frontiers in GENETICS

METHODS ARTICLE
published: 08 August 2014
doi: 10.3389/fgene.2014.00258

Extension of the generalized disequilibrium test to polytomous phenotypes and two-locus models

Alexandre Bureau[1,2]*, Jordie Croteau[2], Yvon C. Chagnon[2], Marc-André Roy[2,3] and Michel Maziade[2,3]

[1] Département de Médecine Sociale et Préventive, Université Laval, Québec, QC, Canada
[2] Centre de Recherche de L'Institut Universitaire en Santé Mentale de Québec, Québec, QC, Canada
[3] Département de Psychiatrie et Neurosciences, Université Laval, Québec, QC, Canada

Edited by:
Rongling Wu, Pennsylvania State University, USA

Reviewed by:
Guohua Zou, Chinese Academy of Sciences, China
Ashok Ragavendran, Massachusetts General Hospital, USA

***Correspondence:**
Alexandre Bureau, Département de Médecine Sociale et Préventive, Université Laval, Pavillon Vandry, 1050 rue de la Médecine, Local 2457, Québec, QC G1V 0A6, Canada
e-mail: alexandre.bureau@fmed.ulaval.ca

We extend the usual logistic model between a dichotomous phenotype and an allele count in two ways: a polytomous phenotype with $K > 2$ levels, and modeling of allele counts at two unlinked marker loci. Inference is based on within-family information to guard against potential bias due to population genetic structure. Score tests of the model coefficients taking into account the correlation between relatives in entire pedigrees are derived as an extension of the Generalized Disequilibrium Test (GDT). Simulations confirm that the tests have the expected statistical properties, and that their power exceeds that of the GDT under a favorable scenario. The score tests are illustrated with candidate genetic markers, a major psychosis phenotype and a cognitive endophenotype in large kindreds from Eastern Quebec.

Keywords: conditional likelihood, endophenotype, family-based association, kinship, major psychosis, polytomous logistic model, score test

1. INTRODUCTION

Studies of the association between a phenotype and genetic markers are commonly performed on the members of families of various sizes. While methods to estimate association parameters and test the null hypothesis of absence of association (possibly coupled with absence of genetic linkage) with dichotomous phenotypes in family samples are well developed (see for instance chapter 12 of Ziegler and König, 2010), methods are lacking to analyze polytomous phenotypes. Such phenotypes can arise when a disease has multiple subtypes (Guey et al., 2010) or when two dichotomous phenotypes are considered simultaneously. The latter occurs when endophenotypes are measured in genetic studies to better capture phenotypic complexity. Endophenotypes are traits related to a disease and believed to be influenced by fewer genes (Gottesman and Gould, 2003). A dichotomous disease status and a dichotomous endophenotype create a four category phenotype. Comparisons between analyzing a polytomous phenotype vs. a dichotomous one have not been done for family studies due to the lack of analysis methods for polytomous phenotypes.

We focus in this paper on a within-family analysis, conditional on phenotype and genotype observed in each family. Such approach is well known to protect against confounding due to population stratification. Families where multiple phenotypic categories are represented provide the most information on the relationship between a polytomous phenotype and genetic markers. Since families extending over multiple generations typically need to be recruited to obtain a large number of phenotyped subjects, we required that the methods for dichotomous traits that we generalize to polytomous traits be applicable to extended families. For a score test of association, we selected the Generalized disequilibrium test (GDT) of Chen et al. (2009).

In previous work, we showed by simulation that conditioning on a marker at a known disease susceptibility locus increased power to detect linkage to new loci interacting with that disease susceptibility locus (Bureau et al., 2009, 2012). Similar power gains are expected in association analysis, as conditioning on a known environmental risk factor increases power to detect loci interacting with the exposure (Kraft et al., 2007). Models involving genetic markers at two distinct loci are needed for analyses conditional on the genotype of known disease susceptibility markers and also to model the relationship between pairs of loci. Multi-category phenotypes present a larger realm of possibilities of interplay between multiple loci than dichotomous traits, making multilocus modeling even more important to capture the actual effects. This is why we derive score tests under two-locus models, with one marker at each locus, in addition to one-locus models. The Type I error and the power of tests of various combinations of regression coefficients are assessed using simulation. The tests are also illustrated with candidate genetic markers, a major psychosis phenotype and a cognitive endophenotype in the Eastern Quebec kindred study.

2. METHODS

We extend the GDT of Chen et al. (2009) in two ways: by allowing the outcome Y to have $K > 2$ levels, and by allowing the odds of the outcome categories to depend on two or more variables X, coding the genotype of markers at two mutually unlinked marker loci. As in the original GDT, X represents the count of a particular form of the DNA sequence at the marker, called allele. We begin by deriving the score statistic from the conditional likelihood for a polytomous outcome Y with a general vector X of allelic count terms (possibly including product terms). Then we derive expressions for particular forms of terms in X.

The polytomous model for subject i with a general X_i vector can be written

$$\log\left(\frac{P[Y_i = k|X_i]}{P[Y_i = K|X_i]}\right) = \mu_k + \underset{\sim}{\beta}'_k X_{ki}, k = 1, \cdots, K-1 \quad (1)$$

where X_{ki} is the sub-vector of X_i containing the allelic terms related to level (category) k and $\underset{\sim}{\beta}_k$ the sub-vector of the full coefficient vector $\underset{\sim}{\beta}$ applicable to level k (in this general formulation, β coefficients can either be distinct for each level k or can be common to multiple levels of k).

