \frac{(2 - (x_i - 1)^2) e^{x_i(\delta + y_i(\mu_0 + \mu_1 w_i))}}{(1 + e^{\delta + y_i(\mu_0 + \mu_1 w_i)})^2}, \quad (3)$$

where x_i represents the genotype score, y_i represents case-control status and w_i represents the disease subtype (cases only) of individual i , and $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{w}$ are vectors of length N . The log ORs for disease subtypes 0 and 1 are represented by μ_0 and $(\mu_0 + \mu_1)$, respectively. For $K - 1$ subtypes, when $K - 1 > 2$, we can use $K - 2$ binary dummy variables in place of the single variable W , replacing $\mu_i w_i$ in equ. 3, by $\mu w_{1i} + \dots + \mu_{(K-2)i} w_{(K-2)i}$. An association test incorporating heterogeneity can be based on testing $\mu_0 = \mu_1 = 0$, using a likelihood ratio test. The relevant test statistic is asymptotically χ^2 distributed with $K - 1$ degrees of freedom. To understand the relative contribution of different disease subtypes on the overall association, it is useful to obtain parameter estimates (together with their standard errors) under the alternative hypothesis. We do this by estimating the observed Fisher information matrix using the negative of the second derivative (the Hessian) matrix of the log likelihood.

We note that for a single $K - 1$ categorical variable W and under the rare disease assumption, as long as multiplicative penetrance holds, our test is approximately equivalent to the polytomous regression model approach described by Morris et al. [6]. The main advantages of our approach are that (i) it conveniently allows an extension to more complex forms of phenotype heterogeneity than is represented by a single categorical variable, and (ii) it can be conveniently adapted for multi-marker testing (below). More than one phenotype heterogeneity characteristic at a time can be handled conveniently by adding new covariates into the 'linear predictor' $(\mu_0 + \mu_1 w_i)$ in equ. 3. For large numbers of disease subtypes or case characteristics measured on a continuous scale (e.g. tumour size in millimetres), powerful tests of association, allowing for heterogeneity, can be obtained by including a continuous variable W in equ. 2. It can be shown that if w_i in equ. 2 is a realisation of a continuous random variable, then the model is equivalent to the polytomous regression model in equ. 1, when $\beta_j - \beta_{j-1}$ is constant for all $j > 1$.

A Multi-SNP Association Test Incorporating Heterogeneity

For sets of SNPs (for example, in genes whose biological functions are related), Tyrer et al. [11] proposed an AML approach to test a global null hypothesis of 'no SNP is associated with the disease'. The test is based on examining the distribution of Cochran-Armitage test statistic values. Under the null hypothesis, the square root of the test statistic follows a standard normal distribution. Under the alternative hypothesis, there exists a proportion of SNPs (α) associated with the disease, and the remaining $(1 - \alpha)$ SNPs are not associated. Under this alternative hypothesis, the Cochran-Armitage test statistic for each SNP will be non-centrally χ^2 distributed with 1 degree of freedom. The AML test assumes the non-centrality parameter to be equal for all associated SNPs, which is equivalent to assuming an equal recurrence rate ratio in the offspring of an affected individual (which can be represented as a function of the SNPs MAF and OR [11]). For the global test, under the alternative hypothesis of association, a mixture of normal distributions is fitted to the distribution of the square root of the Cochran-Armitage test statistic values. To account for both positively and negatively associated SNPs (a priori it is not known which of the two alleles is associated with an increased risk), 3 mixture components are assumed, one representing non-associated SNPs, the other two representing negatively and positively associated SNPs. The negatively and positively associated SNPs are assumed to be equally likely, and the global null hypothesis is assessed by a likelihood ratio test. Typically, some of the SNPs being tested will be in linkage disequilibrium (LD). Because this is not explicitly recognised in the likelihood, the resulting test is termed a 'pseudo-likelihood' ratio test and empirical p values are obtained by simulation. The AML test has been shown to perform well in comparison to other global tests of association (e.g. Wilkinsons's test) across a range of disease-generating mechanisms [11].

The AML test would be cumbersome to extend to incorporate heterogeneity. In principle, the Cochran-Armitage test could be replaced by a likelihood ratio test based on a polytomous regression model; and under the global alternative hypothesis, a mixture of non-central χ^2 distributions could be fitted to the distribution of the SNP test statistics. This approach would, however, have several disadvantages compared to the approach we describe below. For example, extending the AML in this way, it would be unclear what it means to assume that the non-centrality parameter

has the same value for all associated SNPs. Also, it would be difficult to extend the approach to incorporate more than one categorical variable for describing phenotype heterogeneity.

