When is an endophenotype useful to detect association to a disease? Exploring the relationships between disease status, endophenotype and genetic polymorphisms
Running title: Disease, endophenotype and genetic polymorphisms

Alexandre Bureau[1,2,*], Jordie Croteau[2]

Received ____________________; accepted ____________________

[1] Département de médecine sociale et préventive, Université Laval, Québec, Canada

[2] Institut universitaire en santé mentale de Québec du Centre intégré de santé et de services sociaux de la Capitale-Nationale, Québec, Canada (This is the institute where the work was conducted.)

[*] Address for correspondence and reprints: Dr. Alexandre Bureau, 1050 rue de la Médecine, local 2457, Quebec City, QC, CANADA, G1V 0A6. tel.: 418-656-2131, ext.:3342. fax: 418-656-7759. E-mail: alexandre.bureau@msp.ulaval.ca.

Abstract

Objectives: Investigate the conditions and the analysis strategies required so endophenotypes related to a disease help discover genetic variants involved in the disease.
Methods: Association with disease susceptibility variants is examined as a function of the relationships between disease status, endophenotype values and the genotype at another disease or endophenotype susceptibility locus assumed to be previously known, using approximate linear models of allele frequencies as a function of these variables and simulations in the context of family studies when the endophenotype is dichotomous.
Results: Under genetic mechanisms where the risk allele of the tested locus has an effect exclusively in subjects with the endophenotype, the risk allele frequency differences between affected and unaffected subjects are much greater in the subset of subjects with an endophenotype impairment than in those without such impairment, and power gains are obtained when testing association under a joint disease-endophenotype model, both with two-locus or single-locus tests. However, with moderate main effect on risk of disease or endophenotype impairment, testing directly the association between risk allele and disease or endophenotype is more powerful than testing under a joint disease-endophenotype model.
Conclusions: Joint modeling of disease and endophenotype should be used only in parallel with standard disease association testing.
Keywords: expected allele frequency; family-based association; linear model; generalized disequilibrium test; power; quasi-likelihood score test; simulation study

1. Background

In the genetic study of complex diseases, additional phenotypic measurements are increasingly taken to uncover traits related to the disease phenotype, called endophenotypes [1, 2]. The term endophenotype is commonly used to describe traits that are minimally: 1) associated with a disease, 2) heritable, 3) present whether or not the illness is active (state-independent), 4) co-segregating with the disease in families and 5) present in non-affected relatives at a higher rate than in the population [1].

Most often, endophenotypes are used as an univariate or multivariate phenotype in genetic association studies, in place of the disease diagnosis. It has been recognized that crude estimates of association parameters between genetic variants and endophenotypes in samples ascertained conditionally on disease status can be biased when both the genetic variant and the endophenotype are associated with disease status, and corrections for that bias have been proposed for samples of unrelated disease cases and controls (see [3] and [4] for recent proposals and reviews of prior work). One model developed for linkage analysis of a quantitative trait and a dichotomous disease status [5], and applied to test jointly linkage to major depression and quantitative endophenotypes [6], assumes a restrictive latent dependence structure between the two traits and does not take into account ascertainment on the disease status. Others have proposed to test jointly the association to the disease status and an endophenotype in case-control samples [7].

Keeping in mind the ultimate goal to detect genetic variants involved in the etiology of the disease, we adopt the view that the optimal way to use an endophenotype in an association study should be the one that best reaches that goal. It will depend on the relationship between the endophenotype, the disease and genetic variants (those to be discovered as well as those already known to be associated to the disease and/or endophenotype). This is to our knowledge the first study to investigate these relationships with respect to the associations with genetic variants that they generate.

The co-segregation of a disease and an impairment on an endophenotype in densely affected families represents evidence for genetic variants influencing both traits. These genetic variants may differ from the genetic variants underlying the same impairment in a general population, for which recent studies suggest a genetic architecture similar to that of complex traits [8]. This is why the analysis of an endophenotypes to uncover disease-related variants is most relevant in families where the disease of interest is present. This work therefore focuses on the context of familial association studies.

Controlling biases in estimates of parameters related to an endophenotype due to ascertainment of families based on the disease status of their members is much more complex than in case-control samples. The only practical approach that fully protects against ascertainment biases is to condition on the phenotype. In the context of familial association studies, this is currently manageable only with categorical phenotypes. As far as we know, current methods for association to quantitative traits in families assume random sampling of pedigrees, and merely assess the robustness of the methods to ascertainement based on a disease related to the quantitative trait (see [9] for a recent example). In this work we therefore examine a dichotomous endophenotype. For quantitative endophenotypes, a cut-off point below which a subject's score is interpreted as being impaired can be applied to dichotomize it. The bivariate phenotype formed by a dichotomous disease status and a dichotomous endophenotype has thus four categories.

For disease phenotypes, analysis methods conditioning on the phenotype in families essentially compare allele or genotype frequencies between affected and unaffected subjects (or pseudo-controls), either within familly [10, 11] or at the population level [12, 13]. The magnitude of these expected differences determines the power to detect association to the genetic variants. Thus, we examine retrospectively allele frequencies in subjects in the four phenotypic categories formed by the disease and endophenotype to determine the analysis strategy to adopt.

Nevertheless, our starting points are genetic risk mechanisms describing prospectively the probability, or penetrance, of the phenotype categories given the genotype. We begin by reviewing the model forms that have been proposed to represent association with predictor variables applicable to this context. Next, we translate prospective risks into allele frequencies. Allele frequencies and their ratios do not however reduce to simple expressions of the parameters of the prospective risk mechanisms. Our approach has been to express allele frequency as an approximate linear model of the disease and endophenotype status and the genotype at a known disease or endophenotype susceptibility locus. This allowed us to 1) examine relevant special cases of relationships between allele frequencies and these variables and derive the constraints on the risk mechanisms that are required to obtain them and 2) examine the expected allele frequencies under a set of genetic mechanisms all involving a form of modifying effect of the tested locus on a known susceptibility locus, and various additional effects. The approximate linear models of allele frequency were estimated from expected allele frequencies in unrelated individuals, and from simulation in pedigrees ascertained for a minimal number of affected subjects.

Finally, we illustrate our recommendations about the most powerful approach by presenting the power of statistical tests proposed elsewhere by ourselves [14] and others [10, 12] under representative scenarios of genetic mechanisms that we have studied.

2. Methods

2.1. Notation

Let Y_1 and Y_2 be binary variables coding the presence (1) or absence (0) of the endophenotype impairment and the disease, respectively. Let X_l denote the proportion of a given allele in a genotype at the l^{th} biallelic locus, taking values 0, $\frac{1}{2}$ and 1, such that the allele frequency equals the expectation of the genotypic variable, e.g. $E[X_1]$ at locus 1. When two loci are considered, we assume locus 1 is known to be associated with the disease or endophenotype, and locus 2 is being tested. When a single locus is considered, we drop the index. We assume throughout that the allele coded by X_l is the allele increasing risk.