Without loss of generality, we assume that the n genotyped pedigree members with an observed phenotype are ordered such that the first n_1 subjects are in outcome category $Y = 1$, the n_2 following subjects are in outcome category $Y = 2$ and so on up to the last n_K subjects with $Y = K$.

With $K = 2$ and a single X ($\underset{\sim}{\beta}_1 = \beta$ a scalar, without covariates), Chen et al. (2009) showed that the contribution of the family to the score statistic from the conditional likelihood P to test the null hypothesis $\beta = 0$ has the form:

$$S^{GDT} = \frac{\partial \log P}{\partial \beta}\Big|_{\beta=0} = \frac{1}{n}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}(X_i - X_j) \quad (2)$$

We show in Supplementary Material that the contribution of a family to the score statistic for the coefficient β_h component of $\underset{\sim}{\beta}$ when testing the global null hypothesis that the full $\underset{\sim}{\beta} = 0$ under a polytomous model is:

$$S^{(h)} = \frac{\partial \log P}{\partial \beta_h}\Big|_{\underset{\sim}{\beta}=0} = \frac{1}{n}\left[\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}\left(X^{(h)}_{1i} - X^{(h)}_{1j}\right) + \cdots \right.$$
$$\left. + \sum_{i=n-(n_{K-1}+n_K)+1}^{n-n_K}\sum_{j\in E_{K-1}}\left(X^{(h)}_{(K-1)i} - X^{(h)}_{(K-1)j}\right)\right] \quad (3)$$

where $E_{K-1} = \{1, \cdots, n-(n_{K-1}-n_K), n-n_K+1, \cdots, n\}$ and $X^{(h)}_{ai}$ is the slice of X_{ai} related to the coefficient β_h. If β_h is involved only in the logistic function between levels a and K, then the score statistic simplifies to:

$$S^{(h)} = \frac{\partial \log P}{\partial \beta_h}\Big|_{\underset{\sim}{\beta}=0} = \frac{1}{n}\sum_{i\in E_a}\sum_{j\in E_a^c}\left(X^{(h)}_{ai} - X^{(h)}_{aj}\right) \quad (4)$$

where $E_a = \{n_1 + \cdots + n_{a-1} + 1, \cdots, n_1 + \cdots + n_a\}$ for $a > 1$ and $E_1 = \{1 \cdots n_1\}$.

The advantage of expression 3 is that a closed-form expression for the variance of $S^{(h)}$ and the covariance of $S^{(g)}$ and $S^{(h)}$ for coefficients β_g and β_h can be derived, following the steps of Chen et al. It is also easier to interpret. When the tested coefficient belongs to the logistic function attached to a single outcome category and the score statistic reduces to expression 4, it is a contrast of the value of the corresponding X term between subjects in the outcome category and subjects in all other categories.

Letting $v[S]$ be an estimate of the variance-covariance matrix of S, the null hypothesis that $\underset{\sim}{\beta} = 0$ can then be tested with the statistic

$$T = S'v[S]^{-1}S$$

which follows a χ^2 distribution with degrees of freedom equal to the rank of $\underset{\sim}{\beta}$ under the null.

When testing the sub null hypothesis $\beta_{h_1} = \cdots = \beta_{h_m} = 0$ for any subset of indices $h_1, \cdots, h_m$, the other coefficients are free to differ from 0 and the derivation in Supplementary Material no longer applies. We adopt here the approach Chen et al. (2009) apply to model covariates, which is to weight the pairwise differences according to a model of the outcome Y as a function of the predictors with free coefficients under the null hypothesis. The score statistic for the component β_h of the subset of coefficients tested then becomes

$$S^{(h)} = \frac{\partial \log P}{\partial \beta_h}\Big|_{\beta_{h_1}=\cdots=\beta_{h_m}=0} = \sum_{i\in E_a}\sum_{j\in E_a^c} C_{ij}\left(X^{(h)}_{ai} - X^{(h)}_{aj}\right) \quad (5)$$

where the weights C_{ij} can be derived from score equations for β_h under the pairwise formulation of Liang and Stewart (1987) (see Supplementary Material), leading to the following functions of the coefficients $\underset{\sim}{\alpha}$ of a polytomous logistic model of Y as a function of the predictors $X^{(c)}, c = \{l : l \notin (h_1, \cdots, h_m)\}$ when the variability from estimating the $\underset{\sim}{\alpha}$ is neglected:

$$C_{ij} = \frac{2}{N}\frac{1}{\left(1 + exp\left\{\left(X^{(c)}_i - X^{(c)}_j\right)'\left(\underset{\sim}{\alpha}_{Y_i} - \underset{\sim}{\alpha}_{Y_j}\right)\right\}\right)} \quad (6)$$

where $\underset{\sim}{\alpha}_K = 0$.

Adapting Chen et al. (2009)'s Equation 2 from the dichotomous to the polytomous case gives the following expression for the weights instead:

$$C_{ij} = \frac{8}{N}\frac{exp\left\{\left(X^{(c)}_i - X^{(c)}_j\right)'\left(\underset{\sim}{\alpha}_{Y_i} - \underset{\sim}{\alpha}_{Y_j}\right)\right\}}{\left(1 + exp\left\{\left(X^{(c)}_i - X^{(c)}_j\right)'\left(\underset{\sim}{\alpha}_{Y_i} - \underset{\sim}{\alpha}_{Y_j}\right)\right\}\right)^3} \quad (7)$$

We estimate the coefficients $\underset{\sim}{\alpha}$ using generalized estimating equations (GEEs) with an independence working correlation matrix. With this approach the null hypothesis that the component $\beta_h = 0$ can be tested with the statistic

$$Z^{(h)} = S^{(h)}/\sqrt{v[S^{(h)}]}$$

which follows approximately a standard normal distribution under the null, when the weights are defined in such a way that the expectation of $S^{(h)}$ is 0. The weight definition will only have an impact on power. The joint null hypothesis $\beta_{h_1} = \cdots = \beta_{h_m} = 0$ for any subset of indices $h_1, \cdots, h_m$ can be tested with the statistic

$$T = (S_{h_1}, \cdots, S_{h_m})(v[S_{h_1}, \cdots, S_{h_m}])^{-1}(S_{h_1}, \cdots, S_{h_m})' \quad (8)$$

which follows approximately a χ^2 distribution with m degrees of freedom under the null.