Instead of extending the AML, we take an alternative approach, based on extending the single-SNP test (described above in 'A Single-SNP Association Test Incorporating Heterogeneity'). We model the individual level genotype data, conditional on control/case subgroup status, and assume equal ORs across all associated SNPs. We introduce a vector of latent variables, $\mathbf{V} = \{V_1, \dots, V_M\}$, to indicate which of the M markers are associated with the disease; V_j takes value 0 if SNP j is not associated and ± 1 if it is positively (+1) or negatively (-1) associated with the disease. We combine the conditional probabilities of genotypes of each individual SNP into a single likelihood as

$$\sum_{\{\mathbf{v}\}} P(\mathbf{X} = \mathbf{x} | \mathbf{V} = \mathbf{v}, \mathbf{Y} = \mathbf{y}, \mathbf{W} = \mathbf{w}) P(\mathbf{V} = \mathbf{v}), \quad (4)$$

where $\mathbf{v}$ is a vector of values ($v_1, \dots, v_M$) and $\{\mathbf{v}\}$ represents all configurations of $\mathbf{v}$. If we assume $P(\mathbf{V} = \mathbf{v}) = \prod_{j=1}^M P(V_j = v_j)$ and that the SNPs are in linkage equilibrium, then taking the logarithm of equ. 4, we obtain

$$l_A = \sum_{j=1}^M \log \left\{ \sum_{v_j=-1}^1 P(V_j = v_j) \prod_{i=1}^N P(X_{ij} = x_{ij} | W_i = w_i, Y_i = y_i, V_j = v_j) \right\}, \quad (5)$$

where

$$P(X_{ij} = x_{ij} | W_i = w_i, V_j = v_j, Y_i = y_i) = \frac{(2 - (x_{ij} - 1)^2) e^{x_{ij}(\delta_j + v_j y_i (\mu_0 + \mu_1 w_i))}}{(1 + e^{\delta_j + v_j y_i (\mu_0 + \mu_1 w_i)})^2}, \quad (6)$$

for $v_j \in \{-1, 0, 1\}$. In principle, we can specify a parametric form for $P(V_j = v_j)$, for example,

$$P(V_j = v_j) = \frac{I(v_j = 0)}{1 + e^{\xi_{1j}} + e^{\xi_{2j}}} + \frac{I(v_j = -1) e^{\xi_{1j}}}{1 + e^{\xi_{1j}} + e^{\xi_{2j}}} + \frac{I(v_j = 1) e^{\xi_{2j}}}{1 + e^{\xi_{1j}} + e^{\xi_{2j}}}, \quad (7)$$

and optimise equ. 5. If the values of ξ_{1j} and ξ_{2j} are not specified, but are instead included as parameters, along with μ_0 , μ_1 and δ_j for $j \in \{1, \dots, M\}$ over which equ. 5 should be optimised, then the optimal value of equ. 5 is

$$\sum_{j=1}^M \max_{v_j \in \{-1, 0, 1\}} \sum_{i=1}^N \log p(X_{ij} = x_{ij} | W_i = w_i, Y_i = y_i, V_j = v_j), \quad (8)$$

which turns out to be computationally straightforward to evaluate and optimise, with respect to the remaining parameters μ_0 , μ_1 and $\{\delta_j\}_{j=1}^M$. By not specifying the $2M$ values of $\xi \in \mathcal{R}_1^2 \otimes \dots \otimes \mathcal{R}_M^2$, we bypass the need to optimise over the parameters in $P(\mathbf{V} = \mathbf{v})$. At first glance, optimisation of equ. 8 is computationally impractical even for moderate numbers of SNPs, since it includes $M + 2$ (dependent) parameters, the marker-specific δ s and common effect size measures μ_0 and μ_1 . In order to reduce the dimension of the parameter space, we use a reparameterisation. The MAFs in the controls (p_j) that determine the MAFs in the cases, in combination with a given effect size (assumed to be equal for all associated SNPs), can be easily obtained by solving an equation for the combined MAFs, which we denote by r_j , under the different conditions depending on whether each SNP is positively, negatively, or not associated. Similarly to the AML test, we allow positively

and negatively associated SNPs and assume symmetry of effect sizes around an OR of 1. Suppose a log OR of $\mu_0 > 0$ for a first group of cases for all associated SNPs, and a log OR of $\mu_0 + \mu_1$ for a second group of cases for all associated SNPs. We can also suppose that $\mu_1 > 0$, which would be logical if increasing values of W correspond to increasing severity of disease. A priori we do not know whether a SNP is non-associated or in case of association if it is positively or negatively associated, since we do not know which allele is the high-risk allele (0 or 1) for each associated SNP. If a SNP is not associated ($v_j = 0$), then $\delta_j = \log(r_j/(1 - r_j))$ represents the logit transform of the joint MAF. If a SNP is associated and '1' is the high-risk allele, we say the SNP is positively associated and the expected combined MAF in a sample of N_0 controls and N_1 cases is

$$r_j = \frac{1}{N} \left[N_0 p_j^{(+)} + \sum_{i=1}^{N_1} \frac{p_j^{(+)} e^{(\mu_0 + \mu_1 w_i)}}{1 - p_j^{(+)} + p_j^{(+)} e^{(\mu_0 + \mu_1 w_i)}} \right], \quad (9)$$

where $N = N_0 + N_1$. If a SNP is associated and '0' is the high-risk allele, we say the SNP is negatively associated and the expected combined MAF is

$$r_j = \frac{1}{N} \left[N_0 p_j^{(-)} + \sum_{i=1}^{N_1} \frac{p_j^{(-)} e^{-(\mu_0 + \mu_1 w_i)}}{1 - p_j^{(-)} + p_j^{(-)} e^{-(\mu_0 + \mu_1 w_i)}} \right]. \quad (10)$$