2.2. Disease-endophenotype models

2.2.1. Review of models for bivariate dichotomous outcomes

Various types of models for the distribution of a pair of binary outcome variables have been proposed in other contexts of regression on predictor variables, which we denote collectively by the vector X. We review them here in the context of the joint distribution of a dichotomous disease status and a dichotomous endophenotype as a function of the observed genotypes of an individual.

polytomous In this type of models, the probability of each combination of Y_1 and Y_2 other than a reference category (taken to be $Y_1 = 0, Y_2 = 0$) are modeled separately by functions of the predictor variables with each their own coefficients, that is $P[Y_1 = 1, Y_2 = 0|X] = f_1(\beta X), P[Y_1 = 0, Y_1 = 0|X] = f_2(\beta X), P[Y_1 = 1, Y_2 = 1|X] = f_3(\beta X)$ and $P[Y_1 = 0, Y_2 = 0|X] = 1 - P[Y_1 = 1, Y_2 = 0|X] - P[Y_1 = 0, Y_2 = 1|X] - P[Y_1 = 1, Y_2 = 1|X]$. The two variables Y_1 and Y_2 are treated on the same footing [15, Section11.4].

transition They suppose that Y_1 preceeds Y_2, which applies to the context where an endophenotype is an intermediate trait and the disease is the final outcome. The probability of response for Y_1 is modeled as a function of the predictor variables $P[Y_1 = 1|X] = f(\beta X)$, and the second variable Y_2 as a function of Y_1 and the predictor variables $P[Y_2 = 1|Y_1, X] = f(\gamma(X, Y_1, XY_1))$, where (X, Y_1, XY_1) denotes the vector of all terms involving X, Y_1 and XY_1, and $\gamma(.)$ denotes a linear combination [16].

marginal This approach adds to a marginal model of the probability of response for Y_1 and Y_2, $P[Y_1 = 1|X] = f(\beta_1 X)$ and $P[Y_2 = 1|X] = f(\beta_2 X)$, a model for a parameter of association between Y_1 and Y_2, for instance the odds ratio (OR) [17]. This model will not be considered further in this work, since the interest is not in the effect of polymorphisms on the association between the disease status and the endophenotype, but rather in their effect on the risk of the various combinations of these two components of the phenotype.

In the next section, we study properties of general polytomous and transition models and describe statistical tests applicable to analyze a dichotomous disease status and endophenotype. In section 2.2.3, we examine specific scenarios of genetic mechanisms as instances of such models.

2.2.2. Classes of models and statistical tests for dichotomous disease status and endophenotype

Polytomous and transition models that we consider generalize the usual two-locus logistic model for a dichotomous trait:

$$\log\left(\frac{P[Y = 1|X_1, X_2]}{P[Y = 0|X_1, X_2]}\right) = \eta_0 + \eta_1 X_1 + \eta_2 X_2 + \eta_3 X_1 X_2 \qquad (1)$$

In the polytomous model, logistic functions are specified for each phenotype category against the reference category. For the combinations of a dichotomous disease status and dichotomous endophenotype, we have:

$$\begin{aligned}
\log\left(\frac{P[Y_1=1,Y_2=0|X_1,X_2]}{P[Y_1=0,Y_2=0|X_1,X_2]}\right) &= \beta_{10}+\beta_{11}X_1+\beta_{12}X_2+\beta_{13}X_1X_2 \\
\log\left(\frac{P[Y_1=0,Y_2=1|X_1,X_2]}{P[Y_1=0,Y_2=0|X_1,X_2]}\right) &= \beta_{20}+\beta_{21}X_1+\beta_{22}X_2+\beta_{23}X_1X_2 \qquad (2) \\
\log\left(\frac{P[Y_1=1,Y_2=1|X_1,X_2]}{P[Y_1=0,Y_2=0|X_1,X_2]}\right) &= \beta_{30}+\beta_{31}X_1+\beta_{32}X_2+\beta_{33}X_1X_2
\end{aligned}$$

Statistical tests for testing genetic association with categorical traits in family samples fall under two broad classes of approaches: population level analysis and within-family conditional analysis. For the population level analysis, we adopt the Maximum Quasi-Likelihood Score (M_{QLS}) test of [12] which has been developed to optimize power to test for association to dichotomous phenotypes in related individuals but has not been derived for polytomous phenotypes. In our evaluation of power to detect association under various genetic mechanisms (see section 3), we applied it to the disease status (MQLSd), to the endophenotype status (MQLSe) and to endophenotype impairment with disease against all other combinations of disease and endophenotype status (MQLSde). For the within-family conditional analysis, we adopted the score test for dichotomous phenotypes in general pedigrees derived by [10] under the name "Generalized disequilibrium test" (GDT). We have extended the GDT to an outcome with $K > 2$ levels and to two unlinked loci under the polytomous logistic regression model 2 [14]. We applied this new test to interaction coefficients and subsets of coefficients in model 2 and a single locus polytomous model (model 2 with X_2 only). The coefficients in that model are labelled $\beta(1L)$. We note here that the score test of a single coefficient of a logistic function attached to an outcome category under a polytomous model is a contrast between the sum of the corresponding X terms over subjects in the outcome category and over subjects in all other categories. Hence, the test of the coefficient $\beta_3(1L)$ for the endophenotype impairment with disease category is comparable to MQLSde. We denote ”cpoly” the conditional test of locus 2 under model 2, i.e. the 6 d.f. test of the null hypothesis $\beta_{12} = \beta_{13} = \beta_{22} = \beta_{23} = \beta_{32} = \beta_{33} = 0$, and ”cdisease” and ”cendo” the conditional test of locus 2 under model 1 for the disease status and the endophenotype, respectively, i.e. the 2 d.f. test of the null hypothesis $\eta_2 = \eta_3 = 0$. The original GDT was applied to the disease status (GDTd) and to the endophenotype status (GDTe).

The general two-locus transition model for a dichotomous disease status and endophenotype can be written as:

$$\begin{aligned} P[Y_1 = 1|X_1, X_2] &= f(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2) && (3) \\ P[Y_2 = 1|Y_1, X_1, X_2] &= f(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_1 X_2 && (4) \\ &\quad + \gamma_4 Y_1 + \gamma_5 X_1 Y_1 + \gamma_6 X_2 Y_1 + \gamma_7 X_1 X_2 Y_1) \end{aligned}$$

The transition model permits to specify sub-models where either trait is conditionally independent of the allele count at a locus, given the other variables. We consider here two such conditions: conditional independence between the endophenotype and X_2 given X_1 and, in addition, conditional independence between the disease status and X_1 given the endophenotype Y_1 and X_2. The first condition gives the model:

$$\begin{aligned} P[Y_1 = 1|X_1, X_2] &= P[Y_1 = 1|X_1] = f(\beta_0 + \beta_1 X_1) && (5) \\ P[Y_2 = 1|Y_1, X_1, X_2] &= f(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_1 X_2 && (6) \\ &\quad + \gamma_4 Y_1 + \gamma_5 X_1 Y_1 + \gamma_6 X_2 Y_1 + \gamma_7 X_1 X_2 Y_1) \end{aligned}$$

while the two conditions together give:

$$P[Y_1 = 1|X_1, X_2] = P[Y_1 = 1|X_1] = f(\beta_0 + \beta_1 X_1) \qquad (7)$$

$$P[Y_2 = 1|Y_1, X_1, X_2] = P[Y_2 = 1|Y_1, X_2] \;=\; f(\gamma_0 + \gamma_2 X_2 + \gamma_4 Y_1 + \gamma_6 X_2 Y_1) \tag{8}$$