The variance of $S^{(h)}$ depends on whether the null hypothesis refers only to absence of association, or to absence of genetic linkage and association. In the first case, the null distribution of $S^{(h)}$ allows genetic linkage at the locus, and the identical-by-descent (IBD) sharing proportions in the variance estimate must be the actual IBD sharing proportions at the locus π_{hij} (Chen et al., 2009). For the second case, or when IBD is unknown, π_{hij} can be substituted by twice the kinship coefficients ϕ_{ij}, which is constant at all loci. The general expression for the variance of $S^{(h)}$ and covariance between $S^{(h)}$ and $S^{(g)}$ is given in Supplementary Material. When $S^{(h)}$ takes the form 4, $X_{ai}^{(h)}$ is a main effect term, say X_1, and the actual IBD sharing proportions π_{hij} are used then

$$\begin{aligned} Var[S^{(h)}] &= Var\left[\sum_{i\in E}\sum_{j\in E^c} C_{ij}(X_{ai}^{(h)} - X_{aj}^{(h)})\right] \qquad (9)\\ &= \sum_{i,k\in E}\sum_{j,l\in E^c} C_{ij}C_{kl}Cov[X_{ai}^{(h)} - X_{aj}^{(h)}, X_{ak}^{(h)} - X_{al}^{(h)}]\\ &= \sum_{i,k\in E}\sum_{j,l\in E^c} C_{ij}C_{kl}\left(Cov[X_{1i}, X_{1k}] + Cov[X_{1j}, X_{1l}]\right.\\ &\quad \left. - Cov[X_{1i}, X_{1l}] - Cov[X_{1j}, X_{1k}]\right)\\ &= \sum_{i,k\in E}\sum_{j,l\in E^c} C_{ij}C_{kl}\left(\pi_{1ik} + \pi_{1jl} - \pi_{1il} - \pi_{1jk}\right)\sigma_1^2 \end{aligned}$$

The within-family variance of X_1, σ_1^2, is estimated as described in Supplementary Material to obtain the estimate $v[S^{(h)}]$ of $Var[S^{(h)}]$. With equal weights for all pairs, the computation involving the IBD sharing probabilities can be simplified as explained in Supplementary Material.

When $X_{ai}^{(h)}$ is instead a product term, say X_1X_2, then

$$\begin{aligned} Var[S^{(h)}] &= Var\left[\frac{1}{n}\sum_{i\in E}\sum_{j\in E^c}(X_{ai}^{(h)} - X_{aj}^{(h)})\right]\\ &= \frac{1}{n^2}\sum_{i\in E}\sum_{j\in E^c}\sum_{k\in E}\sum_{l\in E^c}\\ &\quad \begin{pmatrix} Cov[X_{1i}X_{2i}, X_{1k}X_{2k}] + Cov[X_{1j}X_{2j}, X_{1l}X_{2l}]\\ -Cov[X_{1i}X_{2i}, X_{1l}X_{2l}] - Cov[X_{1j}X_{2j}, X_{1k}X_{2k}] \end{pmatrix}\\ &= \frac{1}{n^2}\sum_{i\in E}\sum_{j\in E^c}\sum_{k\in E}\sum_{l\in E^c}\\ &\quad \left(\pi_{1ik}\pi_{2ik} + \pi_{1jl}\pi_{2jl} - \pi_{1il}\pi_{2il} - \pi_{1jk}\pi_{2jk}\right)\sigma_{12}^2 \end{aligned}$$

where the within-family variance of the product term X_1X_2, σ_{12}^2, is estimated as described in Supplementary Material.

2.1. APPLICATION TO THE JOINT MODELING OF TWO DICHOTOMOUS TRAITS USING TWO-LOCUS MODELS

The joint analysis of two dichotomous traits represents an important special case of a polytomous phenotype with four categories. We illustrate such a phenotype by referring to a dichotomous disease trait Y_2 and a dichotomous endophenotype Y_1, as defined in the introduction.

We consider here polytomous models for two markers at unlinked loci which may interact to cause the disease and endophenotype impairment. We assume that association of locus 1 to the endophenotype impairment $Y_1 = 1$ and possibly to the disease $Y_2 = 1$ has already been established, and that we want to detect locus 2, which is undetectable in single-locus analyses, by conditioning on locus 1 with which it interacts. This leads to null hypotheses on a subset of coefficients tested with a statistic as defined in Equation 8.

A first option is to use the full model with distinct coefficients for each disease/endophenotype combination contrasted to the reference category of absence of both the disease and endophenotype impairment. This model is:

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

The null hypothesis of the conditional test of locus 2 given locus 1 under the full model is formulated as:

$$\beta_{12} = \beta_{13} = \beta_{22} = \beta_{23} = \beta_{32} = \beta_{33} = 0 \qquad (11)$$

When the null is rejected, insights on the phenotype category driving the signal can be obtained by examining the Z statistics for each coefficient and the p-values associated to the tests of the subsets of coefficients (β_{12}, β_{13}),(β_{22}, β_{23}) and (β_{32}, β_{33}).