If the combined MAFs, r_j ($j \in \{1, \dots, M\}$), are known together with sample sizes and knowledge about whether each SNP is (positively or negatively) associated or not associated, then equ. 8 only needs to be optimised with respect to two parameters, μ_0 and μ_1 , and can be written as

$$\sum_{j=1}^M \max_{v_j \in \{-1, 0, 1\}} \sum_{i=1}^N \log \frac{(2 - (x_{ij} - 1)^2) e^{x_{ij} \eta_j}}{(1 + e^{\eta_j})^2}, \quad (11)$$

where

$$\begin{aligned} \eta_j &= (1 - v_j^2) \delta_j \\ &+ I\{v_j = 1\} (\text{logit}(p_j^{(+)}) + y_i (\mu_0 + \mu_1 w_i)) \\ &+ I\{v_j = -1\} (\text{logit}(p_j^{(-)}) - y_i (\mu_0 + \mu_1 w_i)). \end{aligned} \quad (12)$$

We 'condition' on the observed combined MAFs and use a numerical optimisation routine to localise the stationary point of the log likelihood, and within the call to each optimisation step, for the current estimated value of $\{\mu_0, \mu_1\}$, solve equ. 9 and 10 for $p_j^{(+)}$ and $p_j^{(-)}$. The equations to solve are non-linear, and are solvable using Brent's method for root finding [16], where we are guaranteed a solution in the unit interval by the intermediate value theorem.

We obtain our global test of association allowing for heterogeneity by fitting the model for the conditional distribution of genotypes to the data both under the null and the alternative hypothesis and recording the optimal log-likelihood values. The relevant test statistic is then twice the difference in these log-likelihood values. Under the null hypothesis, closed form solutions for the unknown parameters δ_j ($j \in \{1, \dots, M\}$) exist as the logit of the joint MAF for each marker. Since SNPs will typically not be in linkage equilibrium, the likelihood (in equ. 5) is in fact a pseudo-likelihood. We ignore LD and optimise the pseudo-likelihood

in the same way as is done in the AML approach. In order to assess significance, we use simulation. Empirical levels of significance are obtained by permuting control/disease subgroup status over a large number of simulations. The empirical *p* value is obtained as the fraction of test statistic values based on permuted data which are larger than the test statistic value obtained for the non-permuted data set.

Simulation Studies

Single-SNP Tests

We performed several simulations with the objective of evaluating the performance of our association test under different scenarios. Our single-SNP test for a single categorical covariate is very similar to the polytomous regression-based test described by Morris et al. [6]. To recapitulate the general result of Morris et al. [6], we first carried out a simulation study similar to that described in their article, based on our single-SNP test. We considered a disease for which cases are categorised according to two subtypes, which could, for example, represent the ER status of tumours of breast cancer patients (see ‘A Genetic Association Analysis of Breast Cancer Based on Genes Involved in Oestrogen Metabolism’ below). Under a multiplicative disease model for each disease category, we examined a wide range of association scenarios ranging in a spectrum of disease heterogeneity. Case-control data was simulated by first generating a single-SNP genotype (in HWE) for a large number of individuals and subsequently generating the disease status from a multinomial distribution with 3 categories: the baseline category representing controls and the subsequent 2 categories representing subgroups of cases. Categories of disease were generated such that at baseline (homozygous wild-type) each was equally likely, with a total probability of disease around one percent. The effect of the marker for each generated data set was held constant for the second disease category (S2) such that the natural logarithm of the per-allele OR was 0.1, whereas the effect size for the SNP for the first disease category (S1) was allowed to vary, with log ORs ranging from -0.5 to $+0.5$. From the simulated individuals, we randomly selected 2,000 controls and 2,000 cases to compile the simulated case-control data set. For each scenario, in terms of ORs, we generated 1,000 data sets as described above, and for each simulated data set the ‘population’ MAF was chosen at random from a uniform (0.05, 0.5) distribution. For each simulated data set, we performed the following tests of association:

- All samples – our novel test of heterogeneity: test of association of the SNP with disease, allowing heterogeneity of per-allele ORs between subtypes of disease.
- All samples – logistic: likelihood ratio test of association of the SNP with disease within a logistic regression framework. This test assumes the genetic effect to be the same for both disease subtypes.
- S1 versus controls and S2 versus controls – logistic: tests of association of the SNP stratified by case subtype. Logistic regression (likelihood tests) were used to compare the SNP MAF in subgroups of cases (separately) to controls.
- S1 versus S2 – logistic: a case only analysis of heterogeneity of the effect of the causal SNP between subtypes of disease. Logistic regression with case subgroup as outcome and SNP as covariate (likelihood ratio test).

The results are summarised in figure 1. Empirical power curves (calculated as the proportion of simulated data sets for which the test was significant at the 5% level) are plotted for the 5 tests of association. The vertical line represents the scenario of no heterogeneity. At this point on the x-axis, the most powerful test of association is the test ignoring heterogeneity (‘All samples – logistic’). Moving away from this scenario, in either direction along the x-axis, as the level of heterogeneity increases, other tests become more powerful. The test of association which incorporates heterogeneity (‘All samples’) is consistently one of the most powerful tests, and therefore is useful for reporting a single *p* value, instead of performing a number of different tests and adjusting for multiple testing.

