A mixed-model approach for powerful testing of genetic associations with cancer risk incorporating tumor characteristics

Haoyu Zhang,[1,2] Ni Zhao,[1] Thomas U. Ahearn,[2] William Wheeler,[3] Montserrat García-Closas,[2] and Nilanjan Chatterjee[1,4]

[1]*Department of Biostatistics Johns Hopkins Bloomberg SPH, Baltimore, MD 21205, U.S.A.*

[2]*National Cancer Institute, Division of Cancer Epidemiology and Genetics, Rockville, MD 20850, U.S.A.*

[3]*Information Management Services, Inc., Rockville, MD 20850, USA*

[4]*Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, U.S.A.*

(Dated: 25 September 2019)

Abstract: Cancers are routinely classified into subtypes according to various features, including histo-pathological characteristics and molecular markers. Previous genome-wide association studies have reported heterogeneous association between loci and cancer subtypes. However, it is not evident what is the optimal modeling strategy for handling correlated tumor features, missing data, and increased degrees-of-freedom in the underlying tests of associations. We propose score tests for genetic associations using a mixed-effect two-stage polytomous model (MTOP). In the first stage, a standard polytomous model is used to specify for all subtypes defined by the cross-classification of different markers. In the second stage, the subtype-specific case-control odds ratios are specified using a more parsimonious model based on the case-control odds ratio for a baseline subtype, and the case-case parameters associated with tumor markers. Further, to reduce the degrees-of-freedom, we specify case-case parameters for additional exploratory markers using a random-effect model. We use the Expectation-Maximization (EM) algorithm to account for missing data on tumor markers. Through simulations across a range of realistic scenarios and data from the Polish Breast Cancer Study (PBCS), we show MTOP outperforms alternative methods for identifying heterogeneous associations between risk loci and tumor subtypes.

Key words: Cancer subtypes; EM algorithm; Etiologic heterogeneity; Susceptibility variants; Score tests; Two-stage polytomous model.

I. INTRODUCTION

Genome-wide association studies (GWAS) have identified hundreds of single nucleotide polymorphisms (SNPs) associated with various cancers (MacArthur *et al.*, 2016; Visscher *et al.*, 2017). However, many cancer GWAS have often defined cancer endpoints according to specific anatomic sites, and not according to subtypes of the disease. Many cancers consist of etiologically and clinically heterogeneous subtypes that are defined by multiple correlated tumor characteristics, for instance, breast cancer is routinely classified into subtypes defined by tumor expression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) (Curigliano *et al.*, 2017; Perou *et al.*, 2000; Prat *et al.*, 2015).

Increasing number of epidemiologic studies with tumor specimens are allowing the characteristics of cancers at the histological and molecular levels (Cancer Genome Atlas, 2012; Cancer Genome Atlas Research, 2012, 2014). This provides tremendous opportunities to characterize distinct etiological pathways within cancer subtypes. For example, a breast cancer ER-negative specific GWAS reported 20 SNPs that are more strongly associated with risk of developing ER-negative than ER-positive disease (Milne *et al.*, 2017). Previous studies also suggest traditional breast cancer risk factors, such as age, obesity, and hormone therapy use, may also be heterogeneously associated with breast cancer subtypes (Barnard *et al.*, 2015).

Standard procedure to test associations between a risk factor and cancer subtypes are fitting standard logistic regression for each subtype versus control group, then accounting

for the multiple testing. However, there are several limitations for this procedure. First, its common for cancer cases having missing tumor marker data. This will lead to many cancer cases with no subtype definition. Standard approach usually drops these cases without modeling them. Second, the tumor markers are highly correlated with each other. Testing each subtype separately without modelling the correlation limits the power. Finally, as the number of tumor markers increase, the number of cancer subtypes would dramatically increase. The increased degrees of freedoms would penalize the power of the model.

A two-stage polytomous logistic regression was previously proposed to characterize subtype heterogeneity of a disease using the underlying disease characteristics (Chatterjee, 2004). The first stage of this method uses a polytomous logistic regression (Dubin and Pasternack, 1986; Gortmaker *et al.*, 1994) to model each subtype specific case-control odds ratio. In the second stage, the subtype-specific case-control odds ratios are decomposed to the case-control odds ratio of a reference subtype, case-case odds ratio of tumor characteristic and higher order interactions between the case-case odds ratio of the tumor characteristics. By constraining some higher order interaction terms to be 0, the two-stage model can reduce the degrees of freedom due to the estimation of subtype specific odds ratio. Moreover, the second stage parameters can be interpreted as the case-case odds ratio for tumor characteristics, which measure the etiological heterogeneity between the associations of cancer subtypes.

Although, the two-stage model could improve the power compared to fitting standard logistic regression for each subtype (Chatterjee, 2004; Zabor and Begg, 2017), for several reasons the method has not been widely applied to analyze data on multiple tumor charac-

teristics. First, the two-stage model could not handle missing tumor characteristics problems that are common in epidemiologic studies. Second, the two-stage model estimation algorithm places high demands on computing power and is therefore not readily applicable to large datasets. Finally, although two-stage model could reduce the multiple testing burdens compared to traditional methods, as the number of tumor characteristics increases, the method could still have substantial power loss due to the degrees of freedom penalty.

In this paper, we propose a series of computational and statistical innovations to perform computationally scalable and statistically efficient association tests in large GWAS with cancer data incorporating tumor characteristics. Within the two-stage modeling framework, we propose three alternative types of hypotheses for testing genetic associations in the presence of tumor heterogeneity. As the degrees-of-freedom for the tests can be large in the presence of large number of tumor characteristics, we propose modelling parameters associated with exploratory tumor characteristics using a random-effect model. We then derive score-tests under the resulting mixed effect model while taking into account missing data on tumor characteristics using an efficient EM algorithm (Dempster *et al.*, 1977). All of the different steps combined, our work represent conceptually distinct and practically important extensions of earlier methods for genetic association testing based on random-/fixed-effect models (Lin, 1997; Sun *et al.*, 2013; Wu *et al.*, 2011; Zhang and Lin, 2003) to the novel setting of modeling genetic association with multiple disease characteristics.

We organize the remainder of the paper as follow. In Section II, we describe the two-stage model and methodology underlying the score-tests. In Section III, we study the type one error and power of the proposed methods through extensive simulations across a wide range

of realistic scenarios. The proposed methods are illustrated with two applications using Polish Breast Cancer Study (PBCS) data in Section IV. Finally we discuss the strengths and limitations of the methods, and future research directions in Section V.

II. METHOD

A. Two-stage polytomous model

The details of two-stage polytomous logistic model are described in (Chatterjee, 2004) and we will briefly describe it here. Suppose a disease can be classified using K disease characteristics. Assuming each characteristic k can be classified into M_k categories, then the disease can be classified into $M \equiv M_1 \times M_2 \cdots \times M_K$ subtypes. For example, breast cancer can be classified into eight subtypes by three tumor characteristics (ER, PR, and HER2), each of which is either positive or negative. Note, that we will use this breast cancer example to demonstrate the methods throughout the methods section. Let D_i denote the disease status, taking values in $\{0, 1, 2, \cdots, M\}$, of the ith ($i = 1, \cdots, N$) subject in the study. $D_i = 0$ represents a control, and $D_i = m$ represent a subject with disease of subtype m. Let G_i be the genotype for ith subject and $\mathbf{X}_i$ be a $P \times 1$ vector of other covariates we want to adjust for in the model, where P is the total number of other covariates. we first build the standard "saturated" polytomous logistic regression model as the first stage model,

$$Pr(D_i = m|G_i, \mathbf{X}_i) = \frac{\exp(\beta_m G_i + \mathbf{X}_i^T \boldsymbol{\eta}_m)}{1 + \sum_{m=1}^{M} \exp(\beta_m G_i + \mathbf{X}_i^T \boldsymbol{\eta}_m)}, \quad m \in \{1, 2, \cdots, M\}, \tag{1}$$

where β_m and $\boldsymbol{\eta}_m$ are the regression coefficients for a SNP and other covariates associated with the mth subtype, repsectively.