One can also postulate a model for a particular form of interaction between the two loci. We consider a model which we call the endophenotype-to-disease model where an allele at locus 1 increases susceptibility to the endophenotype impairment $Y_1 = 1$ and possibly to the disease $Y_2 = 1$, and an allele at locus 2 increases susceptibility to the disease in carriers of gene 1 susceptibility genotypes (at higher risk of the endophenotype impairment). For that model we express allele counts as proportion of a given allele in a genotype, taking values 0, $\frac{1}{2}$ and 1. The model is then written as:

$$\begin{aligned} &\log\left(\frac{P[Y_1 = 1, Y_2 = 0|X_1, X_2]}{P[Y_1 = 0, Y_2 = 0|X_1, X_2]}\right) \qquad (12)\\ &\quad = \beta_{10} + \beta_{11}X_1 + \beta_e X_1(1 - X_2)\\ &\log\left(\frac{P[Y_1 = 0, Y_2 = 1|X_1, X_2]}{P[Y_1 = 0, Y_2 = 0|X_1, X_2]}\right)\\ &\quad = \beta_{20} + \beta_{21}X_1 \end{aligned}$$

$$\log\left(\frac{P[Y_1 = 1, Y_2 = 1|X_1, X_2]}{P[Y_1 = 0, Y_2 = 0|X_1, X_2]}\right) = \beta_{30} + \beta_{31}X_1 + \beta_{33}X_1X_2$$

We keep the same notation for the coefficients as in the full model, except for the coefficient β_e, which represents the effect on the risk of the endophenotype impairment in non-carriers of the locus 2 tested allele. When the endophenotype-to-disease model holds, the coefficients β_{33} and β_e are of the same sign. The marginal association of X_2 to the endophenotype impairment under that model will typically be small. Its direction and magnitude depend on the values of β_{33} and β_e and the distribution of X_1.

The null hypothesis of the conditional test of locus 2 given locus 1 under the above model is formulated as:

$$\beta_e = \beta_{33} = 0 \tag{13}$$

The alternative hypothesis can be restricted to

$$\beta_e > 0, \beta_{33} > 0 \cup \beta_e < 0, \beta_{33} < 0$$

or a general alternative can be considered, but the alternative space then contains models outside of the conceptual model formulated above.

Alternatively, detection of locus 2 can be attempted by testing a single interaction parameter between X_1 and X_2, as in the context of a genetic analysis conditional on an environmental exposure (Kraft et al., 2007). Here the interaction parameter for the logistic function contrasting the disease and endophenotype impairment category to the reference category β_{33} is the most promising to test to detect effects on the disease and endophenotype impairment jointly.

2.2. SOFTWARE IMPLEMENTATION

We have implemented the extension of the GDT to polytomous phenotypes and two loci in the R package `fat2Lpoly`, standing for Family-based Association Test for 2 Loci and Polytomous phenotypes available on the CRAN archive at CRAN.R-project.org/package=fat2Lpoly. A function is provided to read phenotype and genotype data, variable names and IBD sharing proportions (if applicable) from input files in the Merlin/QTDT format (www.sph.umich.edu/csg/abecasis/Merlin/tour/input_files.html) and convert them into R objects. Alternatively, R objects made by the user in the same format can be provided as input. Functions are provided to setup design matrices for the full two-locus polytomous model, the one-locus polytomous model and the disease-to-endophenotype model. User-defined functions setting-up customized design matrices can be provided instead of these pre-defined functions.

2.3. EVALUATION BY SIMULATION OF THE PROPOSED HYPOTHESIS TESTS UNDER TWO-LOCUS MODELS

The family structure used in the simulations is a 3-generation 16-member family depicted in **Figure 1**. The disease and endophenotype status of all family members was assumed to be observed. We generated genotype data for genetic variants with two alleles such as single nucleotide polymorphisms (SNPs) at two independent loci. The genotypes of pedigree founders were sampled under Hardy-Weinberg equilibrium using risk allele frequencies (RAFs) of 0.1 at locus 1 and 0.3 at locus 2. The transmission of alleles to their descendants was then simulated following the rules of Mendelian inheritance. Two dichotomous phenotypes Y_1 and Y_2 were generated in a two-step approach: we first simulated from the distribution of Y_{i1} for each subject i by summing over Y_{i2} in a polytomous model, then from the distribution of the vector $Y_2|Y_1$. In the model to simulate $Y_{i2}|Y_1$, Y_1 is treated as a vector of fixed effect, with the effect of the endophenotype of subject h, Y_{h1}, modulated by the kinship coefficient ϕ_{ih} between i and h. An additive polygenic effect on the logit of Y_2 was also included. The model can be written:

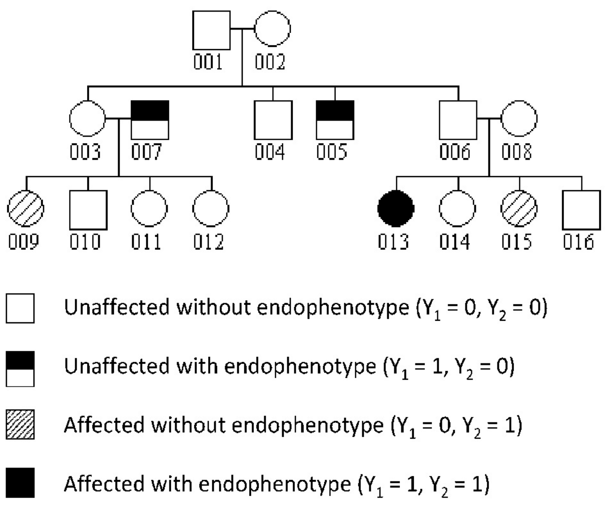
001
002
003
007
004
005
006
008
009
010
011
012
013
014
015
016
□ Unaffected without endophenotype (Y_1 = 0, Y_2 = 0)
Unaffected with endophenotype (Y_1 = 1, Y_2 = 0)
▨ Affected without endophenotype (Y_1 = 0, Y_2 = 1)
■ Affected with endophenotype (Y_1 = 1, Y_2 = 1)