2.2.3. Example scenarios

We consider two-locus mechanisms of the disease and endophenotype where the two loci interact to cause the disease, such that disease risk depends directly on the genotype at both loci (as opposed to mechanisms where the effect of one locus is entirely mediated by the endophenotype as in equation (9)). In scenarios where one locus has an effect on the disease status only (not the endophenotype), we assume that it is locus 2 and focus on methods to detect that locus (see Figure 1). We present examples of scenarios represented by polytomous and transition models with the common characteristics that the A allele of locus 1 is marginally associated to the endophenotype and the presence of risk alleles at locus 2 shifts the effect of locus 1 risk alleles from an effect on the risk of endophenotype impairment without disease ($Y_1 = 1, Y_2 = 0$) to an effect on the risk of endophenotype impairment with disease ($Y_1 = 1, Y_2 = 1$). For the polytomous models, this behaviour is obtained through a negative β_{13} and a positive β_{33}. For transition models, it is obtained through the interaction terms involving locus 1 and 2. The risk allele frequency was set to 0.1 at locus 1 and varied between 0.1 and 0.5 at locus 2. The specificities of each polytomous model representations of genetic mechanisms are presented in Section 4.2 and the values of the coefficients corresponding to each scenario are given in Table 1, while the specificities of each transition model representations of genetic mechanisms are presented in Section 4.3 and the coefficients are given in Table 2.

2.3. Study of allele frequencies implied by the two-locus mechanisms

The variations in allele or genotype frequencies at locus 2 as a function of the disease status and endophenotype value and the genotype at locus 1 determine the power of the analysis strategies applied to detect locus 2. These allele or genotype frequencies, as well as their ratios do not however reduce to simple expressions of the polytomous or transition models above. In particular, ratios still depend on the population genotype frequency at locus 2. To interpret allele frequency at locus 2, we express them using an approximate linear model of the disease and endophenotype status and the genotype at locus 1. Special cases can however be derived under certain conditions, and we present them in section 2.3.2.

2.3.1. Linear models of allele frequency

We have observed that, for a given set of genotype frequencies and a given genetic mechanism, the allele frequency at locus 2 as a function of the disease and endophenotype status and the genotype at locus 1 of an individual is well approximated by a simple linear model. Our approach has therefore been to examine example scenarios of mechanisms and to select and fit parsimonious retrospective linear models of the allele frequency at locus 2 given the disease and endophenotype status and the genotype at locus 1 to them. We also fitted models ignoring locus 1. We used the weighted sum of squared errors (SSE) between the actual and predicted allele frequencies of each combination of disease and endophenotype status and genotype at locus 1 as fit criterion, where each combination is weighted by its prevalence in the population. Details on model selection are given in section 1 of the Appendix.

In families ascertained based on some criterion such as a minimal number of affected subjects, the allele frequency in a member of the family depends on the phenotype of all his relatives. Computing the allele frequency in one subject taking into account the ascertainment criterion for the entire family is much more complicated than in unrelated subjects. Computation of empirical allele frequencies observed in samples of simulated families meeting the ascertainment criterion thus becomes a more practical solution in this case. The process to simulate disease phenotype, endophenotype and genotype data in families is explained in section 2 of the Appendix. The parameters of our simulation study are described in section 3.

2.3.2. Special cases of alleles frequency models and required conditions on risk mechanisms

An important special case is when allele frequencies at locus 2 can be expressed as a function of Y_1 and Y_2 alone. Under conditional independence between the disease status Y_2 and allele proportion at locus 1 X_1 specified in equation (9), X_1 provides no information on X_2 when the endophenotype Y_1 is observed in addition to Y_2. In that case,

$$E[X_2|X_1, Y_1, Y_2] = E[X_2|Y_1, Y_2] = \delta_0 + \delta_1 Y_1 + \delta_2 Y_2 + \delta_3 Y_1 Y_2 \tag{9}$$

The consequence of this result is that association between the phenotype and the genotype at locus 2 can be tested under one-locus models with optimal power. Using the genotype at locus 1 in a two-locus model provides no further information.

Another important special case is when genotype frequencies at locus 2 depend on Y_2 only when the endophenotype impairment is present ($Y_1 = 1$) or, said otherwise, the genotype frequencies at locus 2 do not depend on Y_2 when the endophenotype impairment is absent ($Y_1 = 0$), i.e.

$$P[X_2|Y_1 = 0, Y_2] = P[X_2|Y_1 = 0], Y_2 = 0, 1, \forall X_2 \tag{10}$$

The same property for the allele frequency at locus 2 derives from the definition of expectation:

$$E[X_2|Y_1 = 0, Y_2] = E[X_2|Y_1 = 0], Y_2 = 0, 1$$

A corrolary is that the allele frequency at locus 2 as a function of Y_1 and Y_2 can be described exactly by the linear model $E[X_2|Y_1, Y_2] = \beta_0 + \beta_1 Y_1 + \beta_2 Y_1 Y_2$ i.e. there is no main effect of Y_2.

This special case is obtained when the prospective risk mechanisms satisfy certain conditions, which are specific to each type of model. For a polytomous logistic model, the required condition is that the logistic function contrasting $Y_2 = 1$ and $Y_2 = 0$ when $Y_1 = 0$ is a constant, namely for an arbitrary vector of covariates Z:

$$\log\left(\frac{P[Y_1 = 0, Y_2 = 1|X_2, Z]}{P[Y_1 = 0, Y_2 = 0|X_2, Z]}\right) = \beta_{20} \tag{11}$$

For a transition model, the condition is that X_2 appears in the model only in product terms with Y_1, so that the terms are equal to 0 when $Y_1 = 0$. Mathematically, this can be expressed as:

$$P[Y_2 = 1|Y_1, X_2, Z] = f(\gamma_0 + g_1(Z) + Y_1 g_2(X_2, Z)), \tag{12}$$

where $g_1(Z)$ and $g_2(X_2, Z)$ are arbitrary functions of X_2 and Z.

Conditioning on a fixed level of a covariate vector Z gives a less stringent version of 10. In that case,

$$P[X_2|Y_1 = 0, Y_2, Z = z] = P[X_2|Y_1 = 0, Z = z], Y_2 = 0, 1, \forall X_2 \tag{13}$$

The same property for the allele frequency at locus 2 derives from the definition of expectation:

$$E[X_2|Y_1 = 0, Y_2, Z = z] = E[X_2|Y_1 = 0, Z = z], Y_2 = 0, 1$$

The interpretation of (12) is that the probabilities $P[X_2|Y_1 = 0, Y_2 = 0, Z]$ and $P[X_2|Y_1 = 0, Y_2 = 1, Z]$ are superimposed curves or surfaces as a function of an univariate or multivariate Z. If the probability does not depend on Z, then the curve or surface is flat. Note that Z can be X_1, the allele count at locus 1.