Because each cancer subtype m is defined through a unique combination of the K characteristics, we can always alternatively index the parameters β_m as $\{\beta_{s_1 s_2 \cdots s_K}\}$, where $s_k \in \{0, 1\}$ for binary tumor characteristics, and $s_k \in \{t_1 \leq t_2 \leq \cdots \leq t_{M_k}\}$ for ordinal tumor characteristics with $t_1, \ldots, t_{M_k}$ as a set of ordinal scores for M_k different levels. With this new index, we can represent the log odds ratio defined in first stage model as

$$\beta_{s_1 s_2 \ldots s_K} = \theta^{(0)} + \sum_{k_1=1}^{K} \theta_{k_1}^{(1)} s_{k1} + \sum_{k_1=1}^{K} \sum_{k_2>k_1}^{K} \theta_{k_1 k_2}^{(2)} (s_{k_1} s_{k_2}) + \ldots + \theta_{12\ldots K}^{(K)} (s_1 s_2 \ldots s_K). \tag{2}$$

Here $\theta^{(0)}$ represents the case-control log odds ratio for a reference disease subtype compared to the control, $\theta_{k_1}^{(1)}$ represent the main effect of k_1th tumor characteristic, $\theta_{k_1 k_2}^{(2)}$ represent the second order interaction between k_1th and k_2th tumor characteristics and so on. A reference level could be defined for each tumor characteristic, and the reference disease subtype is jointed defined by the combination of the K tumor characteristics.

The reparameterization in 2 provides a way to decompose the first stage parameters to a lower dimension. We can constrain different main effects or interaction effects to be 0 to specify different second stage models. We can link the first stage parameters and second stage parameters with a matrix form, $\boldsymbol{\beta} = \mathbf{Z}_G \boldsymbol{\theta} = \mathbf{Z}_G \begin{bmatrix} \theta^{(0)} & \boldsymbol{\theta}_{\mathrm{H}}^T \end{bmatrix}^T$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_M)^T$ is a vector of first stage case-control log odds ratios for all the M subtypes, $\theta^{(0)}$ is the case-control log odds ratio for the reference subtype, $\boldsymbol{\theta}_{\mathrm{H}}$ is the vector containing all the main effects and interactions effects in the second stage. We will also refer $\boldsymbol{\theta}_{\mathrm{H}}$ as case-case parameters in the later paper. And $\boldsymbol{\theta} = (\theta^{(0)}, \boldsymbol{\theta}_{\mathrm{H}}^T)^T$ is the vector of second stage parameters. $\mathbf{Z}_G$ is the second stage design matrix connecting the first stage parameters and second stage

parameters. Through specifying different second stage main effects or interaction effects to be 0, we can construct different $\mathbf{Z}_G$ to build different two-stage model.

Up to now, we have only described second stage decomposition for the regression coefficients of $\mathbf{G}$. We could also apply second stage decomposition on the other covariates. The details of this could be found in Supplementary Section 1. We suggest not to perform any second stage decomposition on the regression coefficients of intercepts of first stage polytomous model, since making assumption on the prevalence of different cancer subtypes could potentially yield bias. Moving forward, we use $\mathbf{Z_X}$ to denote the second stage design matrix for the other covariates $\mathbf{X}$, $\boldsymbol{\lambda}$ to denote the second stage parameters for $\mathbf{X}$, and $\mathbf{Z}$ to denote the second stage design matrix for all the covariates.

B. Hypothesis test under two-stage model

We can decompose the first stage case-control log odds ratios of all the subtypes into the second stage case-control log odds ratio of the reference subtype, main effects and interaction effects of tumor characteristics. This decomposition presents multiple options for comprehensively testing the association between a SNP and cancer subtypes. The first hypothesis test is the global association test, $\mathrm{H}_0^{\mathrm{A}} : \boldsymbol{\theta} = \begin{bmatrix}\theta^{(0)} & \boldsymbol{\theta}_{\mathrm{H}}^T\end{bmatrix}^T = \begin{bmatrix}0 & \mathbf{0}^T\end{bmatrix}^T$ versus $\mathrm{H}_1^{\mathrm{A}} : \boldsymbol{\theta} \neq \mathbf{0}$. This tests for an overall association between the SNP and the disease. Because $\boldsymbol{\theta} = \mathbf{0}$ implies $\boldsymbol{\beta} = \mathbf{0}$, rejecting this null hypothesis means the SNP is significantly associated with at least one of the subtypes. The null hypothesis may be rejected if the SNP is significantly associated with a similar effect size across all subtypes (i.e. $\boldsymbol{\theta}^{(0)} \neq 0$, $\boldsymbol{\theta}_{\mathrm{H}} = \mathbf{0}$), or if the SNP has heterogeneous effects on different subtypes ($\boldsymbol{\theta}_{\mathrm{H}} \neq \mathbf{0}$).

The second hypothesis test is the global heterogeneity test, $\mathrm{H}_0^{\mathrm{EH}} : \boldsymbol{\theta}_{\mathrm{H}} = \mathbf{0}$ versus $\mathrm{H}_1^{\mathrm{EH}} : \boldsymbol{\theta}_{\mathrm{H}} \neq \mathbf{0}$. The global heterogeneity test evaluates for etiologic heterogeneity with respect to a SNP and all tumor characteristics simultaneously. Rejecting this null hypothesis indicates that the first stage case-control log odds ratios of at least two different subtypes are significantly different from each other.

Notably, the global heterogeneity test does not identify which tumor characteristic(s) is/are driving the heterogeneity between the subtypes. To identify the tumor characteristic(s) responsible for observed heterogeneity, we propose the individual tumor marker heterogeneity test, $\mathrm{H}_0^{\mathrm{IH}} : \theta_{\mathrm{H}(k)} = 0$ versus $\mathrm{H}_1^{\mathrm{IH}} : \theta_{\mathrm{H}(k)} \neq 0$, where $\theta_{\mathrm{H}(k)}$ is one of the case-case parameters of $\boldsymbol{\theta}_{\mathrm{H}}$. The case-case parameters $\theta_{\mathrm{H}(k)}$ provide a measurement of etiological heterogeneity according to a specific tumor characteristic (Begg and Zhang, 1994). Under the breast cancer example, we could directly test $\mathrm{H}_0^{\mathrm{IH}} : \theta_{\mathrm{ER}}^{(1)} = 0$ versus $\mathrm{H}_1^{\mathrm{IH}} : \theta_{\mathrm{ER}}^{(1)} \neq 0$. Rejecting the null hypothesis provides evidence that the case-control log odds ratios of ER+ and ER- subtypes are significantly different.

C. EM algorithm accounting for cases with incomplete tumor characteristics

In previous sections, we assumed all the tumor characteristics are observed for every case in the study. In epidemiological research it is very common that tumor characteristic data is missing across study participants. This problem becomes exacerbated as the number of analyzed tumor characteristics grows. Restricting to cases with complete tumor characteristics can reduce statistical power and potentially introduce selection bias. To solve this problem, we propose to use the EM algorithm (Dempster *et al.*, 1977) to find the maximum likeli-

hood estimate (MLE) of two-stage model and incorporate all available information from the study cases. Let $Y_{im} = I(D_i = m)$ denote whether the ith subject has subtype m and $\mathbf{T}_{io}$ be the observed tumor characteristics status of the ith subject. Given the observed tumor characteristics, the possible subtypes for subject i would be a limited subset of all possible tumor subtypes, which can be denoted as $\mathcal{Y}_{io} = \{Y_{im} : Y_{im}$ that is consistent with $\mathbf{T}_{io}\}$. We assume that $(Y_{i1}, Y_{i2}, \ldots, Y_{iM}, G_i, \mathbf{X}_i)$ are independently and identically distributed (i.i.d.), and that the tumor characteristics are missing at random. Let $\boldsymbol{\delta} = (\boldsymbol{\theta}^T, \boldsymbol{\lambda}^T)^T$, where $\boldsymbol{\delta}$ represent the second stage parameters of both $\mathbf{G}$ and $\mathbf{X}$. Given the notation above, the EM algorithm at the vth iteration would be:

E step:

$$Y_{im}^{\mathrm{E}} = E(Y_{im}|G_i, \mathbf{X}_i, \mathbf{T}_{io}; \boldsymbol{\delta}^{(v)}) = \frac{Pr(Y_{im} = 1|G_i, \mathbf{X}_i; \boldsymbol{\delta}^{(v)}) I(Y_{im} \in \mathcal{Y}_{io})}{\sum_{Y_{im} \in \mathcal{Y}_{io}} Pr(Y_{im} = 1|G_i, \mathbf{X}_i; \boldsymbol{\delta}^{(v)})}, \tag{3}$$

where Y_{im}^{E} is the probability of the ith person to be of the mth subtype given his observed tumor characteristics, genotype and other covariates. $I(Y_{im} \in \mathcal{Y}_{io})$ denote whether the mth subtype for the ith subject belong to the subsets of possible subtypes given the observed tumor characteristics.