Our approach assumes a multiplicative penetrance model for associated SNPs. To examine how our approach performs under departures from this model assumption, we carried out a simulation study generating data under alternative genetic models. We again assumed the existence of two disease subtypes. As above, the disease subtypes were generated such that at baseline (homozygous wild-type) they were equally likely, with a total probability of disease around one percent. We considered 5 penetrance models, ranging from a recessive to a dominant penetrance model. The log ORs comparing 2 copies of the disease allele against 0 copies were assumed to be the same for all 5 penetrance models. For disease subtype 1, the log ORs (2 vs. 0 rare alleles) was $\ln(1.1)$, whilst for disease subtype 2, the log ORs (2 vs. 0 rare alleles) was $\ln(1.3)$. For the *j*-th penetrance model ($j = 1, \dots, 5$), the log ORs comparing 1 copy of the disease allele against 0 copies were assigned a value k_j times the value of the log OR comparing 2 versus 0 rare alleles. The values of k_j were

Fig. 1. Empirical power (at 5% significance level) of 5 tests of association as a function of phenotype heterogeneity.

chosen as $k_1 = 0$, $k_2 = 0.25$, $k_3 = 0.5$, $k_4 = 0.75$ and $k_5 = 1$, so that the first penetrance model is recessive, the third represents multiplicative penetrance and the fifth is a dominant penetrance model. The penetrance of the second model is between recessive and multiplicative penetrance, and the fourth between dominant and multiplicative penetrance. Aside from the genetic model used to simulate case subgroup/control status, we used the same approach as described earlier in this section to generate 1,000 data sets of 2,000 controls and 2,000 cases.

We performed 6 tests for each simulated data set: (i) our novel association test incorporating heterogeneity, (ii) logistic regression (ignoring heterogeneity) with genotypes coded 0, 1, 2 and treated as a continuous covariate, (iii) polytomous regression (i.e. an association test incorporating heterogeneity) with a genotype coded 0 if homozygous for the common/wild-type allele and otherwise coded 1, (iv) logistic regression with a genotype coded 0 if homozygous for the common/wild-type allele and otherwise coded 1, (v) polytomous regression with a genotype coded 1 if homozygous for the rare allele and otherwise coded 0, and (vi) logistic regression with a genotype coded 1 if homozygous for the rare allele and otherwise coded 0. Tests (i) and (ii), therefore, are based on multiplicative penetrance assumptions, whilst tests (iii) and (iv) are based on dominant penetrance, and tests (v) and (vi) are based on recessive penetrance models. We noted that although Morris et al. [6] discussed the use of polytomous regression under recessive and dominant pene-

trance assumptions, they did not conduct any simulation studies based on these models. All tests carried out in our simulation study were performed as likelihood ratio tests. Empirical power was calculated as the proportion of simulated data sets for which the test was significant at the 5% level. Results are presented in table 1. When data is generated under recessive or dominant penetrance models, the tests based on the correct penetrance assumptions clearly perform best, and for the effect sizes considered in this simulation, the tests incorporating heterogeneity have greatest power. Under multiplicative penetrance and the 2 'hybrid' scenarios, our association test incorporating heterogeneity performs well, with either best or close to best power. In situations where disease loci are not included in the studies but only markers in (incomplete) LD, our approach will perform well. Even if a dominant/recessive model is appropriate for finding the disease-causing locus, SNPs in (incomplete) LD will be associated with penetrance approaching the multiplicative model.

In order to highlight the advantage our approach has, i.e. being able to handle continuous measures of heterogeneity, we carried out a further simulation study based on our single-SNP test. We simulated data under a scenario where cases are divided into larger numbers of subgroups and performed 'trend' tests, treating subgroup as a continuous variable. We first considered situations with S subgroups of cases such that the natural logarithm of the per-allele OR was 0.04 for subgroup 1, and 0.22 for

Table 1. Empirical power (at 5% significance level) of single-marker tests under various genetic models with phenotype heterogeneity

	log ORs (1 vs. 0, 2 vs. 0 rare alleles)				
	(0,L)	(0.25L,L)	(0.5L,L)	(0.75L,L)	(L,L)
Multiplicative penetrance models					
(i) our test*	0.23	0.36	0.50	0.65	0.79
(ii) logistic regression	0.21	0.34	0.46	0.62	0.75
Dominant penetrance models					
(iii) polytomous regression	0.08	0.22	0.40	0.66	0.87
(iv) logistic regression	0.08	0.22	0.39	0.64	0.82
Recessive penetrance models					
(v) polytomous regression	0.45	0.35	0.28	0.21	0.15
(vi) logistic regression	0.41	0.31	0.28	0.21	0.15

L = ln(1.1) for disease subtype 1, ln(1.3) for disease subtype 2.

* For all scenarios, empirical power was the same (up to two decimal places) for our test and for the test based on polytomous regression (assuming multiplicative penetrance).