For a polytomous logistic model, the required condition becomes that the logistic function contrasting $Y_2 = 1$ and $Y_2 = 0$ when $Y_1 = 0$ is a function of Z alone, namely for an arbitrary vector of covariates Z:

$$\log\left(\frac{P[Y_1 = 0, Y_2 = 1|X_2, Z]}{P[Y_1 = 0, Y_2 = 0|X_2, Z]}\right) = g(Z) \tag{14}$$

Proofs that models satisfying the above conditions imply the special cases 10 and 13 are given in section 3 of the Appendix.

3. Simulation study

The first goal of the simulation study was to explore the distribution of disease status and endophenotype resulting from the models described in previous sections in families ascertained based on the disease status of their members. The second goal was to illustrate our recommendations about the most powerful approach by presenting the power of statistical tests to detect locus 2 and relate it to the characteristics of the genetic mechanisms. We simulated the endophenotype Y_1 under the two locus mechanism specified for that trait in transition models or after marginalizing over Y_2 under polytomous models, as in [18]. The computer package Simla [19] was used to simulate Y_1 and marker genotypes of the family members as described in section 2.1 of the Appendix.

In families, the complex phenotypes of relatives depend on one another through multiple genes and possibly common environmental exposures. It is therefore unrealistic to assume that the genotypes at the two loci included in a mechanism suffice to generate a realistic dependance structure between relatives. A more plausible simulation scheme is described in section 2 of the Appendix. We applied it to simulate the disease status Y_2, with the values of γ from the selection of scenarios described in section 2.2.3, a value of $\sigma = 1$ giving a weak dependance between the disease status of relatives in addition to the considered genotypes, a value of $\alpha = \log(4)$ and a value of $\nu = 0.01$.

3.1. Family structure and ascertainment scheme

The family structure used in the simulations is a 3-generation 16-members family depicted in Figure 2, where a disease and endophenotype configuration obtained in our simulation study is depicted as illustration. Our ascertainment criterion required a cousin pair affected by the disease (see section 2.3 of the Appendix for more details). The disease and endophenotype status of all family members was assumed to be observed. For each scenario 10,000 families were generated. For the power comparisons, these families were divided into 100 samples of 100 families.

3.2. Fitting linear models to allele frequencies

The risk allele frequency at locus 2 was estimated for each of the 12 combinations of the disease status and endophenotype levels and the genotype at locus 1 by counting the alleles observed in the subjects falling in these 12 categories among the 160,000 subjects from a sample of 10,000 16-member families. Linear models for the risk allele frequency at locus 2 were then fitted by minimizing the SSE for the 12 combinations weighted by the estimated frequencies as we have done on the theoretical frequencies in unrelated individuals (Section 2.3.1). We elected to use a simple allele count instead of a more sophisticated estimate taking into account dependance between the genotypes of subjects in the same family because allele counts remain unbiased in presence of dependance and we were not concerned with estimating the variability of the estimates given the large size of the sample that we simulated.

4. Results

4.1. Distribution of disease status and endophenotype under the example scenarios

Epidemiological parameters resulting from each scenario when the population risk allele frequency (pRAF) at locus 2 is 0.3 are given in Table 3. The figures were similar with pRAFs of 0.1 and 0.5 (not shown). The intercept parameters in the scenarios were selected to achieve a disease prevalence around 1% and an endophenotype impairment prevalence in unaffected subjects between 10 - 15% in unrelated individuals. Under our ascertainment criterion, disease prevalence reaches about 15% in the family sample.

The endophenotype impairment prevalence in the unaffected subjects does not have the same interpretation in ascertained family samples where unaffected subjects are relatives of affected individuals than in unrelated subjects with no information on the phenotype of relatives, which can be seen as population controls. The prevalence of the impairment in affected and unaffected subjects from the ascertained families (60 - 70% in affected subjects, 15 - 20% in unaffected subjects) was higher than the prevalence in unrelated affected and unaffected subjects (48 - 58% in affected subjects, 10 - 13% in unaffected subjects). This is an effect of the direct dependence of the disease diagnosis on the endophenotype introduced into the simulation scheme. Families containing multiple subjects with an endophenotype impairment then tend to contain more affected subjects, and therefore have a greater probability of being ascertained. This implies that the endophenotype impairment is more prevalent not only in the affected subjects, but also in the unaffected subjects. The higher prevalence of alleles increasing the risk of both disease and endophenotype impairment could also have contributed to the increased prevalence in ascertained families. We found this contribution to be negligible, as the prevalences of the endophenotype impairment in families simulated without direct dependence of the disease diagnosis on the endophenotype status were similar to those from the unrelated subjects (results not shown).

With the simultaneous increase in prevalence of the endophenotype impairment in affected and unaffected subjects, the OR measuring the association between disease diagnosis and endophenotype impairment remained roughly at the same level for the same polytomous model of genetic scenario between the ascertained family samples and the unrelated subjects (7 - 8 with the exception of scenario P4 with an OR around 12). For the genetic scenarios represented by transition models, the ORs between 7 - 9 in the unrelated individuals increased by about one unit in the ascertained families due to a proportionally greater endophenotype impairment prevalence increase in the affected subjects compared to the unaffected ones.

4.2. Selection of polytomous models of genetic scenarios

In what follows, X_1 denotes the A allele proportion at locus 1 and X_2 the B allele proportion at locus 2. The values of the polytomous model coefficients corresponding to each scenario are given in Table 1. Scenario P1 is the base scenario with only interaction terms and the other scenarios were constructed by introducing various main effect terms. Scenarios P1 and P4 satisfy condition 11 in unrelated subjects. Scenario P3 satisfies only the less stringent condition 14.

4.2.1. Scenario P1

Under this scenario, locus 2 influences the risk of disease vs. endophenotype only when a locus 1 endophenotype-related allele is present. The two loci have no effect on the risk of disease in absence of the endophenotype. The phenotype-specific frequency of the risk allele B of locus 2 as a function of locus 1 genotype in unrelated individuals is displayed in Figure 3A when the B allele population frequency is 0.3. On the plot, one can see that phenotypic categories differ in pRAF only when at least one A allele is present at locus 1. In presence of the A allele, the largest pRAF difference is observed between affected and unaffected subjects with the endophenotype. The pRAFs of the two disease statuses do not differ in subjects without the endophenotype.

The B allele frequency $E[X_2|X_1, Y_1, Y_2]$ is well described by the linear models on the left side of table 4 ($SSE \leq 1\%$ in unrelated individuals). As expected from property 14, $E[X_2|X_1, Y_1 = 0, Y_2 = 0]$ and $E[X_2|X_1, Y_1 = 0, Y_2 = 1]$ are identical linear functions of X_1, which equal the population B allele frequency in carriers of the reference aa genotype ($X_1 = 0$). The pRAF difference between affected and unaffected subjects with the endophenotype ($X_1Y_1Y_2$ term) increases with the pRAF, but at a slowing rate. In ascertained families, the pRAF is elevated, particulary in carriers of the AA genotype at locus 1, and the contrast between affected and unaffected subjects is somewhat attenuated compared to unrelated subjects, but the same linear model provides a good fit to the data ($SSE = 3\%$ for a pRAF of 0.3), indicating that property 14 is approximately satisfied.