M step:

$$\boldsymbol{\delta}^{(v+1)} = \arg\max_{\boldsymbol{\delta}} \sum_{i=1}^{N} \left[(1 - \sum_{m=1}^{M} Y_{im}^{E}) \log Pr(D_i = 0|G_i, \mathbf{X}_i) + \sum_{m=1}^{M} Y_{im}^{E} \log \left\{ Pr(D_i = m|G_i, \mathbf{X}_i) \right\} \right]. \tag{4}$$

The M step could be solved through a weighted least square iteration. Let $\mathbf{Y}_m = (Y_{1m}, \ldots, Y_{Nm})^T$, and $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_M^T)^T$. Let $\mathbf{C} = (\mathbf{G}, \mathbf{X})$ and $\mathbf{C}_M = \mathbf{I}_M \otimes \mathbf{C}$. During the tth iteration of the weighted Least square, let $\mathbf{W} = \mathbf{D} - \mathbf{A}\mathbf{A}^T$, with $\mathbf{D} = \mathrm{diag}(\mathbf{P} - \mathbf{P}_{\mathrm{mis}})$,

$\mathbf{P} = E(\mathbf{Y}|\mathbf{C}; \hat{\delta}^{(t)})$, $\mathbf{P}_{\text{mis}} = E(\mathbf{Y}|\mathbf{C}, \mathbf{T}_o; \hat{\boldsymbol{\delta}}^{(t)})$, and $\mathbf{A} = \mathbf{D}(\mathbf{1}_M \otimes \mathbf{I}_N)$. Further, let $\mathbf{Z}$ be the second stage design matrix for $\mathbf{C}$, $\mathbf{Y}^{*(t)} = \mathbf{W}^{(l)}(\mathbf{Y}^{\text{E}} - \mathbf{P}^{(t)}) + \mathbf{C}_M \mathbf{Z} \boldsymbol{\delta}^{(t)}$, where $\mathbf{P}^{(t)}$ and $\mathbf{W}^{(l)}$ uses the same definition as $\mathbf{P}$ and $\mathbf{W}$ evaluated at the $\boldsymbol{\delta}^{(t)}$. The weighted least square updates would be $\boldsymbol{\delta}^{(t+1)} = (\mathbf{Z}^T \mathbf{C}_M^T \mathbf{W}^{(t)} \mathbf{C}_M \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{C}_M^T \mathbf{Y}^{*(t)}$. As $t \to \infty$, the weighted least square interaction converged to $\hat{\boldsymbol{\delta}}^{(v+1)}$, which could be used for next E step. The MLE of the second stage parameters (denoted as $\hat{\boldsymbol{\delta}}$) can be obtained when the EM algorithm converges. More details of EM algorithm procedure could be found in Supplementary Section 2. Following (Louis, 1982), the observed information matrix $\mathbf{I}$ using EM algorithm would be $\mathbf{I} = \mathbf{Z}^T \mathbf{C}_M^T \mathbf{W} \mathbf{C}_M^T \mathbf{Z}$.

With the MLE of the second stage parameters of G as $\hat{\boldsymbol{\theta}}$, the Wald test statistics can be constructed as $\hat{\boldsymbol{\theta}}^{*T} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\theta}}^* \sim \chi_l^2$ for the global association test, global etiological heterogeneity test, and individual tumor characteristic heterogeneity test using the corresponding second stage parameters $\hat{\boldsymbol{\theta}}^*$ and covariance matrix, where the degrees of freedom l equals the length of $\hat{\boldsymbol{\theta}}^*$.

D. Fixed effect two-stage polytomous model score test (FTOP)

Although the hypothesis tests could be implemented through the Wald test, estimating the model parameters for all SNPs in the genome is time consuming and computationally intensive. In this section, we develop a score test for the global association test assuming the second stage parameters as fixed. The score test only needs to estimate the second stage parameters of $\mathbf{X}$ under the null hypothesis once, which makes it much more computational efficient than the Wald test. The EM algorithm is still implemented under the null hypoth-

esis. Since we didn't perform any second stage decomposition on intercept of the regression coefficients of first stage polytomous model, the correlations between the tumor characteristics are kept unspecified. Most of the imputation power are due to the high correlation between the tumor markers. Under the breast cancer example, the correlation between ER and PR is 0.63, ER and HER2 is -0.16, and PR and HER2 is -0.16. Also, the effects of $\mathbf{X}$ are kept in the EM algorithm. Since a single SNP $\mathbf{G}$ usually has small effects in GWAS study, the EM algorithm under the null hypothesis could still efficiently incorporate the information of the cases with incomplete tumor markers.

Let $\mathbf{G}_M = \mathbf{I}_M \otimes \mathbf{G}$, and $\mathbf{X}_M = \mathbf{I}_M \otimes \mathbf{X}$. Under the null hypothesis, $\mathrm{H}_0 : \boldsymbol{\theta} = \mathbf{0}$, let $\hat{\boldsymbol{\lambda}}$ denote the MLE of $\boldsymbol{\lambda}$ under the null hypothesis. The efficient score of $\boldsymbol{\theta}$ is $U_{\boldsymbol{\theta}}(\hat{\boldsymbol{\lambda}}) = \mathbf{Z}_{\mathbf{G}}^T \mathbf{G}_M^T (\mathbf{Y} - \mathbf{P}_{\mathrm{f}})$, where $\mathbf{P}_{\mathrm{f}} = E_{\boldsymbol{\theta}=\mathbf{0}}(\mathbf{Y}|\mathbf{X}; \hat{\boldsymbol{\lambda}})$. Let $\mathbf{W}_{\mathrm{f}} = \mathbf{D}_{\mathrm{f}} - \mathbf{A}_{\mathrm{f}} \mathbf{A}_{\mathrm{f}}^T$, with $\mathbf{P}_{\mathrm{f}} = E_{\boldsymbol{\theta}=\mathbf{0}}(\mathbf{Y}|\mathbf{X}, \mathbf{T}_o; \hat{\boldsymbol{\lambda}})$, $\mathbf{D}_{\mathrm{f}} = \mathrm{diag}(\mathbf{P}_{\mathrm{f}} - \mathbf{P}_{\mathrm{f,mis}})$ and $\mathbf{A}_{\mathrm{f}} = \mathbf{D}_{\mathrm{f}}(\mathbf{1}_M \otimes \mathbf{I}_N)$. The corresponding efficient information matrix of $U_{\boldsymbol{\theta}}(\hat{\boldsymbol{\lambda}})$ is:

$$\tilde{\mathbf{I}} = \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\lambda}} \mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1} \mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\theta}}, \tag{5}$$

where $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}} = \mathbf{Z}_{\mathbf{G}}^T \mathbf{G}_M^T \mathbf{W}_{\mathrm{f}} \mathbf{G}_M \mathbf{Z}_{\mathbf{G}}$, $\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} = \mathbf{Z}_{\mathbf{X}}^T \mathbf{X}_M^T \mathbf{W}_{\mathrm{f}} \mathbf{X}_M \mathbf{Z}_{\mathbf{X}}$, and $\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\theta}} = \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\lambda}}^T = \mathbf{Z}_{\mathbf{X}}^T \mathbf{X}_M^T \mathbf{W}_{\mathrm{f}} \mathbf{G}_M \mathbf{Z}_{\mathbf{G}}$.

The score test statistics $Q_{\boldsymbol{\theta}}$ for fixed-effect two stage model would be:

$$Q_{\boldsymbol{\theta}} = U_{\boldsymbol{\theta}}(\hat{\boldsymbol{\lambda}})^T \tilde{\mathbf{I}}^{-1} U_{\boldsymbol{\theta}}(\hat{\boldsymbol{\lambda}}) \sim \chi_l^2. \tag{6}$$

The degrees freedoms of FTOP is the same as Wald test and they have similar asymptotic power (Yi and Wang, 2011). In GWAS analysis where we need to perform millions of tests, we propose to first use FTOP for global association test to scan the whole genome and select the potential risk regions. Then for each SNP in risk regions, we could estimate $\hat{\boldsymbol{\theta}}$

and use Wald test to perform global heterogeneity test and individual tumor characteristic heterogeneity test.