FIGURE 1 | Structure of simulated families with an example of phenotype realization.

$$\log\left(\frac{P[Y_{i2} = 1|Y_1, X, U]}{P[Y_{i2} = 0|Y_1, X, U]}\right) = \gamma'(X_i, Y_{i1}) + U_i + \alpha\sum_{h \neq i}^{n}(Y_{h1} - \nu)\phi_{ih} \tag{14}$$

$$U \sim N(0, \sigma^2\Phi) \tag{15}$$

where $\gamma'(X_i, Y_{i1})$ in an abbreviated expression of the model for the disease phenotype given the genotype at major loci and endophenotype status of subject i derived from a polytomous model and Φ is the kinship matrix between the family members. The parameter σ^2 controls the degree of polygenic dependence between the disease status Y_2 of the family members and the parameter α the degree of genetic dependance of Y_2 on Y_1 not captured by the genotype at the loci in the model. The parameter ν, between 0 and 1, determines the relative importance of the risk increase $1 - \nu$ due to observing an endophenotype impairment and the risk decrease $-\nu$ due to observing the normal level of the endophenotype in a relative. We note this simulation scheme is meant to reproduce the association between disease phenotype and endophenotype status of relatives, not to represent a causal mechanism. Among the simulated families, we kept those with at least a cousin pair with $Y_2 = 1$, i.e., affected by the disease to mimic the ascertainment process of families in a genetic study.

We simulated two scenarios of population origin of the sample: (1) homogeneity: the sample came from a single population where the phenotypes were generated under the polytomous model presented in **Table 1**. Under this models and with the above RAFs, the disease had a population prevalence of 0.0076 and the endophenotype impairment a prevalence of 0.128; (2) heterogeneity: the sample was a mixture of families from two populations, both represented in equal proportions. In population 1, all intercept coefficients in **Table 1** were reduced by 0.5, while in population 2 they were increased by 0.5. This resulted in disease prevalences of 0.005 in population 1 and 0.012 in population 2, and endophenotype impairment prevalences of 0.082 in population 1 and 0.194 in population 2.

To verify the Type I error of tests of association to locus 2 under the null hypothesis of no association to locus 2, but in presence of genetic linkage at that locus, we generated an additional biallelic variant at locus 2 independent from the causal variant at that locus, i.e., in linkage equilibrium with it. In the homogeneous population, the minor allele frequency of that marker was equal to the RAF of the causal variant, but in the mixture of two populations the minor allele frequency was 0.1 in population 1 and 0.5 in population 2, creating population structure at that locus. For the power evaluation, we tested association to the actual causal variant at locus 2.

The tests evaluated include the tests of the null hypotheses 11 which we denote "cpoly," 13 which we denote "(β_e, β_{33})," and $\beta_{33} = 0$. We also evaluated a single locus polytomous model (model 11 with X_2 only). The coefficients in that model are labeled $\underset{\sim}{\beta}(1L)$, and we tested the null hypotheses $\underset{\sim}{\beta}(1L) = 0$ as well as $\beta_3(1L) = 0$. For the evaluation of the Type I error, Wald tests of the coefficients of the one locus model based on GEEs were also performed. However, these tests were not used for the power comparison, since they had inflated Type I error under our heterogeneity scenario where population stratification was present.

Table 1 | Regression coefficients of the example polytomous model.

Coef.	Value	Coef.	Value	Coef.	Value	Coef.	Value
β_{10}	−2	β_{11}	log (2)	β_{12}	0	β_{13}	− log (2)
β_{20}	−5.5	β_{21}	0	β_{22}	0	β_{23}	0
β_{30}	−5.5	β_{31}	0	β_{32}	0	β_{33}	log (16)

In presence of population stratification, previously available valid tests are restricted to a dichotomous outcome and a single marker. Analysis options are then limited to testing association of a single marker to the dichotomous endophenotype Y_1 and disease status Y_2, either in the full sample or, in the case of Y_2, in a stratum defined by Y_1. This is akin to the strategy for detecting modifier genes conferring susceptibility to a specific phenotype (i.e., the disease) consisting in testing association to the specific phenotype among subjects with a broader phenotype (i.e., the endophenotype impairment) (Bureau et al., 2012). We therefore compared the power of various tests derived under our extension of the GDT against the single marker GDT for dichotomous outcomes applied to the locus 2 causal variant with three phenotype definitions: (1) the disease status Y_2 (standard analysis noted simply GDT), (2) the disease status Y_2 in the subset of subjects with $Y_1 = 1$ (endophenotype impairment), setting the phenotype of other subjects to unknown (GDTc), and (3) the endophenotype status Y_1 (GDTe). We also compared our tests to score tests of coefficients of the usual two-locus logistic model for a dichotomous trait:

$$\log\left(\frac{P[Y=1|X_1,X_2]}{P[Y=0|X_1,X_2]}\right) = \eta_0 + \eta_1 X_1 + \eta_2 X_2 + \eta_3 X_1 X_2 \quad (16)$$

The 2 d.f. test of the null hypothesis $\eta_2 = \eta_3 = 0$ is denoted "cdisease" when the phenotype tested is Y_2 and "cendo" when the phenotype tested is Y_1.