Table 2. Empirical power (at 5% significance level) of single-marker tests in relation to several disease subgroups

	ln OR ₁ = 0.04, ln OR _S = 0.22			ln OR ₁ = -0.15, ln OR _S = +0.15		
	S = 4	S = 8	S = 16	S = 4	S = 8	S = 16
(i) our test, continuous	0.87	0.82	0.80	0.84	0.72	0.66
(ii) our test, S groups	0.78	0.61	0.46	0.73	0.46	0.29
(iii) our test, 4 groups	0.78	0.72	0.71	0.73	0.59	0.52
(iv) our test, 2 groups	0.85	0.79	0.77	0.75	0.60	0.53
(iii) logistic regression	0.78	0.77	0.79	0.05	0.05	0.05

ln OR₁ and ln OR_S are log_e ORs for disease subgroups 1 and several (S), respectively. log_e ORs for disease subgroups $j = 2, \dots, S - 1$ were specified as $\ln OR_1 + (j - 1)(\ln OR_S - \ln OR_1)/(S - 1)$.

subgroup S. For intermediate subgroups, the logarithm of the per-allele OR increased by a constant increment of $0.18/(S - 1)$. We also considered situations with S subgroups of cases such that the natural logarithm of the per-allele OR was -0.15 for subgroup 1, and +0.15 for subgroup S. For intermediate subgroups, the logarithm of the per-allele OR increased by a constant increment of $0.3/(S - 1)$. For both types of scenarios, we allowed the number of subgroups S to be 4, 8 and 16. For each simulation setting we again used the same approach as in our first simulation (aside from the underlying disease model), generating 1,000 data sets of 2,000 controls and 2,000 cases. We performed 6 tests for each simulated data set. All tests assumed multiplicative penetrance. The first test (i) was our novel association test incorporating heterogeneity characterised by a single covariate for subgroup

treated as continuous. Tests (ii–iv) were based on our novel association test treating subgroup as a categorical variable; test (ii) had S categories, test (iii) had 4 categories, each containing (S/4) ‘adjacent’ disease subgroups, and test (iv) had 2 categories (the first and last S/2 disease subgroups). Tests based on polytomous regression with S, 4 or 2 groups could have alternatively been used instead of tests (ii–iv) and would have obtained similar results. Test (v) was based on logistic regression (ignoring heterogeneity). In all settings, our test of association with heterogeneity treated on a continuous scale is clearly more powerful than all other tests (table 2).

Multi-SNP Tests

We performed a number of simulation studies to evaluate the performance of our multi-SNP test of associa-

tion. In the absence of phenotype heterogeneity, we expect our global test (without specifying heterogeneity, i.e. $\mu_1 = 0$) to perform similarly to the AML test. The biggest difference between the tests is that under the alternative hypothesis the AML test assumes that associated SNPs have a common recurrence rate ratio to the offspring of an affected individual, whilst we assume associated SNPs to have a common per-allele OR. These two assumptions are equivalent if all associated SNPs have the same MAF. To see just how similar the performances of the two tests are, we carried out a small simulation study based on generating 300 cases and 300 controls. We first generated 20 SNP genotypes in linkage equilibrium (with different MAFs, drawn from a Uniform (0.05, 0.5) distribution) for a large number of individuals, and subsequently generated a (moderately rare) disease status, such that 8 out of the 20 SNPs were associated with disease, each with a per-allele OR of 1.3 (we assumed a probability of disease of approximately one percent for individuals with the common homozygote genotype on all 8 associated SNPs). The 300 cases and 300 controls were selected randomly from the larger pool of cases and controls, generated as described above. 500 data sets were generated in this manner and empirical power curves were obtained for the AML and for our novel association test (with no heterogeneity specified). The empirical power curves of the two tests were almost indistinguishable from each other (data not shown). We repeated the simulation, but instead of specifying equal ORs for the 8 associated SNPs, we instead generated data under the assumption that the 8 associated SNPs had a common recurrence rate ratio. The empirical power curves of the two approaches were once again extremely close to each other (data not shown).

We next performed a simulation study based on generating repeated data sets in a similar way to that described above, i.e. with 20 SNPs of which 8 are associated, but now with phenotype heterogeneity. Cases were classified as belonging to one of two subgroups. We considered several scenarios of heterogeneity. For each scenario, the cases in the first subgroup were generated such that all of the 8 associated SNPs had the same OR, but the OR was allowed to vary between 1 and 1.3 across the different scenarios of heterogeneity. Cases in the second subgroup were generated such that all of the 8 associated SNPs had an OR of 1.3 for all scenarios. For each heterogeneity scenario, we generated 500 data sets and estimated the empirical power to detect global association at a 5% level of significance using our novel test of association, both allowing heterogeneity and ignoring heterogeneity (for the latter we could,

alternatively, have used the AML test – the two tests perform similarly in terms of power). To assess the particular form of our multi-marker test statistic, i.e. our mixture model approach, we also tested for (global) association, incorporating heterogeneity, using a test statistic based on Wilkinson's test – we evaluated the number of single-marker test statistics from our association test accounting for heterogeneity with (likelihood ratio) p values $< 5\%$. p values for all three multi-marker tests were obtained by permuting the case-control status (500 times). Empirical powers of the multi-marker tests allowing for heterogeneity (our mixture model and the Wilkinson's test based approach) and the multi-marker test which ignores heterogeneity are plotted in figure 2a. As with the single-marker tests in the absence of heterogeneity, the test incorporating heterogeneity is less powerful than the test ignoring heterogeneity, but in the presence of moderate/strong heterogeneity, substantial gains in power are achieved by allowing for heterogeneity in testing for association, particularly for our mixture-model-based approach. Our mixture-model-based test consistently outperforms the ' p value < 0.05 ' multi-marker test.