If we consider only the genotype at locus 2, the one-locus expressions for the B allele frequency are on the right side of table 4. These expressions are exact in unrelated subjects, since scenario P1 satisfies condition 11, and provide an excellent fit in ascertained families ($SSE = 3\%$ for a pRAF of 0.3). Among subjects with the endophenotype ($Y_1 = 0$), the contrast between affected and unaffected subjects is attenuated compared to its maximal value in carriers of the AA genotype at locus 1 (for a pRAF = 0.3, we have 0.061 compared to 0.396 - (-0.070) = 0.466 in unrelated individuals).

As expected from the above observations, a score test of locus 2 conditional on genotype at locus 1 with the joint Y_1, Y_2 phenotype, β_{33}, resulted in the highest power, while tests considering only either the disease status or the endophenotype and locus 2 had low power (Figure 4A).

In summary, scenario P1 exemplifies a mechanism with only interaction effects, where pRAF differences at locus 2 are maximized in carriers of the risk allele at locus 1.

4.2.2. Scenario P2

Scenario P2 differs from scenario P1 in that the locus 2 genotype has an effect on the risk of the disease independently of the locus 1 genotype. This is reflected in the pRAF at locus 2 on Figure 3B and the linear models for $E[X_2|X_1, Y_1, Y_2]$ on the left side of table S1 ($SSE \leq 1\%$ in unrelated individuals). If we consider only the genotype at locus 2, the one-locus expressions for the pRAF are on the right side of table S1 ($SSE \leq 2.1\%$ for unrelated subjects). Both types of models indicate a much higher pRAF in affected than in unaffected subjects, whether an endophenotype impairment is present or not. In ascertained families, the pRAF is greatly elevated and the difference in pRAF is reduced (right side of table S1).

The strong marginal association between the allele proportion X_2 and the disease phenotype Y_2 implies that the power of a standard association test is expected to be good. The much higher contrast in pRAF between affected and unaffected subjects in carriers of two risk alleles at locus 1 (0.241 + 0.161 - (-0.063) = 0.465 with pRAF = 0.3 in unrelated individuals) could provide a boost in power under two-locus models considering locus 1, although all pRAF differences are attenuated in ascertained families when the pRAF = 0.3. We observe on Figure 4B that tests of association to the dichotomous disease status have 100% power, whether they are single-locus (MQLSd, GDTd) or conditionnal on locus 1 (cdisease). Tests involving Y_1 and Y_2 (MQLSde, $\beta_3(1L)$, cpoly) have only slightly lower power, but the small difference may be misleading given the ceiling effect on power under this scenario.

4.2.3. Scenario P3

Scenario P3 differs from scenario P1 in that the locus 1 genotype has an effect on the risk of the disease independently of the locus 2 genotype, in addition to its effect on the risk of the endophenotype impairment. This locus 1 effect on the risk of disease has little impact on the pRAF at locus 2 given the disease and endophenotype status and the genotype at locus 1, as can be seen by comparing Figure 3C and 3A. Indeed, the best fitting linear models of the B allele frequency are similar to the corresponding expression for Scenario P1, except for the magnitude of the interaction term (left side of table S2, $SSE \leq 0.2\%$ for unrelated subjects).

As expected from property 14, $E[X_2|X_1, Y_1 = 0, Y_2 = 0]$ and $E[X_2|X_1, Y_1 = 0, Y_2 = 1]$ are identical linear functions of X_1. If we consider only the genotype at locus 2, the one-locus expression for the pRAF is also similar to that for Scenario P1, except here it is not an exact expression since scenario P3 does not satisfy condition 11(Right side of table S2. $SSE \leq 0.02\%$ for unrelated subjects).

As with Scenario P1, pRAF differences at locus 2 are maximized in carriers of the risk allele at locus 1 and a score test of locus 2 conditional on genotype at locus 1 with the joint Y_1, Y_2 phenotype, β_{33}, resulted in the highest power.

4.2.4. Scenario P4

Scenario P4 differs from scenario P1 in that the locus 2 genotype has an effect on the risk of endophenotype impairment. The scenario is such that risk alleles at either locus 1 or locus 2 increase the risk of endophenotype impairment, but having both results in no further risk increase. Figure 3D illustrates the implications for the pRAF at locus

2. In carriers of the reference aa genotype at locus 1, there is a difference in pRAF at locus 2 between subjects with and without the endophenotype impairment, while in carriers of the AA genotype at locus 1, only the affected subjects with an endophenotype impairment differ in pRAF at locus 2 from the other three groups of subjects. The B allele frequency $E[X_2|X_1, Y_1, Y_2]$ is well described by the linear model on the left side of table S3 ($SSE \leq 0.5\%$ for unrelated subjects). As expected from property 14, $E[X_2|X_1, Y_1 = 0, Y_2 = 0]$ and $E[X_2|X_1, Y_1 = 0, Y_2 = 1]$ are identical linear functions of X_1.

If we consider only the genotype at locus 2, the one-locus expressions for the pRAF are on the right side of table S3. These expressions are exact in unrelated subjects, since scenario P4 satisfies condition 11. The marginal association between the allele count X_2 and the endophenotype provides power to detect locus 2. The largest effect size is however obtained when considering locus 1 and 2 together (0.246 - (-0.069) = 0.32 with pRAF = 0.3). In the simulations, power is highest for the tests of association with the dichotomous endophenotype, either the single locus MQLSe and GDTe tests or the cendo test conditionning on locus 1 (Figure 4D).

4.2.5. Scenario P5

Scenario P5 differs from scenario P1 in that the two loci interact to cause the disease, irrespective of the endophenotype status, as reflected in the pRAF at locus 2 presented on Figure 3E. The B allele frequency $E[X_2|X_1, Y_1, Y_2]$ is well described by the linear models on the left side of table S4 ($SSE = 0.03\%$ for unrelated subjects). The one-locus models best fitting the B allele frequency considering only the genotype at locus 2 are on the right side of table S4 ($20\% < SSE < 32\%$ for unrelated subjects).

In the two linear models, terms involving the endophenotype are smaller than the terms involving the disease. This suggests that a classical interaction analysis of locus 1 and 2 with respect to the disease alone would be powerful, and that the modeling of the endophenotype would provide no benefit. This is indeed what the simulations show, with the highest power being achieved by a conditional test of locus 2 given locus 1 under the classical two-locus model for the dichotomous disease phenotype (cdisease) (Figure 4D).

4.3. Selection of transition models of genetic scenarios

As we saw in section 2.2.2, mechanisms where the disease status Y_2 does not depend directly on the genotype at locus 1 represented by X_1 as specified in equation 9 imply that the pRAF at locus 2 does not depend on X_1. Therefore, when the goal is to detect association to locus 2, X_1 brings no information given the endophenotype Y_1 and should not be included in the analysis. The analysis strategies should then be selected among one-locus models.