3. RESULTS

3.1. EVALUATION OF THE TYPE I ERROR

The Type I error was evaluated on 1000 replicate samples of 100 families. The results of the simulation under the null hypothesis in **Table 2** show that the nominal Type I error rate was respected

Table 2 | Estimations of Type I error on 1000 replicate samples of 100 families.

	GEE		Conditional likelihood				
	Single locus		Single locus		Given other locus[a]		
	$\beta_3(1L)$	$\underset{\sim}{\beta}(1L)$	$\beta_3(1L)$	$\underset{\sim}{\beta}(1L)$	β_{33}	(β_e, β_{33})	cpoly
HOMOGENEOUS POPULATION							
$\alpha = 0.01$:	0.009	0.015	0.006	0.012	0.001	0.001	0.007
$\alpha = 0.05$:	0.053	0.060	0.051	0.045	0.019	0.003	0.029
MIXTURE OF TWO POPULATIONS							
$\alpha = 0.01$:	0.102	0.762	0.012	0.010	0.002	0.001	0.007
$\alpha = 0.05$:	0.237	0.906	0.048	0.053	0.025	0.003	0.033

[a] *subject pairs were weighted using expression 6.*

under both scenarios for all test statistics from our polytomous extension of the GDT. The Type I error rates of the tests conditional on locus 1 were similar for weight definitions 6 and 7, so only results for the former are shown. They were both below the nominal level, making these tests conservative. By contrast, the Type I error of the Wald tests based on GEE estimates were at nominal level only under the homogeneous sample scenario, and were severely inflated under the heterogeneous sample scenario.

3.2. EVALUATION OF THE POWER

Under the simulated scenario the endophenotype-to-disease model holds. While the test of the null hypothesis 13 has some power, testing $\beta_{33} = 0$ (the interaction parameter for the combination of disease and endophenotype impairment) achieves the highest power among the tests considered (**Figure 2**). Using weight definition 7 instead of 6 led to nearly identical power (results not shown). Under this scenario, testing association for the same phenotypic category of the allele count at locus 2 $\beta_3(1L) = 0$ or the entire vector $\underset{\sim}{\beta}(1L) = 0$ does not provide a measurable power improvement over the GDT applied to the disease status in the subset of subjects with endophenotype impairment. Further comparisons of testing strategies under a variety of scenarios will be reported elsewhere.

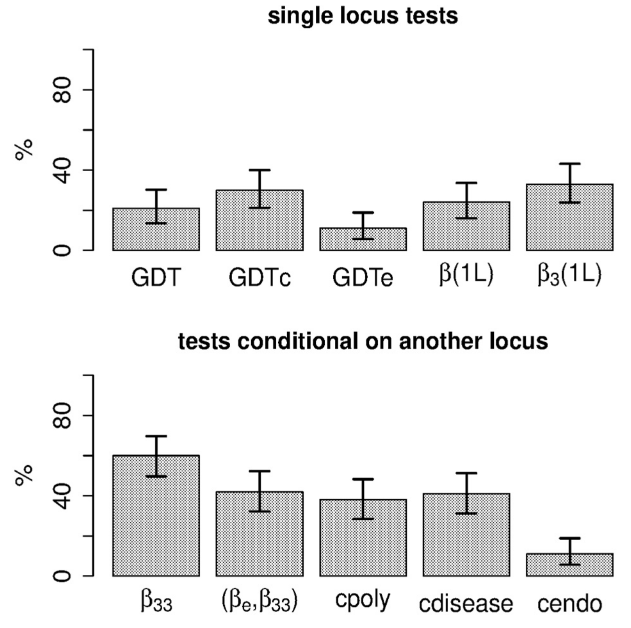

FIGURE 2 | Power of various within-family score tests to detect locus 2. See text for definitions of the acronyms of the tests. For tests conditional on another locus, subject pairs were weighted using expression 6.

3.3. APPLICATION TO MAJOR PSYCHOSIS AND VISUAL EPISODIC MEMORY

Schizophrenia (SZ) and bipolar disorder (BP) are two forms of the spectrum of major psychosis (MP), which also includes schizo-affective disorder. SZ and BP co-aggregate in families (Van Snellenberg and de Candia, 2009), and share genetic liability (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). Various cognitive domains are widely recognized as endophenotypes of MP (Bora et al., 2009; Ivleva et al., 2010). In the Eastern Quebec kindred study, visual episodic memory (VisEM) was found to be impaired in both SZ and BP patients and non-affected adult relatives of these patients (Maziade et al., 2011). In that same family sample, we recently replicated an association between the T allele of SNP rs1156026 and SZ that we had previously detected in another sample (Bureau et al., 2013). All the elements required for the application of our extension of the GDT to markers genotyped in the family sample are present: a diagnosis within the spectrum of MP as the disease phenotype, a VisEM mesurement dichotomized as presence/absence of deficit as the endophenotype and the SNP rs1156026 as the established risk locus. Given the small number of subjects with cognitive measurements, this analysis is not sufficiently powered to draw conclusions and must be considered illustrative. The small sample size also limited us to an analysis of MP globally, without separating SZ and BP.