We then repeated the above simulation, but instead of generating cases and controls under the condition that all associated SNPs have the same OR under each level of heterogeneity, we allowed the ORs to vary. We used OR values drawn from normal distributions with a standard deviation of 0.05. Mean values were chosen to be the same as the fixed values used in the first multi-marker simulation study (fig. 2a). Also, in this simulation we allowed the mean OR value for subgroup 1 (across scenarios of heterogeneity) to vary between 1 and 1.25 (instead of between 1 and 1.3), which was sufficient for understanding the performances of the 3 multi-SNP tests. Results for the 3 multi-marker tests (our mixture-model-based test, our p value < 0.05 test and the test ignoring heterogeneity) are displayed in figure 2b. The improvements in power of our mixture model over both the p value < 0.05 test and the test ignoring heterogeneity in the presence of moderate/strong heterogeneity, which were seen in figure 2a, are still evident.

Finally, we performed a simulation study for the multi-marker tests, similar to those described above, but with the SNPs in LD. We generated a LD structure among the simulated genotypes using the procedure described by Jiang et al. [17]. We simulated 8 LD block, each containing 5 SNPs. The haplotypes in each block were generated by first simulating vectors from a multivariate normal distribution with a variance-covariance matrix with elements c_{ij} , and then thresholding the vector components

Fig. 2. Power of multi-SNP tests of association. **a** Associated SNPs with common ORs in linkage equilibrium. **b** ORs of associated SNPs distributed normally in linkage equilibrium. **c** ORs of associated SNPs distributed normally in LD.

according to pre-specified allele frequencies. Within each block, the LD structure was defined by specifying $c_{ij} = 0.8^{|i-j|}$, $i, j \in \{1, \dots, 5\}$. In each simulated data set, allele frequencies of the 40 markers were drawn from a uniform distribution on the interval $[0.2, 0.8]$. In this simulation, we let only 4 SNPs chosen as ‘middle’ SNPs from 4 different blocks be associated with disease. Disease status (and subgroup for cases) was generated with ORs drawn from normal distributions, in a similar way to the simulation presented in figure 2b. We assumed slightly larger effect sizes for the second subgroup (mean OR of 1.4 for the associated SNPs for the second subgroup) in order to obtain

levels of power similar to those observed in figures 2a and b. This was necessary because we removed the 4 SNPs that were directly associated with disease before testing for association in each simulated data set, leaving data sets with 36 SNPs, some of which are in indirect association with disease and disease subgroup status. We observed the same patterns of performance under LD as under linkage equilibrium. Because ORs for the second subgroup were higher than in earlier simulations (fig. 2a, b), the power curve of the no heterogeneity test crosses the power curves of the heterogeneity tests at a higher subgroup 1 OR value.

A Genetic Association Analysis of Breast Cancer Based on Genes Involved in Oestrogen Metabolism

A Swedish case-control study, CAHRES (Cancer And Hormone REplacement Study), was initiated in the early 1990s to examine the effect of menopausal hormone use on breast cancer risk. The nation-wide case-control study encompassed all Swedish-born women between 50 and 74 years of age and resident in Sweden between October 1993 and March 1995 with breast cancer. Approximately 1,500 breast cancer cases and 1,500 healthy controls were randomly selected from CAHRES and included in a molecular study in which 239 SNPs, distributed among 34 genes within the oestrogen metabolic 'pathway', were selected for genotyping [12]. By performing a progressive pathway-based association analysis from the whole metabolic pathway to 3 metabolic subpathways, using the AML approach, we were able to associate the metabolic pathway to breast cancer risk, and also to refine the association to the androgen-to-oestrogen conversion component of the pathway, in which we had typed 120 SNPs. We further demonstrated that genetic determinants relating to the androgen-to-oestrogen conversion are particularly important for the development of ER-positive tumours in postmenopausal women. A subset of the 1,500 cases and 1,500 controls, approximately 750 cases and 750 controls, was also included in a Swedish-Finnish genome-wide association scan [13], which highlighted 10 pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [18]. One of these pathways was the androgen and oestrogen metabolic pathway, containing 199 SNPs. Only 14 of the 239 SNPs included in the oestrogen metabolic pathway and 11 of the 120 SNPs included in the androgen-to-oestrogen pathway, typed on the 1,500 cases and 1,500 controls, were included amongst the 199 androgen and oestrogen metabolic SNPs typed in the genome-wide study (of course, many pairs of SNPs, made up of SNPs from different subsets, are in LD).

We first applied our novel multi-SNP association test incorporating heterogeneity in terms of ER status to the sample of cases and controls included in Low et al. [12], based on both the full oestrogen metabolic pathway (239 SNPs) and the androgen-to-oestrogen subpathway (120 SNPs). We also applied our multi-SNP test to the subset of approximately 750 cases and controls included in the genome-wide association scan [13], using the 199 SNPs involved in androgen and oestrogen metabolism. On the basis of our first single-marker simulation study, we argued that our association test incorporating heterogeneity provides a good summary (single) p value, avoiding

Table 3. Multi-locus tests of association for the oestrogen metabolic pathway and the androgen-to-oestrogen conversion subpathway

Subgroup	Cases	Controls	No hetero- geneity (AML)	Our hetero- geneity
<i>239 SNPs – oestrogen metabolic pathway study</i>				
ER-positive	811	1,527	0.012	0.004
ER-negative	220	1,527	0.785	
Both	1,031	1,527	0.136	
<i>120 SNPs – androgen-to-oestrogen subpathway study</i>				
ER-positive	811	1,527	0.002	0.005
ER-negative	220	1,527	0.693	
Both	1,031	1,527	0.007	
<i>199 SNPs – oestrogen metabolism GWAS study</i>				
ER-positive	473	764	0.005	0.006
ER-negative	108	764	0.847	
Both	581	764	0.012	

p values in columns 3 and 4 are based on 5,000 permutations.