E. Mixed effect two-stage polytomous model score test (MTOP)

The two-stage model decreases the degrees of freedom compared to the polytomous logistic regression; however, the power gains in the two-stage model can be lost as additional tumor characteristics are added into the model. We further propose a mixed-effect two-stage model by modeling some of the second stage case-case parameters as a random effect. Let $\mathbf{u} = (u_1, \ldots, u_s)^T$, where each u_j follows an arbitrary distribution F with mean zero and variance σ^2. The mixed effect second stage model links the first and second stage parameters via the following:

$$\boldsymbol{\beta} = \mathbf{Z}_{\mathrm{f}}\boldsymbol{\theta}_{\mathrm{f}} + \mathbf{Z}_{\mathrm{r}}\mathbf{u}, \tag{7}$$

where $\mathbf{Z}_{\mathrm{f}}$ is the second stage design matrix of fixed effect, $\mathbf{Z}_{\mathrm{r}}$ is the second stage design matrix of random effect, and $\boldsymbol{\theta}_{\mathrm{f}}$ are the fixed-effect second stage parameters. Let $\boldsymbol{\theta}_{\mathrm{f}} = (\theta^{(0)}, \boldsymbol{\theta}_{\mathrm{fH}}^T)^T$, where $\theta^{(0)}$ is the case-control log odds ratio of the reference subtype and $\boldsymbol{\theta}_{\mathrm{fH}}$ are the fixed case-case parameters. The baseline effect $\theta^{(0)}$ is always kept fixed, since the baseline effect parameter captures the SNP's overall effect on all the cancer subtypes.

The fixed case-case parameters $\boldsymbol{\theta}_{\mathrm{fH}}$ can be used for the tumor characters with prior information suggesting that they are a source of heterogeneity. And the random effect case-case parameters $\mathbf{u}$ can be used for tumor characteristics with little or no prior information to suggest that they are a source of heterogeneity. Under the breast cancer example, the baseline parameter $\theta^{(0)}$ and the main effect of ER θ_{fH} could be modeled fixed effects, since

previous evidence indicates ER as a source breast cancer heterogeneity (Garcia-Closas *et al.*, 2013). And the main effects of PR and HER2 and other potential interactions effects can be modeled as random effect ($\mathbf{u}$).

Under the mixed effect two-stage model, the global association test would be: $\mathrm{H}_0^{\mathrm{A}} : \boldsymbol{\theta}_{\mathrm{f}} = \mathbf{0}$, $\sigma^2 = 0$ versus $\mathrm{H}_1^{\mathrm{A}} : \boldsymbol{\theta}_{\mathrm{f}} \neq \mathbf{0}$ or $\sigma^2 \neq 0$. And the global etiology heterogeneity test would be: $\mathrm{H}_0^{\mathrm{EH}} : \boldsymbol{\theta}_{\mathrm{fH}} = \mathbf{0}$, $\sigma^2 = 0$ versus $\mathrm{H}_1^{\mathrm{EH}} : \boldsymbol{\theta}_{\mathrm{fH}} \neq \mathbf{0}$ or $\sigma^2 \neq 0$.

To derive the score statistics for under the global null $\mathrm{H}_0^{\mathrm{A}} : \boldsymbol{\theta}_{\mathrm{f}} = \mathbf{0}$, $\sigma^2 = 0$, the common approach is to take the partial derivatives of loglikelihood with respective to $\boldsymbol{\theta}_{\mathrm{f}}$ and σ^2 respectively. However, under the null hypothesis, the score for $\boldsymbol{\theta}_{\mathrm{f}}$ follows a normal distribution and for σ^2 follows a mixture of chi-square distribution (see more details in Supplementary Section 3). With the correlation between the two scores, getting the joint distribution between the two becomes very complicated. Inspired by methods for the rare variants testing (Sun *et al.*, 2013), we propose to modify the derivations of score statistics so that we can get two independent scores. First for $\boldsymbol{\theta}_{\mathrm{f}}$, we derive the score test statistics $Q_{\boldsymbol{\theta}_{\mathrm{f}}}$ under the global null hypothesis $\mathrm{H}_0^{\mathrm{A}} : \boldsymbol{\theta}_{\mathrm{f}} = \mathbf{0}$, $\sigma^2 = 0$ as usual. But for the score test statistics for σ^2, we derived the the score statistics Q_{σ^2} under the null hypothesis $\mathrm{H}_0 : \sigma^2 = 0$ without constraining $\boldsymbol{\theta}_{\mathrm{f}}$. Through this procedure, the two score test statistics $Q_{\boldsymbol{\theta}_{\mathrm{f}}}$ and Q_{σ^2} can be proved as independent (proof in Supplementary Section 4). With the independence property and the asymptotic distribution of the two score test statistics, the Fisher's procedure (Koziol and Perlman, 1978) can be used to combine the p-value coming out from the two independent tests.

The score statistics of fixed effect $\boldsymbol{\theta}_{\mathrm{f}}$ under the global null $\mathrm{H}_0^{\mathrm{A}} : \boldsymbol{\theta}_{\mathrm{f}} = \mathbf{0}$, $\sigma^2 = 0$ would be:

$$Q_{\boldsymbol{\theta}_{\mathrm{f}}} = (\mathbf{Y} - \mathbf{P}_{\mathrm{f}})^T \mathbf{G}_M \mathbf{Z}_{\mathrm{f}} \tilde{\mathbf{I}}_{\mathrm{f}}^{-1} \mathbf{Z}_{\mathrm{f}}^T \mathbf{G}_M^T (\mathbf{Y} - \mathbf{P}_{\mathrm{f}}) \sim \chi^2_{l_{\mathrm{f}}}, \tag{8}$$

where $\mathbf{P}_{\mathrm{f}} = E_{\boldsymbol{\theta}_{\mathrm{f}}=\mathbf{0},\sigma^2=0}(\mathbf{Y}|\mathbf{X};\hat{\boldsymbol{\lambda}})$. Here $\tilde{\mathbf{I}}_{\mathrm{f}}$ has the same definition as Equation 5, but substitute $\mathbf{Z}_{\mathbf{G}}$ with $\mathbf{Z}_{\mathrm{f}}$. Under the null hypothesis, $Q_{\boldsymbol{\theta}_{\mathrm{f}}}$ follows a χ^2 distribution, and the degrees of freedom l_{f} is the same as the length of $\boldsymbol{\theta}_{\mathrm{f}}$.

To explicitly express the Q_{σ^2}, let $\boldsymbol{\tau} = (\boldsymbol{\theta}_{\mathrm{f}}^T, \boldsymbol{\lambda}^T)^T$ be the second stage fixed effect, and $\mathbf{Z}_{\boldsymbol{\tau}}$ is the corresponding second stage design matrix. The variance component score statistics of σ^2 under the null hypothesis: $\mathrm{H}_0 : \sigma^2 = 0$ without constraining $\boldsymbol{\theta}_{\mathrm{f}}$ would be:

$$Q_{\sigma^2} = (\mathbf{Y} - \mathbf{P}_{\mathrm{r}})^T \mathbf{G}_M \mathbf{Z}_r \mathbf{Z}_r^T \mathbf{G}_M^T (\mathbf{Y} - \mathbf{P}_{\mathrm{r}}) \sim \sum_{i=1}^{s} \rho_i \chi^2_{i,1}, \tag{9}$$

where $\mathbf{P}_{\mathrm{r}} = E_{\sigma^2=0}(\mathbf{Y}|\mathbf{G},\mathbf{X};\hat{\boldsymbol{\tau}})$, and $\hat{\boldsymbol{\tau}}$ is the MLE under the null hypothesis: $\mathrm{H}_0 : \sigma^2 = 0$. Under the null hypothesis, Q_{σ^2} follows a mixture of chi square distribution (derivation details in Supplementary Section 3), where $\chi^2_{i,1}$ i.i.d. follows χ^2_1. $(\rho_1, \ldots, \rho_s)$ are the eigenvalues of $\tilde{\mathbf{I}}_{\mathrm{r}} = \mathbf{I}_{\mathbf{uu}} - \mathbf{I}_{\mathbf{u}\boldsymbol{\tau}}^T \mathbf{I}_{\boldsymbol{\tau}\boldsymbol{\tau}}^{-1} \mathbf{I}_{\boldsymbol{\tau}\mathbf{u}}$, with $I_{\mathbf{uu}} = \mathbf{Z}_r^T \mathbf{G}_M^T \mathbf{W}_{\mathrm{r}} \mathbf{G}_M \mathbf{Z}_{\mathrm{r}}$, $\mathbf{I}_{\boldsymbol{\tau}\boldsymbol{\tau}} = \mathbf{Z}_{\boldsymbol{\tau}}^T \mathbf{C}_M^T \mathbf{W}_{\mathrm{r}} \mathbf{C}_M \mathbf{Z}_{\boldsymbol{\tau}}$ and $\mathbf{I}_{\boldsymbol{\tau}u} = \mathbf{I}_{u\boldsymbol{\tau}}^T = \mathbf{Z}_{\boldsymbol{\tau}}^T \mathbf{C}_M^T \mathbf{W}_{\mathrm{r}} \mathbf{G}_M \mathbf{Z}_r$. The weighted matrix $\mathbf{W}_{\mathrm{r}}$ has the same definition as in Section II C , but evaluated under the null hypothesis $\mathrm{H}_0 : \sigma^2 = 0$. The Davies exact method (Davies, 1980) is used here to calculate the p-value of the mixture of chi square distribution.