We consider here the alternative situation where the disease status Y_2 depends directly on X_1, X_2 and Y_1. We examine transition models where Y_1 depends only on X_1 (equation 8), a restriction that is not possible with polytomous models. The values of the coefficients corresponding to each transition model are given in Table 2. Scenarios T1 and T3 satisfy condition 11.

4.3.1. Scenario T1

Like in scenario P1, locus 2 influences disease risk only in presence of an endophenotype impairment and a locus 1 risk allele, through the triple interaction term. The phenotype-specific frequency of the risk allele B of locus 2 when the B allele population frequency equals 0.3 is displayed in Figure 5A. On the plot, one can see that, in presence of the A allele, affected subjects with the endophenotype impairment differ from the other three groups which all have an approximately constant pRAF of 0.3. The B allele frequency $E[X_2|X_1, Y_1, Y_2]$ is well described by the linear model reported in table 5 ($SSE \leq 0.5\%$ for unrelated subjects). In ascertained families, the contrast between affected and unaffected subjects is somewhat attenuated compared to unrelated subjects, but the same linear model provides a good fit to the data ($SSE = 17\%$ for a pRAF of 0.3).

If we consider only the genotype at locus 2, the one-locus expressions for the pRAF are on the right side of table 5. These expressions are exact in unrelated subjects, since scenario T1 satisfies condition 12, and provide an excellent fit in ascertained families ($SSE = 4\%$ for a pRAF of 0.3). Among subjects with the endophenotype impairment ($Y_1 = 1$), the contrast between affected and unaffected subjects (0.027 - (-0.001) = 0.028 with pRAF = 0.3) is attenuated compared to its maximal value in carriers of two A risk alleles at locus 1 (0.17 - (-0.01) = 0.18). Even though the true underlying mechanism is a transition model, the existence of a genetic effect on disease only in presence of an endophenotype impairment gives the test of the interaction term β_{33} for the category $Y_1 = 1, Y_2 = 1$ under a polytomous model greater power than tests of association to the dichotomous disease status or endophenotype (Figure 6A).

4.3.2. Scenario T2

Under this scenario, there is no triple interaction $X_1X_2Y_1$ on the multiplicative risk of disease. The interaction term between locus 1 and 2 X_1X_2 is present in subjects with and without an endophenotype impairment, but multiplies a larger risk in subjects with an impairment than in those without. The pRAF at locus 2 does not differ substantially between subjects with and without an impairment on the endophenotype (Figure 5B). Consequently, the best fitting linear model ($SSE \leq 3.3\%$ for unrelated subjects) does not involve Y_1, as can be seen on the left side of table S5.

If we consider only the genotype at locus 2, the one-locus expression for the pRAF is a simple function of Y_2 (see right side of table S5). The absence of Y_1 indicates that the endophenotype Y_1 need not be considered in the analysis while the higher contrast between affected and unaffected subjects in carriers of the AA genotype at locus 1 than in the general population (0.160 - (-0.003) = 0.163 compared to 0.019 with pRAF=0.3) suggests that conditioning on locus 1 will provide a power gain. In the simulation study, the higher power of the conditional test of locus 2 given locus 1 under the classical two-locus model for the dichotomous disease phenotype (cdisease) or the polytomous model (β_{33}) compared to the single locus GDTd counterpart is consistent with this observation, but we also notice that single locus M_{QLS} tests achieve a level of power similar to the conditional tests. (Figure 6B).

4.3.3. Scenario T3

Scenario T3 differs from scenario T1 in that the locus 2 genotype has an effect on the risk of the disease independently of the locus 1 genotype. This is reflected in the pRAF presented on Figure 5C. The B allele frequency $E[X_2|X_1, Y_1, Y_2]$ is well described by the linear models reported in table S6 ($SSE \leq 1.0\%$ for unrelated subjects).

If we consider only the genotype at locus 2, the one-locus expressions for the pRAF are exact, since scenario T3 satisfies condition 12(see right side of table S6), and provide an excellent fit in ascertained families ($SSE = 3\%$ for a pRAF of 0.3). Among subjects with the endophenotype ($Y_1 = 1$), the contrast between affected and unaffected subjects (0.111 -

(-0.004) = 0.12 with pRAF = 0.3) remains sufficiently important compared to its maximal value in carriers of the AA genotype at locus 1 (0.18 + 0.08 - (- 0.016) = 0.28) to confer good power to a single-locus analysis of X_2. This is what we observe in the simulations, where a test of $\beta_3(1L)$ in the polytomous model and the MQLSde test with X_2 achieving both a power of 0.87 (Figure 6C). Tests of association with the disease status ignoring the endophenotype have lower power than the tests taking the endophenotype into account.

5. Discussion

We have addressed in this work a topic that has been little studied before: the genetic association patterns arising when a disease status and a dichotomous endophenotype are considered jointly, in function of the relationships between the disease, the endophenotype and genetic variants. We described these relationships using two established types of prospective models for bivariate dichotomous outcomes: polytomous models and transition models. We have focused on two-locus systems where the presence of risk alleles at locus 2 shifts the effect of locus 1 from a risk of endophenotype impairment without disease to a risk of endophenotype impairment with disease.

We examined expected allele frequency at locus 2 in function of the disease and endophenotype statuses and genotype at locus 1, since it is driving power to detect a locus. First, we presented special cases of allele frequency models, in particular when locus 2 has an effect on the risk of disease only in presence of endophenotype impairment. We derived conditions that polytomous and transition models must satisfy to obtain these special cases. Next, since allele frequencies or frequency ratios at locus 2 are not interpretable explicit expressions of other parameters under the two-locus two-phenotype system considered, we examined a selection of scenarios of polytomous and transition models that represent variations of our base scenario. For each of these, we approximated allele frequencies at locus 2 in unrelated subjects by a linear function of the disease and endophenotype statuses and allele proportion at locus 1. We also simulated data in families ascertained for the presence of multiple affected subjects. In these simulations, we included familial and disease-endophenotype dependence. This simulation scheme produced higher prevalences of endophenotype impairment in both affected and unaffected members of disease-ascertained families compared to unrelated subjects because of familial correlation, the association parameters for the susceptibility loci remaining unchanged. This increased prevalence is consistent with epidemiological evidence for various diseases and endophenotypes (see [20] for a review covering schizophrenia and bipolar disorder).

We draw two general observations from our examination of a selection of genetic scenarios represented by polytomous and transition models. First, under mechanisms where the locus 2 risk allele has an effect exclusively in subjects with the endophenotype, the risk allele frequency differences between affected and unaffected subjects are much greater in the subset of subjects with an endophenotype impairment than in those without such impairment, and power gains are obtained when testing association under a joint disease-endophenotype model, either two-locus (scenarios P1, P3, T1 and T3) or single locus (scenario T3) compared to testing association to the disease status alone. This advantage vanishes however when locus 2 has a main effect on the risk of disease, even though the allele frequency at locus 2 still depends strongly on the presence or not of the endophenotype (scenarios P2, P5 and T2). The power advantage also vanishes when locus 2 has a main effect on the risk of endophenotype impairment, but this time to the benefit of tests of association of the endophenotype status alone (scenario P4). When detecting association to the endophenotype alone, inference of an effect of the locus on disease status can only be made indirectly, through the endophenotype - disease association.

Second, when interaction effects between locus 1 and 2 are the only effects of these loci on the phenotype, the locus 2 allele frequency differences between affected and unaffected subjects are much greater in carriers of risk genotypes at locus 1, and tests of association conditioning on locus 1 genotype are more powerful than association tests with locus 2 alone. This observation was made previously with dichotomous disease outcomes in the context of gene-environment interaction where the environmental exposure plays the role of locus 1 [21]. Our contribution is to show that the higher power of conditional tests applies under joint modeling of a disease and endophenotype (scenarios P1, P3, T1 and T3), and not only when the strongest association is seen with the disease alone (scenarios P2, P5 and T2) or the endophenotype alone (scenario P3). When locus 2 has a main effect on the risk of disease, the conditional tests retain a power similar to the tests of locus 2 alone, an observation also in line with the study of [21].

Our observations are subject to a number of limitations. Given the high dimension of the classes of two-locus mechanisms of a disease and an endophenotype that we considered, the parameter space could not be explored exhaustively, and different behaviours may be observed with other combinations of parameter values. In all mechanisms considered the effect of each locus was a function of the risk allele proportion at the locus (multiplicative odds ratios model). We anticipate that most of our observations will hold under other mechanisms of locus effects, in particular dominant and recessive effects, where the variables X_1 and X_2 become indicators of the presence of risk genotypes. The application of allelic models to data arising from other forms of locus effects may nevertheless have unexpected impacts on the relative power of the testing strategies that we considered. Multiple endophenotypes may also be related to the same disease. If each endophenotype represents a distinct biological trait, one can hope that each will be influenced by distinct loci, so that each endophenotype with the locus pairs associated to it and the disease will be well described by two-locus mechanisms described in this work. It is however possible for the same locus to be associated to multiple endophenotypes, such that models for a single endophenotype and the disease do not result in a signal sufficient to detect new loci by conditioning on a known one. We also assumed that the markers tested were the actual risk variants or in perfect linkage disequilibrium (LD) with the risk variant. Imperfect LD between a marker and the causal variant typically attenuates genetic effects; certain effects may be attenuated to a greater extent than others, altering the relative power of the testing strategies. Our study is also limited to two risk loci in linkage equilibrium and does not address the impact of LD when the two risk loci are nearby. Finally, our joint disease - endophenotype analysis under a two-locus model may not have been the most powerful possible, since it was limited to within-family score tests, which tend to be less powerful than population-level tests, as can be noted in this simulation study when comparing M_{QLS} and the GDT. The unavailability of a two-locus analysis approach with the optimality properties of M_{QLS} prevented us from performing such population-level analysis.

6. Conclusions

In conclusion, we showed that jointly modeling a disease, an endophenotype and a marker at a known susceptibility locus improves the power to detect another locus whose effect on disease susceptibility is expressed only in subjects with an endophenotype impairment, but that modest main effects on the risk of disease suffice to revert the advantage to association tests with the dichotomous disease status alone. Joint modeling of disease and endophenotype should be used only in parallel with standard association testing with the disease status.

7. List of abbreviations used

OR: Odds Ratio; SSE: Sum of Squared Errors; GDT: Generalized Disequilibrium Test; M_{QLS}: Maximum Quasi-Likelihood Score; pRAF: population Risk Allele Frequency ; LD: linkage disequilibrium

Author's contributions

AB was the conceptor and designer of the study and drafted the manuscript. JC performed the data simulation and statistical analyses. He also participated in the creation of the tables and figures.

Acknowledgements

Financial support from the Canadian Institutes of Health research (CIHR, grant MOP-106448) is gratefully acknowledged. A. Bureau was supported by a research fellowship from the Fonds de recherche du Québec - Santé.

References

1. Gottesman II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. Am J Psychiatry. 2003;160(4):636–45.

2. Szatmari P, Maziade M, Zwaigenbaum L, Merette C, Roy MA, Joober R, et al.

Informative phenotypes for genetic studies of psychiatric disorders. Am J Med Genet B Neuropsychiatr Genet. 2007;144B(5):581–8.

3. Chen HY, Kittles R, Zhang W. Bias correction to secondary trait analysis with case-control design. Stat Med. 2013;32(9):1494–508.

4. Ghosh A, Wright FA, Zou F. Unified Analysis of Secondary Traits in Case-Control Association Studies. J Am Stat Assoc. 2013;108(502).

5. Williams JT, Van Eerdewegh P, Almasy L, Blangero J. Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results. Am J Hum Genet. 1999;65(4):1134–47.

6. Glahn DC, Curran JE, Winkler AM, Carless MA, Kent J J W, Charlesworth JC, et al. High dimensional endophenotype ranking in the search for major depression risk genes. Biol Psychiatry. 2012;71(1):6–14.

7. Liu J, Pei Y, Papasian CJ, Deng HW. Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. Genet Epidemiol. 2009;33(3):217–27.

8. Wilhelmsen KC. The feasibility of genetic dissection of endophenotypes. Psychophysiology. 2014;51(12):1337–8.

9. Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SL, Peyser PA, et al. SNP set association analysis for familial data. Genet Epidemiol. 2012;36(8):797–810.

10. Chen WM, Manichaikul A, Rich SS. A generalized family-based association test for dichotomous traits. Am J Hum Genet. 2009;85(3):364–76.

11. Cordell HJ. Properties of case/pseudocontrol analysis for genetic association studies:

Effects of recombination, ascertainment, and multiple affected offspring. Genet Epidemiol. 2004;26(3):186–205.

12. Thornton T, McPeek MS. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet. 2007;81(2):321–37.

13. Thornton T, McPeek MS. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am J Hum Genet. 2009;86(2):172–84.

14. Bureau A, Croteau J, Chagnon YC, Roy MA, Maziade M. Extension of the generalized disequilibrium test to polytomous phenotypes and two-locus models. Frontiers in Genetics. 2014;5:258.

15. Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. 2nd ed. Wiley series in probability and statistics. Hoboken, N.J.: Wiley; 2011.

16. Diggle P. Analysis of longitudinal data. 2nd ed. Oxford statistical science series. Oxford ; New York: Oxford University Press; 2002.

17. Fitzmaurice GM, Laird NM, Zahner GE, Daskalakis C. Bivariate logistic regression analysis of childhood psychopathology ratings using multiple informants. Am J Epidemiol. 1995;142(11):1194–203.

18. Bureau A, Croteau J, Couture C, Vohl MC, Bouchard C, Perusse L. Estimating genetic effect sizes under joint disease-endophenotype models in presence of gene-environment interactions. Front Genet. 2015;6:248.

19. Schmidt M, Hauser ER, Martin ER, Schmidt S. Extension of the SIMLA package for generating pedigrees with complex inheritance patterns: environmental covariates, gene-gene and gene-environment interaction. Stat Appl Genet Mol Biol. 2005;4:Article15.

20. Ivleva EI, Morris DW, Moates AF, Suppes T, Thaker GK, Tamminga CA. Genetics and intermediate phenotypes of the schizophrenia–bipolar disorder boundary [Journal Article]. Neurosci Biobehav Rev. 2010;34(6):897–921.

21. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. Hum Hered. 2007;63(2):111–9.

Figure legends

Fig. 1.— **Graphical models.** Directed graphical representation of models of the relationships between disease status Y_2, endophenotype value Y_1 and the genotype at two loci represented by X_1 and X_2. A: Endophenotype Y_1 depends only on locus 1 (condition 5); B: General model, only assuming marginal independence, or linkage equilibrium, between locus 1 and locus 2.

Fig. 2.— **Structure of simulated families.** A phenotypic configuration typical of the simulated families is depicted.

Fig. 3.— **Theoretical risk allele frequencies under polytomous models.** Theoretical frequency of risk allele B at locus 2 by number of A alleles at a known susceptibility locus 1, disease and endophenotype status, when the population B allele frequency = 0.3. Panels A to E represent polytomous models P1 to P5.

Fig. 4.— **Power of family-based association for polytomous models.** Power of the various family-based association testing strategies described in section 2.2.2 at a significance level $\alpha = 0.05$. Panels A to E represent polytomous models P1 to P5 with a population risk allele frequency at locus 2 = 0.3.

Fig. 5.— **Theoretical risk allele frequencies under transition models.** Theoretical frequency of risk allele B at locus 2 by number of A alleles at a known susceptibility locus 1, disease and endophenotype status, when the population B allele frequency = 0.3. Panels A to C represent transition models T1 to T3.

Fig. 6.— **Power of family-based association for transition models.** Power of the various family-based testing strategies described in section 2.2.2 at a significance level $\alpha = 0.05$. Panels A to C represent transition models T1 to T3 with a population risk allele frequency at locus 2 = 0.3.

Tables

Table 1: Regression coefficients of the example polytomous model scenarios

	β_{11}	β_{12}	β_{13}	β_{21}	β_{22}	β_{23}	β_{31}	β_{32}	β_{33}
scenario P1	log(2)	0	-log(2)	0	0	0	0	0	log(16)
scenario P2	log(2)	0	-log(2)	0	log(4)	0	0	log(4)	log(4)
scenario P3	log(2)	0	-log(2)	log(4)	0	0	log(4)	0	log(4)
scenario P4	log(2)	log(2)	-log(2)	0	0	0	0	log(2)	log(4)
scenario P5	log(2)	0	-log(2)	0	0	log(16)	0	0	log(16)

Table 2: Regression coefficients of the example scenarios of the disease status within transition models [a]

	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
scenario T1	0	0	0	log(8)	0	0	log(4)
scenario T2	0	0	log(4)	log(8)	0	0	0
scenario T3	0	0	0	log(5)	0	log(2)	log(4)

[a] The model for the endophenotype is of the form 5 with $\beta_1 = \log(2)$.

Table 3: Disease and endophenotype impairment prevalences in unrelated individuals and ascertained families.

	Unrelated individuals				Ascertained families			
	$P(Y_2)$[a]	$P_0(Y_1)$[b]	$P_1(Y_1)$[c]	OR[d]	$P(Y_2)$[a]	$P_0(Y_1)$[b]	$P_1(Y_1)$[c]	OR[d]
polytomous scenarios								
scenario P1	0.0076	0.125	0.534	8.04	0.147	0.191	0.646	7.83
scenario P2	0.0076	0.125	0.520	7.60	0.149	0.189	0.629	7.37
scenario P3	0.0089	0.125	0.524	7.70	0.153	0.191	0.628	7.27
scenario P4	0.0086	0.098	0.573	12.38	0.148	0.152	0.690	12.68
scenario P5	0.0081	0.125	0.500	7.00	0.149	0.192	0.612	6.76
transition scenarios								
scenario T1	0.0106	0.120	0.540	8.61	0.160	0.173	0.678	10.21
scenario T2	0.0108	0.120	0.528	8.18	0.162	0.174	0.666	9.57
scenario T3	0.0094	0.120	0.486	6.93	0.156	0.181	0.633	7.92

[a] $P(Y_2 = 1)$, the disease prevalence.

[b] $P(Y_1 = 1|Y_2 = 0)$, the endophenotype prevalence in the unaffected subjects.

[c] $P(Y_1 = 1|Y_2 = 1)$, the endophenotype prevalence in the affected subjects.

[d] $P(Y_1=1|Y_2=1) \Big/ P(Y_1=1|Y_2=0)$

Table 4: Values of the coefficients of linear models of the population risk allele frequency at locus 2 for polytomous phenotype scenario P1.

		Model conditional on X_1				Model ignoring X_1		
	pRAF[a]	Intercept	X_1	X_1Y_1	$X_1Y_1Y_2$	Intercept	Y_1	Y_1Y_2
Unrelated[b]	0.1	0.100	0.004	-0.029	0.206	0.100	-0.004	0.024
	0.3	0.300	0.009	-0.070	0.396	0.301	-0.008	0.061
	0.5	0.500	0.009	-0.086	0.404	0.501	-0.010	0.076
Ascertained families[c]	0.3	0.315	0.050	-0.083	0.316	0.322	-0.010	0.048

[a] Population risk allele frequency

[b] Least square fit to expected allele frequency at locus 2 for each combination of explanatory variables

[c] Least square fit to empirical allele frequency at locus 2 for each combination of explanatory variables in data simulated under the familial dependence simulation model with $\sigma_2 = 1.0, \alpha = \log(4) = 1.386$ and $\nu = 0.01$ in 3-generation 16-members family depicted in Figure 2 ascertained for at least a cousin pair affected by the disease

Table 5: Values of the coefficients of linear models of the population risk allele frequency at locus 2 for transition scenario T1.

		Model conditional on X_1			Model ignoring X_1		
	pRAF[a]	Intercept	X_1Y_1	$X_1Y_1Y_2$	Intercept	Y_1	Y_1Y_2
Unrelated[b]	0.1	0.100	-0.004	0.078	0.100	-5e-4	0.011
	0.3	0.300	-0.010	0.170	0.300	-0.001	0.027
	0.5	0.500	-0.013	0.180	0.501	-0.002	0.034
Ascertained families[c]	0.3	0.309	-0.010	0.143	0.310	-0.003	0.025

[a] Population risk allele frequency

[b] Least square fit to expected allele frequency at locus 2 for each combination of explanatory variables

[c] Least square fit to empirical allele frequency at locus 2 for each combination of explanatory variables in data simulated under the familial dependence simulation model with $\sigma_2 = 1.0, \alpha = \log(4) = 1.386$ and $\nu = 0.01$ in 3-generation 16-members family depicted in Figure 2 ascertained for at least a cousin pair affected by the disease

Additional Files

Appendix revision.pdf

Appendix

Section about simulating a bidimensional dichotomous phenotype in families, simulation of Y_1 and genotypes, family ascertainment, and proof of properties 10 and 13.

SupplementaryTables.pdf

Supplementary Tables

Tables of coefficients of linear models for polytomous scenarios P2, P3, P4 and P5, and transition scenarios T2 and T3 (tables S1, S2, S3, S4, S5 and S6).