the need to run several tests (e.g. subgroups versus controls, one-at-a-time or case-only analyses) and adjust for multiple testing. Here, for illustrative purposes we did, however, compare subgroups (separately) to cases. We performed AML tests based on comparing subgroups of cases (defined by ER status) with controls as well as comparing all cases with controls. For all 3 data sets, our test incorporating heterogeneity yields a p value close to the 'best' of the 3 AML tests (without adjustment for multiple testing). For the data set of 239 SNPs, our test incorporating heterogeneity yields a lower p value than the 3 AML tests (even prior to multiple testing adjustments; table 3).

A gene-based analysis of the androgen-to-oestrogen conversion subpathway, based on the sample of 1,500 cases and 1,500 controls, revealed CYP19A1 (aromatase) to be a major player in the association with breast cancer risk [12]. One of the SNPs, rs4646, in the 3'-untranslated region of this gene has been the subject of several breast cancer studies of prognosis, and its common variant has recently been found to be associated with an improved progression-free survival compared with carriers of the minor allele of this SNP in a sample of women treated with letrozole, an anti-aromatase therapy [19]. Testing of this SNP has been proposed as a predictive tool for breast cancer patients on anti-aromatase therapy. The p value for the Cochran-Armitage trend test of association (with

breast cancer risk) for this SNP was 0.008 in our study [12]. We recently found this SNP to also be associated with both tumour grade and tumour size amongst the cases in CAHRES [20], with p values of 0.002 and 0.030, based on fitting an ordinal logistic model and a linear regression model, respectively.

We used our novel single-SNP test of association, allowing for heterogeneity, to study the association of rs4646 with breast cancer risk, tumour grade and tumour size. For the current analysis, we performed 2 tests based on equ. 6, firstly with *W* representing grade, coded 0, 1, 2 and treated as continuous, and secondly with *W* representing tumour size, measured in centimetres and treated as continuous (table 4). The resulting p values of 9.1×10^{-5} and 6×10^{-3} can be viewed as p values combining previous results from case-control and case-only analyses, which were reported above (although different statistical modelling assumptions are made). We also carried out a test of association incorporating heterogeneity using both tumour grade and tumour size as covariates in a regression model for the conditional distribution of genotypes (last column in table 4). The major allele of rs4646 appears to be associated with an increased risk of breast cancer, particularly that of low grade and small tumour size (although when adjusting for grade, the association with tumour size was diminished).

Discussion

In this article, we have proposed a novel and flexible test of genetic association incorporating phenotype heterogeneity. Single- and multi-SNP versions of the method have been described. We have shown that tests which incorporate phenotype heterogeneity can be more powerful than tests which do not and that a test which incorporates heterogeneity can perform consistently well, across a range of scenarios for phenotype heterogeneity. In the absence of heterogeneity, our multi-SNP/global test is closely related and performs similarly to the AML approach [11], which has been shown to perform well in terms of statistical power compared to alternative global tests of association (for a homogenous phenotype) proposed in the genetic epidemiological literature [10, 21].

Our novel approach for testing association, incorporating heterogeneity, is flexible in several ways. By modelling the distribution of genotypes, conditional on phenotypic characteristics, our approach allows heterogeneity to be described in terms that are more complex than can be described by a single categorical variable, e.g. by using

Table 4. Single-SNP analysis of marker rs4646 in gene CYP19A1 located on chromosome 15q21.1 – association tests incorporating heterogeneity in terms of tumour grade and size

Covariate(s)	Grade	Tumour size	Tumour grade + size
Controls	1,527	1,527	1,527
Cases	1,013	1,470	1,000
m_0 (SE)	−0.496 (0.120)	−0.267 (0.084)	−0.546 (0.132)
m_{grade} (SE)	0.228 (0.076)	–	0.235 (0.079)
m_{size} (SE)	–	0.067 (0.034)	0.022 (0.042)
p value	9.1×10^{-5}	6×10^{-3}	1.7×10^{-4}

a mixture of categorical and continuous variables. We note that our approach can adjust for population stratification by including axes of genetic variation, obtained from for example a principal component analysis [22], as covariates in the model for the conditional distribution of genotypes. At present, our multi-SNP method does not assume/incorporate prior knowledge about the individual SNPs. Given the importance of replicating reported genetic associations in independent samples, an interesting extension of our approach would be to allow for prior specification of $P(V)$, either in the current likelihood or in an empirical Bayesian framework.

We have applied our method in a re-analysis of Swedish case-control study data, studying the association of breast cancer risk/prognosticators with genetic variability in the oestrogen metabolic pathway. We were able to confirm previous findings and strengthen previous evidence that the polymorphism rs4646, in the 3'-untranslated region of CYP19A1, is (jointly) associated with breast cancer risk and tumour characteristics in postmenopausal women. The minor variant of rs4646 has previously been shown to be associated with a shorter progression-free survival [19] and with the HER2 status of the tumour [23]. Testing for this specific gene polymorphism may help to improve the clinical management of breast cancer patients.

In summary, we have developed a new powerful association method for detecting single variants or sets of variants with related biological function contributing to the disease association, taking into account the phenotype heterogeneity that accompanies complex disease aetiologies. By taking subtypes of a disease into account when relating disease risk to genetic factors, specific aetiological pathways may become identifiable.

Acknowledgements

Financial support from the Swedish Research Council (521-2008-2335 and 523-2006-972), Cancerfonden (contract No. 09 0420), the Märit and Hans Rausing's Initiative against Breast Cancer, and the Susan G. Komen Breast Foundation is gratefully acknowledged. We would like to thank colleagues at Karolinska

Institutet and the Genome Institute of Singapore, who work on the genetic association study of breast cancer (described in 'A Genetic Association Analysis of Breast Cancer Based on Genes Involved in Oestrogen Metabolism'), for data and comments. We are grateful to Jingmei Li for helping us with the pathway data from the genome-wide study. We would also like to thank referees for valuable comments.

References

- 1 Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP: A powerful and flexible multilocus association test of quantitative traits. *Am J Hum Genet* 2008;82:386–397.
- 2 Sagoo GS, Little J, Higgins JP: Systematic reviews of genetic association studies. *Human Genome Epidemiology Network. PLoS Med* 2009;6:e28.
- 3 Zhou H, Pan W: Binomial mixture model-based association tests under genetic heterogeneity. *Ann Hum Genet* 2009;73:614–630.
- 4 Bertucci F, Birnbaum D: Reasons for breast cancer heterogeneity. *J Biol* 2008;7:6.
- 5 Garcia-Closas M, et al: Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS Genet* 2008;4:e1000054.
- 6 Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, Hattersley AT, McCarthy MI: A powerful approach to sub-phenotype analysis in population-based genetic association studies. *Genet Epidemiol* 2010;34:335–343.
- 7 Hoh J, Ott J: Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet* 2003;4:701–709.
- 8 Thomas DC: The need for a systematic approach to complex pathways in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* 2005;14:557–559.
- 9 Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, de Andrade M, Henley JR, Rocca WA, Ahlskog JE, Maraganore DM: A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. *PLoS Genet* 2007;3:e98.
- 10 Dudbridge F, Koeleman BP: Rank truncated product of P-values, with application to genomewide association scans. *Genet Epidemiol* 2003;25:360–366.
- 11 Tyrer J, Pharoah PD, Easton DF: The admixture maximum likelihood test: a novel experiment-wise test of association between disease and multiple SNPs. *Genet Epidemiol* 2006;30:636–643.
- 12 Low YL, Li Y, Humphreys K, Thalamuthu A, Li Y, Darabi H, Wedren S, Bonnard C, Czene K, Iles MM, Heikkinen T, Aittomäki K, Blomqvist C, Nevanlinna H, Hall P, Liu ET, Liu J: Multi-variant pathway association analysis reveals the importance of genetic determinants of estrogen metabolism in breast and endometrial cancer susceptibility. *PLoS Genet* 2010;6:e1001012.
- 13 Li J, Humphreys K, Heikkinen T, Aittomäki K, Blomqvist C, Pharoah PDP, Dunning AM, Ahmed S, Hoening MJ, Martens JWM, van den Ouweland AMW, Alfredsson L, Palotie A, Peltonen-Palotie L, Irwanto A, Low HQ, Teoh GHK, Thalamuthu A, Easton DF, Nevanlinna H, Liu J, Czene K, Hall P: A combined analysis of genome-wide association studies in breast cancer. *Breast Cancer Res Treat* 2010, E-pub ahead of print.
- 14 Sasieni PD: From genotypes to genes: doubling the sample size. *Biometrics* 1997;53:1253–1261.
- 15 Clayton D: A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* 1999;65:1170–1177.
- 16 Brent RP: Algorithms for Minimization Without Derivatives. Englewood Cliffs, N.J., Prentice-Hall, Inc., 1973.
- 17 Jiang R, Dong J, Dai Y: Improving power in genetic association studies via wavelet transformation. *BMC Genet* 2009;10:53.
- 18 Kanehisa M, Goto S: KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27–30.
- 19 Garcia-Casado Z, Guerrero-Zotano A, Llombart-Cussac A, Calatrava A, Fernandez-Serra A, Ruiz-Simon A, Gavila J, Clement MA, Almenar S, Cervera-Deval J, Campos J, Albaladejo CV, Llombart-Bosch A, Guillem V, Lopez-Guerrero JA: A polymorphism at the 3'-UTR region of the aromatase gene defines a subgroup of postmenopausal breast cancer patients with poor response to neoadjuvant letrozole. *BMC Cancer* 2010;10:36.
- 20 Darabi H, Czene K, Wedren S, Li Y, Liu J, Hall P, Humphreys K: Genetic variation in the androgen estrogen conversion pathway in relation to breast cancer prognosticators. *Breast Cancer Res Treat* 2010, E-pub ahead of print.
- 21 O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, Corvin A: The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics* 2009;25:2762–2763.
- 22 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.
- 23 Fasching PA, Loehberg CR, Strissel PL, Lux MP, Bani MR, Schrauder M, Geiler S, Ringeleff K, Oeser S, Weihbrecht S, Schulz-Wendtland R, Hartmann A, Beckmann MW, Strick R: Single nucleotide polymorphisms of the aromatase gene (CYP19A1), HER2/neu status, and prognosis in breast cancer patients. *Breast Cancer Res Treat* 2008;112:89–98.