To use Fisher's procedure to combine the p-value coming out from the two indepent score test statistics. Let $P_{\boldsymbol{\theta}_{\mathrm{f}}} = Pr(Q_{\boldsymbol{\theta}_{\mathrm{f}}} \geq \chi^2_{l_{\mathrm{f}}})$ and $P_{\sigma^2} = Pr(Q_{\sigma^2} \geq \sum_{i=1}^{s} \rho_i \chi^2_{i,1})$. Under the null hypothesis $\mathrm{H}_0^{\mathrm{A}} : \boldsymbol{\theta}_{\mathrm{f}} = \mathbf{0}$, $\sigma^2 = 0$, $-2\log(P_{\boldsymbol{\theta}_{\mathrm{f}}}) - 2\log(P_{\sigma^2})$ follows χ^2_4. Then the p-value P_{mix} of mixed effect two-stage model score test under the null hypothesis would be:

$$P_{\mathrm{mix}} = Pr\left\{-2\log(P_{\boldsymbol{\theta}_{\mathrm{f}}}) - 2\log(P_{\sigma^2}) \geq \chi^2_4\right\}. \tag{10}$$

The extension of the score statistics to global test for etiology heterogeneity, $\mathrm{H}_0^{\mathrm{EH}} : \boldsymbol{\theta}_{\mathrm{fH}} = \mathbf{0}$, $\sigma^2 = 0$, would be straightforward.

III. SIMULATION EXPERIMENTS

In this section, large scale simulations across a wide range of practical scenarios are conducted to evaluate the type I error and power of the fixed effect and mixed effect two-stage models. Data were simulated to mimic the PBCS. Four tumor characteristics were simulated: ER (positive vs. negative), PR (positive vs negative), HER2 (positive vs. negative) and grade (ordinal 1, 2, 3). This defined a total of $2^3 \times 3 = 24$ breast cancer subtypes.

In each case control simulation, genotype data $\mathbf{G}$ was assumed to be under Hardy-Weinberg equilibrium in the underlying population with a minor allele frequency (MAF) of 0.25. An additional covariate, $\mathbf{X}$, was simulated as a standard normal distribution independent of $\mathbf{G}$. We used polytomous logistic regression model as Equation 11 to simulate a multinomial outcome with 25 groups, one for the control subjects, and the other 24 for different cancer subtypes.

$$Pr(D_i = m|X_i) = \frac{\exp(\alpha_m + \beta_m G_i + 0.05X_i)}{1 + \sum_{m=1}^{M} \exp(\alpha_m + \beta_m G_i + 0.05X_i)}, \tag{11}$$

where β_m is the log OR of G for mth subtype v.s. control. The effect of covariate $\mathbf{X}$ was set as 0.05 across all the subtypes. By using the frequency of 24 breast cancer subtypes estimated from Breast Cancer Association Consortium data (Supplementary Table 1), we computed the corresponding polytomous logistic regression intercept parameters α_m . The cases and controls ratio was set to be around 1:1, and the proportion of ER+, PR+ and

HER2+ were set to be 0.81, 0.68, and 0.17, respectively. The proportion of grade 1, 2, and 3 were 0.20, 0.48, and 0.32, respectively. The missing tumor markers were randomly selected and the missing rate of ER, PR, HER2, and grade were set to be 0.17, 0.25, 0.42, and 0.27, respectively. Under this simulation setting, around 70% breast cancer cases had at least one missing tumor characteristic.

A. Type I error

In this subsection, we evaluated the type I error of global association test, global heterogeneity test, and individual tumor characteristic heterogeneity test under the global null hypothesis. We generated data by setting $\beta_m = 0$ in Equation 11, where none of the subtypes is associated with genotypes. The total sample size n was set to be 5,000, 50,000, and 100,000. And 2.4×10^7 simulations were conducted to evaluate the type I error at $\alpha = 10^{-4}, 10^{-5}$ and, 10^{-6} level.

We applied both MTOP and FTOP with an additive two-stage model as in Equation 12, where the subtype-specific case-controls log ORs were specified into the case-control log OR of a baseline disease subtype (ER- , PR- , HER2-, grade 1) and the main effects associated with the four tumor markers. All of the second stage interactions parameters were constrained to be 0. Furthermore, the MTOP assumed the baseline parameter and the ER case-case parameter as fixed effects, and the PR, HER2, and grade case-case parameters as random effects. The global association test and global heterogeneity test were implemented using both MTOP and FTOP, but the individual tumor characteristic heterogeneity test

could only be implemented with FTOP.

$$\beta_{s_1 s_2 \ldots s_K} = \theta^{(0)} + \sum_{k_1=1}^{4} \theta_{k_1}^{(1)} s_{k1}. \tag{12}$$

Table I presents the simulated estimated type I error under the global null hypothesis. As expected, the type I error for both MTOP and FTOP tended to be lower for the simulated sample size of 5000, but with larger samples sizes all tests report nearly correct type I error, demonstrating the validity of MTOP and FTOP.

B. Statistical power

In this subsection, we present the statistical power of MTOP and FTOP under three different scenarios using our breast cancer example: I. no heterogeneity between tumor markers, II. heterogeneity according to one tumor marker, and III. heterogeneity according to multiple tumor markers. We generated the subtypes through Equation 11. Under the scenario I, we set β_m as 0.05 for all the subtypes, thus no heterogeneity of the ORs between $\mathbf{G}$ and the subtypes. For scenarios II and III, it was assumed that β_m followed the additive second stage structure as in Equation 12. Under scenarios II, we simulated a situation with only ER heterogeneity by setting the baseline effect $\theta^{(0)}$ to be 0, the case-case parameter for ER $\theta_1^{(1)}$ was set to be 0.05, and the PR, HER2, and grade case-case parameters to be 0. For scenario III, we simulated a situation with heterogeneity according to all 4 tumor markers by setting the baseline effect $\theta^{(0)}$ set to be 0, the ER $\theta_1^{(1)}$ case-case parameter to be 0.05, and all the other three case-case parameters were set to follow a normal distribution with mean 0 and variance 4×10^{-4}. Under this scenario, all the tumor characteristics contribute

subtype-specific heterogeneity. The total sample size n was set to be 25,000, 50,000, and 100,000. We performed 10^5 simulations to evaluate the power at $\alpha < 10^{-3}$ level.

We compared the statistical power to detect the genetic association between MTOP, FTOP, a standard logistic regression, polytomous logistic regression, and a two-stage model that only uses cases with complete tumor characteristics. The same additive second stage structure as Section III A was used for MTOP and FTOP. When we applied MTOP, FTOP and polytomous model, we removed all the subtypes with fewer than 10 cases to avoid unconvergence of the model. All the methods were set to test the overall association between $\mathbf{G}$ and the risk of the cancer.

To evaluate the different methods under a larger number of tumor characteristics, we added two additional binary tumor characteristics to the previous breast cancer example. This defined a total of $2^5 \times 3 = 96$ cancer subtypes. Similar to the four tumor characteristics simulations, we generated subtypes by polytomous model as in Equation 11, and simulated data under three different scenarios: I. no heterogeneity II. one tumor marker drove the heterogeneity, and III. multiple tumor markers driving tumor heterogeneity. Total sample size was set to be 25,000, 50,000, and 100,000. The two additional tumor characteristics were randomly selected to missing with 5% missing rate. Under this setting, around 77% of the cases have at least one tumor characteristics missing. We performed 10^5 simulations to evaluate the power at $\alpha < 10^{-3}$ level.

Figure 1 and 2 shows the power comparison between the five methods under different scenarios. Overall, MTOP had robust power under all the heterogeneity scenarios. Under scenario I with no subtype-specific heterogeneity, standard logistic regression had the

highest power, but suffered from substantial power loss when heterogeneity existed between subtypes. When heterogeneity was introduced in scenarios II and III, MTOP, followed by FTOP, consistently demonstrated the highest power among the five methods. The higher power observed in MTOP, relative to FTOP, ranged from 102% to 168%. Under the scenarios with four tumor characteristics (Figure 1) the difference in degrees of freedom between MTOP and FTOP were small, therefore MTOP had only a slight power advantage. However, with six tumor markers (Figure 2), the differences in degrees of freedom between MTOP and FTOP becomes more apparent, as does the greater power of MTOP. FTOP is least efficient in scenarios of none or little heterogeneity, such as scenarios I and II, but with increasing sources of heterogeneity, such-as scenario III, the power of MTOP and FTOP are more similar.

Simulation study also shows that incorporation of cases with missing tumor characteristics significantly increased the power of the methods. Under the four tumor markers setting with around 70% incomplete cases, the power of MTOP was between 202% to 905% greater compared to the original two-stage model using only complete data. Under the six tumor markers setting with around 77% incomplete cases, the two-stage model with only complete data lost more power compared to the four tumor markers setting; however, MTOP maintained similar power.

Overall, in the scenario of no heterogeneity the standard logistic regression demonstrated the most powerful. However, in the presence of subtype heterogeneity, MTOP was the most powerful method. The polytomous model had the lowest power across all of the settings.

IV. APPLICATION TO THE POLISH BREAST CANCER STUDY (PBCS)

In this section, we used data from the PBCS, a population-based breast cancer case-control study conducted in Poland between 2000 and 2003 (García-Closas *et al.*, 2006). The study population consisted of 2,078 cases of histologically or cytologically confirmed invasive breast cancer, and 2,219 women without a history of breast cancer at enrollment. Tumor characteristic information on ER, PR, and grade were available from pathology records (García-Closas *et al.*, 2006) and information on HER2 status was available from immunohistochemical staining of tissue microarray blocks (Yang *et al.*, 2007). We used genome-wide genotying data to compare MTOP, FTOP, standard logistic regression, and polytomous logistic regression to detect SNPs associated with breast cancer risk.

Table II presents the sample size distribution of the tumor characteristics. Combinations of the four tumor characteristics define 24, mutually exclusive breast cancer subtypes. Subtypes with less than 10 cases were excluded, leaving 17 subtypes that were evaluated. Both MTOP and FTOP used additive second stage design structure as in Equation 12. In addition, MTOP assumed the baseline parameter and the ER case-case parameter as fixed effects, and the case-case parameters of PR, HER2 and grade as random effects. We put ER as a fixed effect because of the previously reported heterogeneity of genetic association by ER status (Garcia-Closas *et al.*, 2013). Genotype imputation was done using IMPUTE2 based on 1000 Genomes Project dataset as reference panel (Michailidou *et al.*, 2017; Milne *et al.*, 2017). In total, 7,017,694 common variants on 22 auto chromosomes with MAF $\geq 5\%$

were included in the analysis. In the four models, we adjusted for age, and the first four genetic principal components to account for population stratification bias.

Figure 3 shows the Manhattan plot of genome-wide associations analysis with PBCS using the four different methods. MTOP, FTOP and standard logistic regression identified a known susceptibility variant in the FGFR2 locus on chromosome 10 (Michailidou *et al.*, 2013), with the most significant SNP being rs11200014 ($P < 5 \times 10^{-8}$). Further, both MTOP and FTOP identified a second known susceptibility locus on chromosome 11 (CCND1) (Michailidou *et al.*, 2017), with the most significant SNP in both models being rs78540526 ($P < 5 \times 10^{-8}$). The individual heterogeneity test of this SNP showed evidence for heterogeneity by ER (P=0.011) and grade (P=0.024). Notably, the CCND1 locus was not genome-wide significant in standard logistic regression or polytomous models. The QQ plots of the four GWAS analysis can be found in Supplementary Figure 1.

Next, we compared the ability of the same MTOP and standard logistic regressions to detect 178 previously identified breast cancer susceptibility loci (Michailidou *et al.*, 2017). As shown in Table III, for eight of the 178 loci, the MTOP global association test p value was more than ten fold lower compared to the standard logistic regression p value. In the MTOP model, these eight loci all had significant global heterogeneity tests ($P < 0.05$). Confirming these results, in a previous analysis applying MTOP to 106,571 breast cancer cases and 95,762 controls, these eight loci were reported to have significant global heterogeneity (Ahearn *et al.*, 2017).

V. DISCUSSION

We present a series of novel methods for performing genetic association testing for cancer outcomes accounting for potential heterogeneity across subtypes defined by multiple, correlated, tumor characteristics. These methods efficiently account for multiple testing, correlation between markers, and missing tumor data. Under the model framework, we develop a computational efficient score statistics (FTOP) and propose a new mixed effect two-stage model (MTOP) which could model the subtypes heterogeneity effect (response variables) as random effects. We demonstrate that MTOP has greater statistical power in the presence of subtype heterogeneity than either standard case-control logistic regression or polytomous logistic regression.

Several statistical methods have been proposed to study the etiological heterogeneity of cancer subtypes (Chatterjee, 2004; Rosner *et al.*, 2013; Wang *et al.*, 2015). A recent review showed the well controlled type-one error and good statistical power of two-stage model among these methods (Zabor and Begg, 2017). However, previous two-stage models have not accounted for missing tumor marker, a common problem of epidemiological studies. We show that by incorporating the EM algorithm into the two-stage model we can take advantage of all available information and make substantial gains in statistical power (as shown in Figure 1). Moreover, the newly proposed mixed effect model can mitigate the degrees of freedom penalty caused by analyzing a large number of tumor characteristics.

Notably, the computation time of MTOP is markedly greater than FTOP due to estimating the coefficients of covariates. To construct the score tests in FTOP, the coefficients of

covariates need to be estimated once under the null hypothesis, while for MTOP they need to be estimated for every SNP. The computational complexity of FTOP is O(NMS_θ), with S_θ as the number of second stage parameters of $\mathbf{G}$; whereas the computational complexity for MTOP is O($N^3M^2P^2S_{\boldsymbol{\tau}}$), with $S_{\boldsymbol{\tau}}$ as the number of fixed effect second stage parameters of $\mathbf{G}$ and $\mathbf{X}$.

Currently, we have only implemented the linear kernel in MTOP, but other common kernel functions that capture the similarity between tumor characteristics could be implemented in the future. If there is prior knowledge about the overlapping genetic architecture across different tumor subtypes, this could help to choose the kernel function and improve the power of the methods.

In conclusion, we have proposed an efficient and systematic approach for incorporating tumor characteristics information to identify genetic associations in the presence of subtype heterogeneity. The methods leverage all available tumor information and have robust statistical power. We have limited our demonstration of the benefit of these methods to analyzing the association between genetic variants and breast cancer subtypes; however these methods can easily be applied to the analysis of other non-genetic risk factors and/or other endpoints characterized by subtypes. The proposed methods have been implemented in a user-friendly and high-speed R statistical package called TOP (https://github.com/andrewhaoyu/TOP), which includes all the core functions implemented in C code.

VI. SUPPLEMENTARY MATERIAL

In Supplementary Section 1, we describe generalizing two-stage polytomous model to multiple variates with different second stage design matrix. In Supplementary Section 2, we derive the EM algorithm under two-stage model. In Supplementary Section 3, we derive the variance component score statistics in two-stage model. In Supplementary Section 4, we prove the independence between $Q_{\boldsymbol{\theta}_{\mathrm{f}}}$ and Q_{σ^2}. The Supplementary Table 1 contains the 24 breast cancer subtypes frequency estimated from Breast Cancer Association Consortium data. The Supplementary Figure 1 is the QQ plot of GWAS with PBCS.

ACKNOWLEDGMENTS

This work was supported by funds from the NCI Intramural Research Program and Bloomberg Distinguished Professorship endowment. The simulation experiments and data analysis were implemented using the high performance computation Biowulf cluster at National Institutes of Health, USA.

Ahearn, T. *et al.* (**2017**). "Novel analysis incorporating multiple tumor characteristics provide evidence of highly heterogeneous associations for known breast cancer risk loci," The American Society of Human Genetics poster .

Barnard, M. E., Boeke, C. E., and Tamimi, R. M. (**2015**). "Established breast cancer risk factors and risk of intrinsic tumor subtypes," Biochim Biophys Acta **1856**(1), 73–85.

Begg, C. B., and Zhang, Z. (**1994**). "Statistical analysis of molecular epidemiology studies employing case-series.," Cancer Epidemiology and Prevention Biomarkers **3**(2), 173–175.

Cancer Genome Atlas, N. (**2012**). "Comprehensive molecular portraits of human breast tumours," Nature **490**(7418), 61–70.

Cancer Genome Atlas Research, N. (**2012**). "Comprehensive genomic characterization of squamous cell lung cancers," Nature **489**(7417), 519–25.

Cancer Genome Atlas Research, N. (**2014**). "Comprehensive molecular profiling of lung adenocarcinoma," Nature **511**(7511), 543–50.

Chatterjee, N. (**2004**). "A two-stage regression model for epidemiological studies with multivariate disease classification data," Journal of the American Statistical Association **99**(465), 127–138.

Curigliano, G., Burstein, H. J., P Winer, E., Gnant, M., Dubsky, P., Loibl, S., Colleoni, M., Regan, M. M., Piccart-Gebhart, M., Senn, H.-J. *et al.* (**2017**). "De-escalating and escalating treatments for early-stage breast cancer: the st. gallen international expert consensus conference on the primary therapy of early breast cancer 2017," Annals of Oncology **28**(8), 1700–1712.

Davies, R. B. (**1980**). "Algorithm as 155: The distribution of a linear combination of χ 2 random variables," Journal of the Royal Statistical Society. Series C (Applied Statistics) **29**(3), 323–333.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (**1977**). "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society. Series B **39**(1), 1–38.

Dubin, N., and Pasternack, B. S. (**1986**). "Risk assessment for case-control subgroups by polychotomous logistic regression," American journal of epidemiology **123**(6), 1101–1117.

García-Closas, M. *et al.* (**2006**). "Established breast cancer risk factors by clinically important tumour characteristics," British Journal of Cancer **95**(1), 123–129.

Garcia-Closas, M. *et al.* (**2013**). "Genome-wide association studies identify four er negative-specific breast cancer risk loci," Nat Genet **45**(4), 392–8, 398e1–2.

Gortmaker, S. L., Hosmer, D. W., and Lemeshow, S. (**1994**). "Applied Logistic Regression.," Contemporary Sociology .

Koziol, J. A., and Perlman, M. D. (**1978**). "Combining independent chi-squared tests," Journal of the American Statistical Association **73**(364), 753–763.

Lin, X. (**1997**). "Variance component testing in generalised linear models with random effects," Biometrika **84**(2), 309–326.

Louis, T. A. (**1982**). "Finding the observed information matrix when using the em algorithm," Journal of the Royal Statistical Society. Series B (Methodological) 226–233.

MacArthur, J. *et al.* (**2016**). "The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog)," Nucleic acids research **45**(D1), D896–D901.

Michailidou, K. *et al.* (**2013**). "Large-scale genotyping identifies 41 new loci associated with breast cancer risk," Nature Genetics .

Michailidou, K. *et al.* (**2017**). "Association analysis identifies 65 new breast cancer risk loci," Nature **551**(7678), 92–94.

Milne, R. L. *et al.* (**2017**). "Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer," Nature genetics **49**(12), 1767.

Perou, C. M. *et al.* (**2000**). "Molecular portraits of human breast tumours," Nature **406**(6797), 747–52.

Prat, A. *et al.* (**2015**). "Clinical implications of the intrinsic molecular subtypes of breast cancer," Breast .

Rosner, B. *et al.* (**2013**). "Breast cancer risk prediction with heterogeneous risk profiles according to breast cancer tumor markers," American Journal of Epidemiology **178**(2), 296–308.

Sun, J., Zheng, Y., and Hsu, L. (**2013**). "A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies," Genetic Epidemiology **37**(4), 334–344.

Visscher, P. M. *et al.* (**2017**). "10 years of gwas discovery: Biology, function, and translation," Am J Hum Genet **101**(1), 5–22.

Wang, M., Kuchiba, A., and Ogino, S. (**2015**). "A meta-regression method for studying etiological heterogeneity across disease subtypes classified by multiple biomarkers," American Journal of Epidemiology .

Wu, M. C. *et al.* (**2011**). "Rare-variant association testing for sequencing data with the sequence kernel association test," American Journal of Human Genetics **89**(1), 82–93.

Yang, X. R. *et al.* (**2007**). "Differences in risk factors for breast cancer molecular subtypes in a population-based study," Cancer Epidemiology and Prevention Biomarkers **16**(3), 439–443.

Yi, Y., and Wang, X. (**2011**). "Comparison of wald, score, and likelihood ratio tests for response adaptive designs," Journal of Statistical Theory and Applications **10**(4), 553–569.

Zabor, E. C., and Begg, C. B. (**2017**). "A comparison of statistical methods for the study of etiologic heterogeneity," Statistics in medicine **36**(25), 4050–4060.

Zhang, D., and Lin, X. (**2003**). "Hypothesis testing in semiparametric additive mixed models.," Biostatistics **4**(1), 57–74.

TABLE I. Type one error estimates of MTOP, FTOP with 2.4×10^7 randomly simulated samples. Global test for association and global test for heterogeneity were applied with FTOP and MTOP. Heterogeneity test for a tumor marker was applied with only FTOP. All of the type error rates are divided by the α level.

		MTOP			FTOP		
Interested tests	Total sample size	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-6}$	$\alpha = 10^{-4}$	$\alpha = 10^{-5}$	$\alpha = 10^{-6}$
Global association test	5,000	.99	.97	.88	.91	.91	.67
	50,000	.98	1.0	1.0	.99	1.0	.93
	100,000	1.0	.94	1.0	1.0	1.0	1.0
Global heterogeneity test	5,000	1.0	.97	.89	.92	.85	.55
	50,000	1.0	1.0	1.0	1.0	1.0	1.0
	100,000	1.0	.94	1.0	1.0	.98	.97
Heterogeneity test for	5,000				.92	.93	.76
a tumor marker	50,000				.98	.97	1.0
	100,000				1.0	.97	1.0

TABLE II. Sample size of four tumor characteristics in Polish Breast Cancer Study

	ER	PR	HER2		Grade
Positive	1316	1056	1246	Grade 1	356
Negative	594	847	254	Grade 2	968
Missing	168	157	578	Grade 3	554
				Missing	200

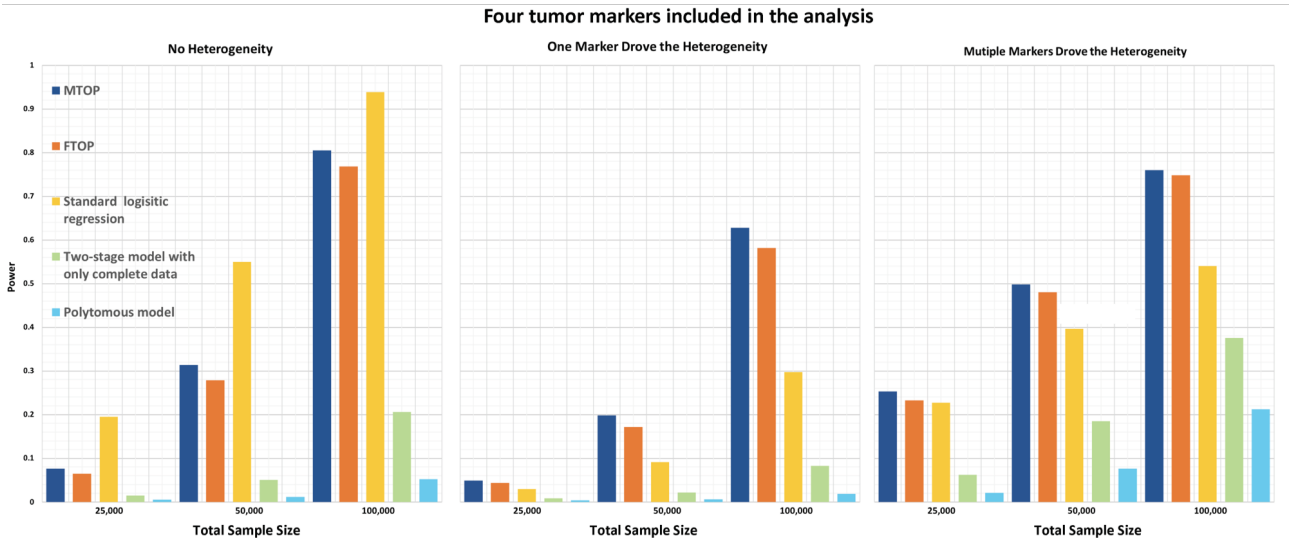

FIG. 1. Power comparison among MTOP, FTOP, standard logistic regression, two-stage model with only complete data and polytomous model with 10^5 random samples. Four tumor markers were included in the analysis. Three binary tumor marker and one ordinal tumor marker defined 24 cancer subtypes. The missing rate for the four markers were set as 0.17, 0.25, 0.42, and 0.27 respectively. Around 70% cases would be incomplete. The total sample size was set as 25,000, 50,000 and 100,000. The case control ratio was 1:1.

TABLE III. Analysis results of previously identified susceptibility loci. For the listed eight loci, MTOP global association test p value decreased more than ten fold compared to the standard logistic regression p value. All of the loci are significant in global heterogeneity test ($P < 0.05$).

SNP	Chr.[a]	Position	MAF[b]	Global association p	Standard analysis p	Global heterogeneity p
rs4973768	3	27,416,013	.47	3.12×10^{-2}	9.53×10^{-1}	9.48×10^{-3}
rs10816625	9	110,837,073	.06	4.98×10^{-2}	9.79×10^{-1}	2.22×10^{-2}
rs7904519	10	114,773,927	.46	6.51×10^{-2}	8.48×10^{-1}	3.07×10^{-2}
rs554219	11	69,331,642	.13	7.34×10^{-11}	1.42×10^{-7}	5.13×10^{-6}
rs11820646	11	129,461,171	.40	1.48×10^{-2}	8.62×10^{-1}	4.53×10^{-3}
rs2236007	14	37,132,769	.21	2.10×10^{-3}	1.93×10^{-1}	3.49×10^{-3}
rs1436904	18	24,570,667	.40	7.17×10^{-4}	6.61×10^{-2}	9.69×10^{-4}
rs1436904	22	29,121,087	.01	9.83×10^{-3}	1.61×10^{-1}	2.32×10^{-2}

[a]Chr. chromosome. [b] MAF, minor allele frequency.

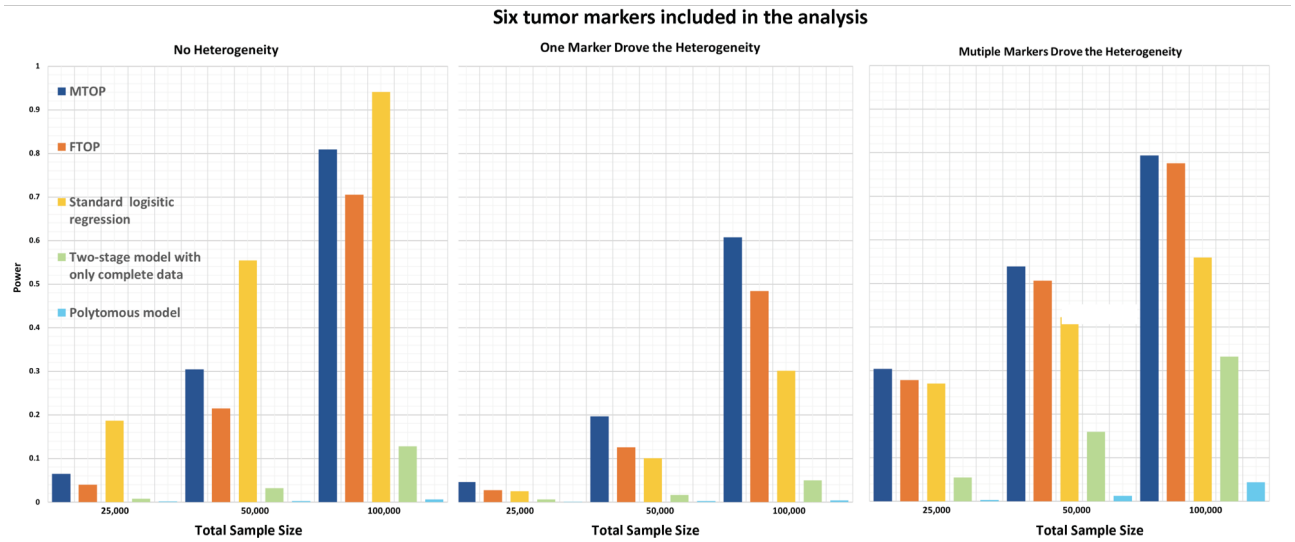

FIG. 2. Power comparison among MTOP, FTOP, standard logistic regression, two-stage model with only complete data and polytomous model with 10^5 random samples. Six markers were included in the analysis. Compared to the four markers analysis, two extra binary tumor markers were included in the analysis. The six tumor markers defined 96 subtypes. The missing rate of the two extra markers were 0.05. Around 77% cases would be incomplete. The total sample size was set as 25,000, 50,000 and 100,000. The power was estimated by controlling the type one error $\alpha < 10^{-3}$.

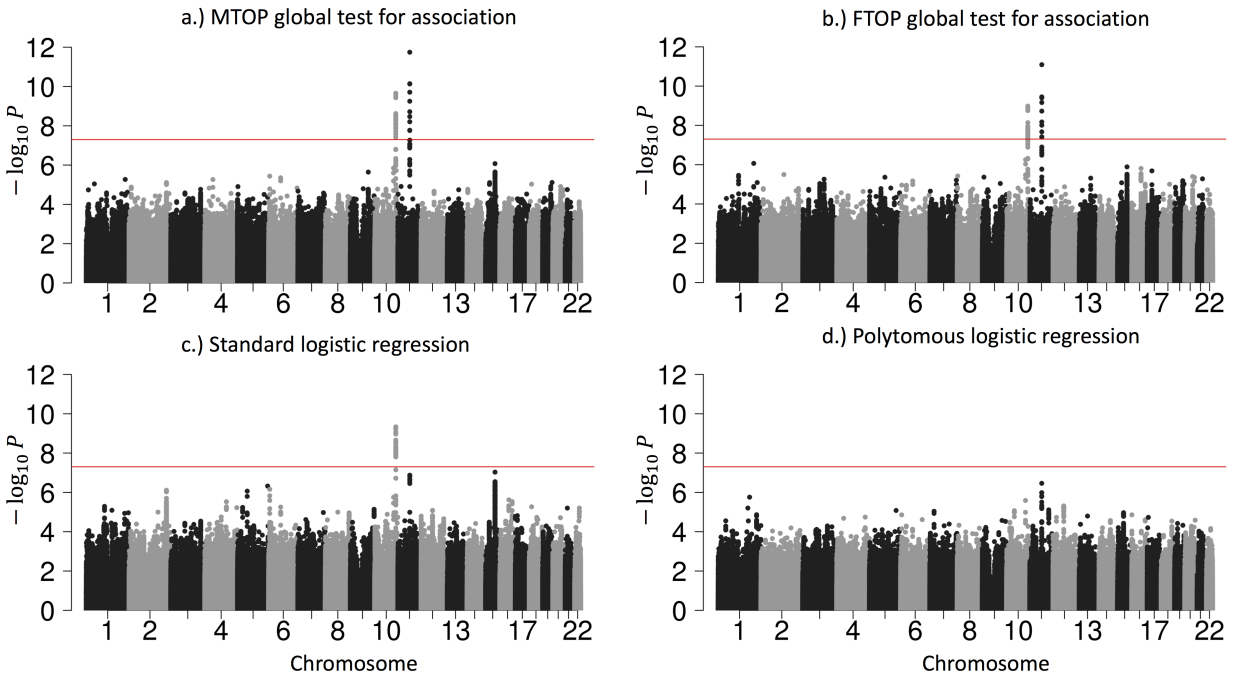

FIG. 3. Manhattan plot of genome-wide association analysis with PBCS using four different methods. PBCS have 2,078 invasive breast cancer and 2,219 controls. In total, 7,017,694 SNPs on 22 auto chromosomes with MAF more than 5% were included in the analysis. ER, PR, HER2 and grade were used to define breast cancer subtypes.