VisEM was measured by the performance on the delayed recall of the Rey figure task (Meyers and Meyers, 1995) defining the affected status as being the 4th percentile of the distribution of age and gender matched controls. We retained the 14 informative families defined as containing at least one MP affected subject with a visual memory measurement and subjects in at least one other phenotypic category. **Table 3** presents the joint distribution of MP and VisEM in the 133 genotyped subjects from these families along with the frequency of the rs1156026 T allele. Although the frequency of the T allele is greatly increased in subjects with MP and the VisEM impairment compared to normal subjects (and this increase is statistically significant in a population-level comparison) the within-family score test of the corresponding coefficient has a high p-value, suggesting that the difference in T allele frequency is mostly between families and not so much within families.

We tested association to 80 SNPs in genomic regions where genetic linkage to SZ, BP, or MP was previously detected in that family sample on the p arm of chromosomes 6, 8, and 16 and the q arm of chromosomes 12 and 18 (Maziade et al., 2005). We applied the same tests as in the simulation study. SNPs where a p-value < 0.05 was obtained in at least one analysis are shown in **Table 4**.

The results for rs7500550 illustrate that tests of the joint MP-VisEM phenotype conditional on the rs1156026 T allele count can detect associations to SNPs where the test of the MP or VisEM phenotype alone did not. In this case, the rare allele was negatively associated to MP with VisEM impairment with Z statistics of -2.66 for the X_2 and -2.34 for the X_1X_2 terms ($p = 0.0019$ for the test of the coefficients of both terms) while it was positively associated to a lesser extent to MP without VisEM impairment with Z statistics of 2.54 for the X_2 and 2.07 for the X_1X_2 terms ($p = 0.005$ for the test of the coefficients of both terms). The signal was thus driven by opposite associations to these two phenotypic categories. The signal at rs1087266 was detected by single locus tests with lower p-values than by tests conditioning on rs1156026. In that case, testing association with VisEM status was the key to detect the signal. Nonetheless, the conditional test of the

Table 3 | Joint distribution of major psychosis and visual episodic memory deficits along with the frequency of the rs1156026 T allele.

		VisEM <= 4th perc				VisEM > 4th perc				Total	
		n_1	Freq T	p_{GEE}[a]	p_{1L}[b]	n_0	Freq T	p_{GEE}[a]	p_{1L}[b]	$n_.$	Freq T
MP	Yes	21 (41%)	0.52	0.0011	0.34	30	0.40	0.040	0.310	51 (38%)	0.45
	No	13 (16%)	0.31	0.97	0.72	69	0.30			82 (62%)	0.30
	Total	34 (26%)	0.44			99	0.33			133	0.36

[a] *p-values of Wald tests of the coefficients of the one locus polytomous model estimated using generalized estimating equations (GEE).*

[b] *p-values of within-family score tests of the coefficients of the one locus polytomous model.*

Table 4 | Results for SNPs where a *p*-value < 0.05 was obtained in at least one analysis[a].

SNP	Chr	Pos (Mb)		MAF (*n*)			
				$Y_1 = 0, Y_2 = 0$	$Y_1 = 0, Y_2 = 1$	$Y_1 = 1, Y_2 = 0$	$Y_1 = 1, Y_2 = 1$
rs1087266	6	24.4		0.39 (42)	0.26 (25)	0.60 (5)	0.55 (19)
rs7500550	16	19.1		0.11 (41)	0.16 (25)	0.17 (6)	0.03 (18)
TESTS *p*-VALUES							
SNP	**GDT**	**GDTc**	**GDTe**	$\underset{\sim}{\beta}(1L)$	β_{33}	(β_e, β_{33})	**cpoly**
rs1087266	0.48	0.25	0.005	0.005	0.032	0.085	0.006
rs7500550	0.52	0.17	0.57	0.040	0.019	0.064	0.015

[a] *For tests conditioning on rs1156026 genotypes, subject pairs were weighted using expression 6.*

polytomous phenotype provides a *p*-value similar to the standard GDT. Given the limited power of the analysis and the number of SNPs tested, these results cannot be considered statistically significant once multiple testing is taken into account.

4. DISCUSSION

We have extended the GDT, a score test of genetic association applicable with extended families, to enable testing association with a polytomous phenotype. Another extension is the use of a model of association with two genetic loci, allowing to test association at a locus conditional on the genotype of a marker at a known risk locus, to exploit interaction between the two. A software implementation in the form of a R package has been made freely available. The within-family analysis framework that we adopted has the advantage of protecting against Type I error inflation due to population stratification. Polytomous phenotypes can be more informative than dichotomous ones to detect genetic associations, as illustrated in our simulation study.

The proposed score tests also suffer from limitations. First, score tests provide no estimates of the regression parameters being tested. Conditional maximum likelihood estimation would be applicable only with exchangeable relatives, which is not required for the GDT as explained in Supplementary Material. We are exploring the robustness and power of conditional maximum likelihood estimation in sibships from extended families.

Second, within-family analysis tends to be less powerful than population-level analysis which also exploits between family information. Furthermore, the Type I error of score tests for one locus conditionning on another tends to be conservative even with the weight definition 6 neglecting variability from estimating the $\underset{\sim}{\alpha}$. Our simulation studies illustrate that power remains limited despite large sample sizes (1600 subjects in 100 families) and large effect sizes (interaction odds ratios of 16). Extracting the most power from the data is particularly important when phenotypic measures are expensive to obtain, such as the cognitive measurements in our example. Population analyses are then attractive, with an adjustment for population structure using genomewide SNP genotypes (Price et al., 2006). Methods for population analysis of polytomous phenotypes are not well developed, and will be the object of future work.

AUTHOR CONTRIBUTIONS

Alexandre Bureau defined the research questions, derived the proposed statistical test, wrote part of the R implementation, conceived the simulation study, oversaw the analysis of the major psychosis data and drafted the manuscript. Jordie Croteau wrote part of the R implementation, performed the simulation study and the analysis of the major psychosis data, and created figures and tables. Yvon C. Chagnon oversaw the genotyping of the Eastern Quebec kindred study and contributed to the design of the genetic aspects of that study. Marc-André Roy contributed to the design of the genetic and clinical aspects of the Eastern Quebec kindred study, established diagnosis of patients and made substantial revisions to the manuscript. Michel Maziade designed the genetic and clinical aspects of the Eastern Quebec kindred study, oversaw clinical data collection and established diagnosis of patients. All authors approved the version submitted for publication and agree to be accountable for all aspects of the work.

ACKNOWLEDGMENTS

This work was funded by the Canadian Institutes of Health research (CIHR, grant MOP-106448) and by a research fellowship from the Fonds de recherche du Québec - Santé to A. Bureau. The Eastern Quebec Kindred Study is funded by CIHR (grants MOP-74430, MOP-119408 and MOP-114988) and by a Canada Research Chair (950-200810) in the genetics of neuropsychiatric disorders of which M. Maziade is the Chair.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00258/abstract

REFERENCES

Bora, E., Yucel, M., and Pantelis, C. (2009). Cognitive endophenotypes of bipolar disorder: a meta-analysis of neuropsychological deficits in euthymic patients and their first-degree relatives. *J. Affect. Disord.* 113, 1–20. doi: 10.1016/j.jad.2008.06.009

Bureau, A., Chagnon, Y. C., Croteau, J., Fournier, A., Roy, M. A., Paccalet, T., et al. (2013). Follow-up of a major psychosis linkage site in 13q13-q14 reveals significant association in both case-control and family samples. *Biol. Psychiatry* 74, 444–450. doi: 10.1016/j.biopsych.2013.03.004

Bureau, A., Croteau, J., Merette, C., Fournier, A., Chagnon, Y. C., Roy, M. A., et al. (2012). Detection of phenotype modifier genes using two-locus linkage analysis in complex disorders such as major psychosis. *Hum. Hered.* 73, 195–207. doi: 10.1159/000341392

Bureau, A., Merette, C., Croteau, J., Fournier, A., Chagnon, Y. C., Roy, M. A., et al. (2009). A new strategy for linkage analysis under epistasis taking into account genetic heterogeneity. *Hum. Hered.* 68, 231–242. doi: 10.1159/000228921

Chen, W. M., Manichaikul, A., and Rich, S. S. (2009). A generalized family-based association test for dichotomous traits. *Am. J. Hum. Genet.* 85, 364–376. doi: 10.1016/j.ajhg.2009.08.003

Cross-Disorder Group of the Psychiatric Genomics Consortium, A. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide snps. *Nat. Genet.* 45, 984–994. doi: 10.1038/ng.2711

Gottesman, I., and Gould, T. D. (2003). The endophenotype concept in psychiatry: etymology and strategic intentions. *Am. J. Psychiatry* 160, 636–645. doi: 10.1176/appi.ajp.160.4.636

Guey, L. T., Garcia-Closas, M., Murta-Nascimento, C., Lloreta, J., Palencia, L., Kogevinas, M., et al. (2010). Genetic susceptibility to distinct bladder cancer subphenotypes. *Eur. Urol.* 57, 283–292. doi: 10.1016/j.eururo.2009.08.001

Ivleva, E. I., Morris, D. W., Moates, A. F., Suppes, T., Thaker, G. K., and Tamminga, C. A. (2010). Genetics and intermediate phenotypes of the schizophrenia–bipolar disorder boundary. *Neurosci. Biobehav. Rev.* 34, 897–921. doi: 10.1016/j.neubiorev.2009.11.022

Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J., and Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Hum. Hered.* 63, 111–119. doi: 10.1159/000099183

Liang, K.-Y., and Stewart, W. F. (1987). Polychotomous logistic regression methods for matched case-control studies with multiple case or conrol groups. *AJE* 125, 720–730.

Maziade, M., Rouleau, N., Merette, C., Cellard, C., Battaglia, M., Marino, C., et al. (2011). Verbal and visual memory impairments among young offspring and healthy adult relatives of patients with schizophrenia and bipolar disorder: selective generational patterns indicate different developmental trajectories. *Schizophr. Bull.* 37, 1218–1228. doi: 10.1093/schbul/sbq026

Maziade, M., Roy, M. A., Chagnon, Y. C., Cliche, D., Fournier, J. P., Montgrain, N., et al. (2005). Shared and specific susceptibility loci for schizophrenia and bipolar disorder: a dense genome scan in eastern quebec families. *Mol. Psychiatry* 10, 486–499. doi: 10.1038/sj.mp.4001594

Meyers, J., and Meyers, K. (1995). *Rey Complex Figure Test and Recognition Trial (RCFT)*. Odessa, FL: Psychological Assessment Resources.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Van Snellenberg, J. X., and de Candia, T. (2009). Meta-analytic evidence for familial coaggregation of schizophrenia and bipolar disorder. *Arch. Gen. Psychiatry* 66, 748–755. doi: 10.1001/archgenpsychiatry.2009.64

Ziegler, A., and König, I. R. (2010). *A statistical Approach to Genetic Epidemiology, 2nd Edn.* Weinheim: Wiley-VCH. doi: 10.1002/9783527633654

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 03 April 2014; accepted: 14 July 2014; published online: 08 August 2014.

*Citation: Bureau A, Croteau J, Chagnon YC, Roy M-A and Maziade M (2014) Extension of the generalized disequilibrium test to polytomous phenotypes and two-locus models. Front. Genet. **5**:258. doi: 10.3389/fgene.2014.00258*

This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics.