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: 23 October 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 all possible subtypes defined by the cross-classification of the tumor characteristics of interest. 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 *and others*, 2016). 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) (Perou *and others*, 2000; Prat *and others*, 2015).

Increasing numbers of epidemiologic studies with tumor specimens are allowing the characterization of cancers at the histological and molecular levels (Cancer Genome Atlas Research, 2014; Network, 2012). This provides tremendous opportunities to investigate distinct etiological pathways within cancer subtypes. For example, a breast cancer ER-negative specific GWAS reported 20 SNPs that were more strongly associated with risk of developing ER-negative than ER-positive disease (Milne *and others*, 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 *and others*, 2015).

The most common procedure for testing for associations between risk factors and cancer subtypes is fitting a standard logistic regression for each subtype versus a control group, then accounting for multiple testing. However, this procedure has several limitations. First, it's

common for cancer cases to have 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 can be highly correlated with each other. Testing each subtype separately without modeling the correlation limits the power. Finally, as the number of tumor markers increases, the number of cancer subtypes dramatically increases, thus the increased degrees of freedom penalizes 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) to model subtype-specific case-control odds ratios. In the second stage, the subtype-specific case-control odds ratios are decomposed to a case-control odds ratio for a reference subtype, case-case odds ratios of each tumor characteristic and higher order interactions between the case-case odds ratios for tumor characteristics. The two-stage model can reduce the degrees of freedom by constraining some or all of the higher order interactions to be 0. Moreover, the second stage parameters can be interpreted as the case-case odds ratios for tumor characteristics, which can measure the etiological heterogeneity between the associations of cancer subtypes.

Although, the two-stage model can improve the power compared to fitting standard logistic regression for each subtype (Chatterjee, 2004; Zabor and Begg, 2017), the method has not been widely applied to analyze data on multiple tumor characteristics. First, the two-stage model can not handle missing tumor characteristics, which is 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 can reduce the multiple testing burdens compared to traditional methods, as the number of tumor characteristics increases, the method can 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 cancer GWASs incorporating tumor characteristic data. Within this 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 many tumor characteristics, we propose modeling 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 *and others*, 1977). All combined, our work represents a conceptually distinct and practically important extension of earlier methods based on mixed-/fixed-effect models (Lin, 1997; Sun *and others*, 2013; Wu *and others*, 2011; Zhang and Lin, 2003) to the novel setting of modeling genetic associations with multiple tumor characteristics.

The paper is organized as follows. In Section II, we describe the proposed three different hypothesis tests, the missing data algorithm and two score tests. In Section III, we present the simulation results for type I error, power and computaion time. In Section IV, the proposed methods are illustrated with applications using Polish Breast Cancer Study (PBCS)

data. In Section V, we discuss the strengths and limitations of the methods, and future research directions.

II. METHOD

A. Two-stage polytomous logistic model

The details of two-stage polytomous logistic model are described earlier (Chatterjee, 2004) and we briefly summarize it for completeness. Suppose a disease can be classified using K disease characteristics and 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. Let D_i denote the disease status of subject i in the study that $D_i \in \{0, 1, 2, \cdots, M\}$, $i \in 1, \cdots, N$. $D_i = 0$ represents a control, and $D_i = m$ represents a case with disease subtype m. Let G_i be the genotype for subject i and $\mathbf{X}_i$ be a $P \times 1$ vector of other covariates we want to adjust for, 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 the SNP and other covariates 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 ratios defined in first stage model as

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

Here $\theta^{(0)}$ represents the case-control log odds ratio for a reference disease subtype compared to the control, $\theta_{k_1}^{(1)}$ represents the main effect of k_1th tumor characteristic, $\theta_{k_1 k_2}^{(2)}$ represents 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. The first stage and second stage parameters can be linked 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 a reference subtype, $\boldsymbol{\theta}_{\mathrm{H}}$ is a vector containing the main effects and interactions effects in the second stage. We will refer to $\boldsymbol{\theta}_{\mathrm{H}}$ as case-case parameters, 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 and second stage parameters. By constraining 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, details

of which could be found in Supplementary Section 1. We suggest not to perform second stage decomposition on the intercepts parameters of the first stage polytomous model, i.e., the coefficients of intercepts are saturated, because decomposing the intercepts equates to making assumptions on the prevalence of different cancer subtypes, which can potentially lead to 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 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 $\mathrm{H}_1^{\mathrm{EH}} : \boldsymbol{\theta}_{\mathrm{H}} \neq \mathbf{0}$. This test evaluates the 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 are significantly different for at least two different subtypes.

Notably, the global heterogeneity test does not identify which tumor characteristic(s) is/are driving the heterogeneity. 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 parameter ($\theta_{\mathrm{H}(k)}$) 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 the previous sections, we assume all tumor characteristics are observed for all cases. In epidemiological research it is very common to have missing tumor characteristics. This problem becomes exacerbated as the number of 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 *and others*, 1977) to find the maximum likelihood estimate (MLE) of the two-stage model, incorporating all available information from the study. Let $\mathbf{T}_{io}$ be the observed tumor characteristics of subject i and $Y_{im} = I(D_i = m)$ denote whether this subject is of disease subtype m. Given $\mathbf{T}_{io}$, the possible subtypes for subject i would be a limited subset of

all possible tumor subtypes, 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 $\boldsymbol{\delta} = (\boldsymbol{\theta}^T, \boldsymbol{\lambda}^T)^T$ 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. $I(Y_{im} \in \mathcal{Y}_{io})$ denotes 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 $\mathbf{C}_M = \mathbf{I}_M \otimes \mathbf{C}$. Let $\mathbf{W} = \mathbf{D} - \mathbf{A}\mathbf{A}^T$, with $\mathbf{D} = \mathrm{diag}(\mathbf{P} - \mathbf{P}_{\mathrm{mis}})$, $\mathbf{P} = E(\mathbf{Y}|\mathbf{C}; \delta)$, $\mathbf{P}_{\mathrm{mis}} = E(\mathbf{Y}|\mathbf{C}, \mathbf{T}_o; \delta)$, and $\mathbf{A} = \mathbf{D}(\mathbf{1}_M \otimes \mathbf{I}_N)$. During the tth iteration of the weighted Least square, $\mathbf{Y}^{*(t)} = \mathbf{W}^{(t)}(\mathbf{Y}^{\mathrm{E}} - \mathbf{P}^{(t)}) + \mathbf{C}_M \mathbf{Z} \boldsymbol{\delta}^{(t)}$, where $\mathbf{P}^{(t)}$ and $\mathbf{W}^{(t)}$ 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 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 can 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, making it much more computational efficient than the Wald test. Moreover, we only need to implement the EM algorithm once under the null hypothesis.Since we don't perform any second stage decomposition on the intercept parameters in first stage polytomous model, the correlations between the tumor characteristics are kept close to the empirical correlations for tumor markers. Most of the imputation power are due to the high correlation between the tumor markers. In 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 association of $\mathbf{X}$ with the tumor markers can improve the power of the EM algorithm. Since a single SNP $\mathbf{G}$ usually has a small effect, the fact that the effect of individual $\mathbf{G}$ is not incorporated in the EM algorithm itself doesn't result in much loss of efficiency.

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 $\mathbf{P}_{\mathrm{f}} = E_{\boldsymbol{\theta}=\mathbf{0}}(\mathbf{Y}|\mathbf{X}, \mathbf{T}_o; \hat{\boldsymbol{\lambda}})$, $\mathbf{P}_{\mathrm{f,mis}} = E(\mathbf{Y}|\mathbf{X}, \mathbf{T}_o; \hat{\boldsymbol{\lambda}})$, $\mathbf{D}_{\mathrm{f}} = \mathrm{diag}(\mathbf{P}_{\mathrm{f}} - \mathbf{P}_{\mathrm{f,mis}})$ 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 statistic $Q_{\boldsymbol{\theta}}$ for fixed-effect two stage model is:

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

The degree of freedom of FTOP is the same as the Wald test and they have similar asymptotic power (Yi and Wang, 2011). In GWAS where we need to perform millions of tests, we propose to first use FTOP to scan the whole genome for global association test and select the potential risk regions. Then for each SNP in the risk regions, we could estimate $\hat{\boldsymbol{\theta}}$ and use Wald test for 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 reduced 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 it 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. Under the breast cancer example, the baseline parameter ($\theta^{(0)}$) and the main effect of ER (θ_{fH}) can be modeled as fixed effects, since previous evidence indicates ER as a source breast cancer heterogeneity (García-Closas *and others*, 2013; Milne *and others*, 2017; Zhang *and others*, 2019), and the main effects of PR and HER2 and other potential interactions effects

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

To derive the score statistic for 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 (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 *and others*, 2013), we propose to modify the derivations of score statistic so that we can get two independent scores. First for $\boldsymbol{\theta}_{\mathrm{f}}$, we derive the score test statistic $Q_{\boldsymbol{\theta}_{\mathrm{f}}}$ 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 statistic for σ^2, we derive the the score statistic 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 (Supplementary Section 4). And we can use the Fisher's procedure (Koziol and Perlman, 1978) to combine the p-value coming out from the two independent tests. Similarly as FTOP, the EM algorithm under the null hypothesis of MTOP can efficiently handle the missing tumor marker problems given the high correlations between the tumor characteristics. However, since MTOP needs to estimate $\boldsymbol{\theta}_{\mathrm{f}}$ under the null hypothesis $\mathrm{H}_0 : \sigma^2 = 0$ for every single SNP, the computation speed for MTOP is slower than FTOP.

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

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

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 $\mathbf{Z_G}$ 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 statistic of σ^2 under the null hypothesis: $\mathrm{H}_0 : \sigma^2 = 0$ without constraining $\boldsymbol{\theta}_{\mathrm{f}}$ is:

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

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 $\tilde{\mathbf{I}}_{\mathrm{r}} = \mathbf{I_{uu}} - \mathbf{I}_{\mathbf{u}\boldsymbol{\tau}}^T \mathbf{I}_{\boldsymbol{\tau}\boldsymbol{\tau}}^{-1} \mathbf{I}_{\boldsymbol{\tau}\mathbf{u}}$, 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 $\mathbf{I}_{\boldsymbol{\tau}u} = \mathbf{I}_{u\boldsymbol{\tau}}^T = \mathbf{Z}_{\boldsymbol{\tau}}^T \mathbf{C}_M^T \mathbf{W}_{\mathrm{r}} \mathbf{G}_M \mathbf{Z}_r$, where $\mathbf{W}_{\mathrm{r}} = \mathbf{D}_{\mathrm{r}} - \mathbf{A}_{\mathrm{r}} \mathbf{A}_{\mathrm{r}}^T$, with $\mathbf{P}_{\mathrm{r}} = E_{\sigma^2=0}(\mathbf{Y}|\mathbf{G},\mathbf{X},\mathbf{T}_o;\hat{\boldsymbol{\tau}})$, $\mathbf{P}_{\mathrm{r,mis}} = E(\mathbf{Y}|\mathbf{G},\mathbf{X},\mathbf{T}_o;\hat{\boldsymbol{\tau}})$, $\mathbf{D}_{\mathrm{r}} = \mathrm{diag}(\mathbf{P}_{\mathrm{r}} - \mathbf{P}_{\mathrm{r,mis}})$ and $\mathbf{A}_{\mathrm{r}} = \mathbf{D}_{\mathrm{r}}(\mathbf{1}_M \otimes \mathbf{I}_N)$. The Davies exact method (Davies, 1980) is used here to calculate the p-value of the mixture of chi square distribution.

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

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

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

Large scale simulations across a wide range of practical scenarios were conducted to evaluate the type I error (Section III A), statistical power (Section III B) and and computation time (Supplementary Section 5) 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 simulation, genotype data $\mathbf{G}$ was simulated under Hardy-Weinberg equilibrium with a minor allele frequency (MAF) of 0.25. An additional covariate, $\mathbf{X}$, was simulated as standard normal independent of $\mathbf{G}$. We used polytomous logistic regression model (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 versus control. The effect of $\mathbf{X}$ was set as 0.05 for all subtypes. Using the frequency of the breast cancer subtypes from Breast Cancer Association Consortium (Supplementary Table 1) (Michailidou *and others*, 2017), we computed the corresponding polytomous logistic regression intercept parameters α_m. The case-to-control ratio was set around 1:1, and the proportions of ER+, PR+ and HER2+ were 0.81, 0.68, and 0.17. The proportions of grade 1, 2, and 3 were 0.20, 0.48, and 0.32. The missing tumor markers were selected randomly with missing rates of 0.17, 0.25, 0.42,

and 0.27 for ER, PR, HER2 and grade, respectively. Under this simulation, aproximately 70% cases had at least one missing tumor characteristic.

A. Type I error

We evaluated the type I error of the global association test, global heterogeneity test, and individual tumor marker 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. We conducted 2.4×10^7 simulations to evaluate the type I error at $\alpha = 1.0\times10^{-4}, 1.0\times\times10^{-5}$ and, 1.0×10^{-6} level.

We applied both MTOP and FTOP with an additive two-stage model as in Equation 12, where the subtype-specific case-control 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 and 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.

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

Table I presented the estimated type I errors under the global null hypothesis. Both MTOP and FTOP correctly controlled the type I errors, especially for the larger sample sizes. FTOP

is conservative with sample size of 5000, especially for $\alpha = 1.0 \times 10^{-6}$, however, the method is still valid. In the later sections, we generally used the additive second stage structure as specified in Equation 12 for both MTOP and FTOP unless otherwise specified.

B. Statistical power

We assessed the statistical power of all the proposed methods using various simulation settings at sample sizes of 25,000, 50,000, and 100,000. For each setting, we performed 2×10^5 simulations to evaluate the power at $\alpha = 5.0 \times 10^{-8}$ level.

1. Global association test

We simulated the data with three different scenarios: I. no heterogeneity between tumor markers, II. heterogeneity according to one tumor marker, and III. heterogeneity according to multiple tumor markers. We generated disease subtypes through Equation 11. Under scenario I, we set β_m as 0.08 for all the subtypes. For scenarios II and III, β_m was simulated following the additive second stage structure as in Equation 12. Under scenarios II, datasets were simulated with only ER heterogeneity by setting the case-case parameter for ER ($\theta_1^{(1)}$) as 0.08 and all the other as 0. For scenario III, we simulated a scenario with heterogeneity according to all 4 tumor markers by setting the baseline effect ($\theta^{(0)}$) to be 0, the ER case-case parameter ($\theta_1^{(1)}$) to be 0.08, and all the case-case parameters following a normal distribution with mean 0 and variance 4.0×10^{-4}. Under this scenario, all tumor characteristics contribute to the subtype-specific heterogeneity. Moreover, to evaluate different methods under a larger number of tumor characteristics, we conducted additional simulations by

adding two additional binary tumor characteristics to the previous four tumor characteristic setting. This defined a total of $2^5 \times 3 = 96$ cancer subtypes. The two additional tumor characteristics were randomly selected to be missing with 5% missing rate. Under this setting, around 77% of the cases have at least one tumor characteristics missing. We compared the statistical power to detect the overall association among FTOP, MTOP, standard logistic regression, FTOP with only complete data and polytomous logistic regression. For MTOP, FTOP and polytomous model, we removed all the subtypes with fewer than 10 cases to avoid potential nonconvergence of the model.

Overall, MTOP had robust power under all scenarios (Figure 1). Standard logistic regression had the highest power when there was no subtype-specific heterogeneity (Scenario I), but suffered from substantial power loss when heterogeneity existed between subtypes. MTOP, followed by FTOP, consistently demonstrated the highest power among the five methods when subtype-specific heterogeneity existed (scenarios II and III). The power gain of MTOP over FTOP ranged from 2% to 49%. The power gain was small when there were four tumor characteristics, because the difference in degrees of freedom between MTOP and FTOP was small. However, with six tumor markers, the power gain of MTOP was more apparent because the difference in degrees of freedom between MTOP and FTOP becomed larger. FTOP was least efficient in scenarios with no or little heterogeneity, such as scenarios I and II, but with increasing heterogeneity, such-as scenario III, the powers of MTOP and FTOP were more similar.

Simulation study also showed that incorporation of cases with missing tumor characteristics significantly increases the power of the methods. Under the four tumor markers

setting with around 70% incomplete cases, the power gain of FTOP was at least 200% compared to FTOP using only the complete data. Under the six tumor markers setting with around 77% incomplete cases, FTOP with only complete data lost more power compared to the four tumor markers setting; however, MTOP maintained higher power. Additional power simulation with sample size at 5,000 and a larger effect size was described in Supplementary Section 6. MTOP was the most powerful method when there were heterogeneous associations across cancer subtypes.

The previous simulations mainly focused on two-stage model with additive effects. Additional simulations were also implemented with pairwise interactions in the model. We simulated data with β_m following a second stage model with that included pairwise interactions (Equation 13) with the case-case parameter for ER ($\theta_1^{(1)}$) as 0.08, the pairwise interaction effect between ER and HER2 ($\theta_{13}^{(2)}$) as 0.04, and all the other parameters as 0. Four methods were evaluated including FTOP with/without pairwise interactions, MTOP with/without pairwise interactions (baseline and ER fixed). FTOP without interaction terms still had high power (Figure 2). However, FTOP with pairwise interaction structure had limited power due to incorporation of the interaction terms as fixed effects. On the other hand, MTOP and MTOP with pairwise interactions maintained a high power even when there were underlying interaction effect.

$$\beta_{s_1 s_2 \ldots s_K} = \theta^{(0)} + \sum_{k_1=1}^{K} \theta_{k_1}^{(1)} s_{k1} + \sum_{k_1=1}^{K} \sum_{k_2>k_1}^{K} \theta_{k_1 k_2}^{(2)} (s_{k_1} s_{k_2}) \tag{13}$$

2. Global heterogeneity test

Supplementary Figure 3 shows the simulation results for global heterogeneity tests under similar simulation settings as global association test. MTOP had highest power when there were heterogeneous associations across the subtypes.

3. Individual tumor marker heterogeneity test

We further evaluated the power of individual tumor marker heterogeneity test. The data was generated with four tumor characteristics with the ER case-case parameter ($\theta_1^{(1)}$) as 0.08 and all other parameters as 0. ER was randomly selected to be missing with the rate of 0.17, 0.30 and 0.50. We compared two different methods including FTOP with all four tumor characteristics and polytomous model. Polytomous model was set up to test each marker at a time. In the one marker at a time polytomous analysis, we removed cases which had missing data only on the relevant tumor marker. This way gave polytomous model more power since it avoided removing the cases which missed other tumor markers not in the analysis. FTOP with all four tumor characteristics had smaller power compared to polytomous model in testing the effect of ER (Supplementary Figure 3). Since FTOP included all four tumor characteristics, and the tumor markers were highly correlated, variability of underlying parameters were larger. However, the type I errors of polytomous model in testing PR, HER2 and grade were inflated under this case (Supplementary Figure 4). Under this simulation, these three markers had no effect. On the other hand, FTOP controlled the type error of all the tests.

Overall, for global test for association and global test for heterogeneity, when there was no heterogeneity, the standard logistic regression was the most powerful method. However,

in the presence of subtype heterogeneity, MTOP was the most powerful method. Even with large number pairwise interactions terms included, MTOP had stable power. The polytomous model had the lowest power across all of the settings.

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

We applied our proposed methods to the PBCS, a population-based breast cancer case-control study conducted in Poland between 2000 and 2003 (García-Closas *and others*, 2006). The study 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. Information on ER, PR, and grade were available from pathology records (García-Closas *and others*, 2006) and information on HER2 was available from immunohistochemical staining of tissue microarray blocks (Yang *and others*, 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.

Supplementary Table 2 presents the sample size distribution of the tumor characteristics. The four tumor characteristics defined 24 mutually exclusive breast cancer subtypes. Subtypes with less than 10 cases were excluded, leaving 17 subtypes in analysis. Both MTOP and FTOP used the additive second stage design (Equation 12). In addition, we modeled the baseline and ER case-case parameter as fixed effects in MTOP, and all other effects as random effects. We put ER as a fixed effect because of the previously reported heterogeneity in genetic association by ER (García-Closas *and others*, 2013; Milne *and others*, 2017; Zhang *and others*, 2019). Genotype imputation was done using IMPUTE2 based on 1000

Genomes Project as reference (Michailidou *and others*, 2017; Milne *and others*, 2017). In total, 7,017,694 common variants on 22 auto chromosomes with MAF $\geq 5\%$ were included in the analysis. In all the models, we adjusted for age and the first four genetic principal components to account for population stratification.

As Figure 3 shows, MTOP, FTOP and standard logistic regression all identified a known susceptibility variant in the FGFR2 locus on chromosome 10 (Michailidou *and others*, 2017), with the most significant SNP being rs11200014 ($P < 5.0 \times 10^{-8}$). Further, both MTOP and FTOP identified a second known susceptibility locus on chromosome 11 (CCND1) (Michailidou *and others*, 2017), with the most significant SNP in both models being rs78540526 ($P < 5.0 \times 10^{-8}$). 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 type I error of the four methods were well controlled (Supplementary Figure 6).

Additional sensitivity analysis of MTOP was implemented by specifying baseline, ER and grade as fixed effects, and PR and HER2 as random effects (Supplementary Figure 7). The results for MTOP with grade as fixed vs. random effect were similar. We also implemented MTOP and FTOP incorporating pairwise interactions in the second stage model (Supplementary Figure 8-9). With pairwise interactions, both MTOP and FTOP could detect FGFR2 and CCND1 with genome-wide significant threshold. However, the P-value of FTOP with pairwise interactions was less significant compared to FTOP without these interaction terms (for rs11200014, $P = 4.3 \times 10^{-8}$ vs. $P = 1.0 \times 10^{-9}$; for rs78540526, $P = 2.7 \times 10^{-10}$ vs. $P = 8.1 \times 10^{-12}$). The P-value of MTOP with pairwise interactions

was also less significant compared to MTOP without interaction terms (for rs11200014, P= 1.0×10^{-9} vs. P = 2.2×10^{-10}; for rs78540526, P = 1.7×10^{-11} vs. P = 1.8×10^{-12}). However, the power loss was smaller.

Next, we compared the ability of MTOP and standard logistic regressions to detect 178 previously identified breast cancer susceptibility loci (Michailidou *and others*, 2017). 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 (Table II). 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 *and others*, 2019).

V. DISCUSSION

We present a series of novel methods for performing genetic association testing for cancer outcomes accounting for potential heterogeneity across subtypes. These methods efficiently account for multiple testing, correlation between markers, and missing tumor data. Under the model framework, we develop two computationally efficient score tests, FTOP and MTOP, which models the underlying heterogeneity parameters in terms of fixed effects or mixed effects, respectively. We demonstrate these methods have greater statistical power in the presence of subtype heterogeneity than either standard or polytomous logistic regression analysis.

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

Incorporating missing tumor characteristics based on the proposed EM algorithm requires the assumption of missing at random (MAR), i.e. the mechanism of missing of the individual tumor characteristics can depend only on the other observed tumor characteristics and covariates, but not on the unobserved missing value themselves. For the analysis of tumor heterogeneity, it is possible that information on aggressive types of tumor may be systematically missing more often and if the missing tumor characteristics are important determinant of aggressiveness then the underlying assumption is violated. In general, dealing with non-ignorable missing data is a complex problem and certain sensitivity analysis could be performed to explore the degree of bias (Little and Rubin, 2019). In the context of genetic association testing, non-ignorable missingness can lead to inflated type I error only if missingness mechanism itself is related to the genetic variant under study. Further research is merited to explore the complex effects of non-ignorable missingness in type I error and power of the proposed tests.

The computation time of MTOP is greater than FTOP (Supplementary Section 5). To construct the score tests in FTOP, the coefficients of covariates need to be estimated once under the null hypothesis, while in MTOP they need to be estimated for every SNP. The computational complexity of FTOP is O(NM^2P^2), with P as the number of other covariates $\mathbf{X}$. The computational complexity for MTOP is hard to decide since it relates to the convergence threshold for EM algorithm and weighted least square.

Currently, we only implemented the linear kernel in MTOP, but other common kernels that capture the similarity between tumor characteristics could be used 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.

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 MATERIALS

In Supplementary Section 1, we generalized two-stage polytomous model to multivariate. In Supplementary Section 2, we derived the EM algorithm under two-stage model. In Supplementary Section 3, we derived the variance component score statistic in two-stage model. In Supplementary Section 4, we proved the independence between $Q_{\boldsymbol{\theta}_{\mathrm{f}}}$ and Q_{σ^2}. In Supplementary Section 5, we conducted computation time simulations. In Supplementary Section 6, we described the simulation results with sample size at 5,000. Supplementary Table 1 contained the 24 breast cancer subtypes frequency.Supplementary Table 2 presented sample

size of tumor characteristics in PBCS. Supplementary Figure 1 showed the computation time simulation results. Supplementary Figure 2-4 respectively presented power analysis of global association test with 5,000 subjects, global heterogeneity test and individual tumor marker heterogeneity test. Supplementary Figure 5 showed the type I error of individual tumor heterogeneity test for PR, HER2 and grade. Supplementary Figure 6 was the QQ plot of GWAS with PBCS. Supplementary Figure 7-9 respectively presented the GWAS with PBCS using MTOP with/without pairwise interactions and FTOP with/without pairwise interactions.

ACKNOWLEDGEMENTS

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. U. *and others*. (2019). Common breast cancer risk loci predispose to distinct tumor subtypes. *bioRxiv*, 733402.

Barnard, M. E., Boeke, C. E. and Tamimi, R. M. (2015). Established breast cancer risk factors and risk of intrinsic tumor subtypes. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1856**(1), 73–85.

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

Cancer Genome Atlas Research, Network. (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.

Davies, R. B. (1980). 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 (Methodological)* **39**(1), 1–22.

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. *and others*. (2006). Established breast cancer risk factors by clinically important tumour characteristics. *British journal of cancer* **95**(1), 123.

García-Closas, M. *and others*. (2013). Genome-wide association studies identify four er negative–specific breast cancer risk loci. *Nature genetics* **45**(4), 392.

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.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, Volume 793. John Wiley & Sons.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **44**(2), 226–233.

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

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

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

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

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

Prat, A. *and others*. (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast* **24**, S26–S35.

Rosner, B. *and others*. (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.

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* **182**(3), 263–270.

Wu, M. C. *and others*. (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. *and others*. (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.

Zhang, H. *and others*. (2019). Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *bioRxiv*, 778605.

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. 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	3.1×10^{-2}	9.5×10^{-1}	9.5×10^{-3}
rs10816625	9	110,837,073	.06	5.0×10^{-2}	9.8×10^{-1}	2.2×10^{-2}
rs7904519	10	114,773,927	.46	6.5×10^{-2}	8.5×10^{-1}	3.1×10^{-2}
rs554219	11	69,331,642	.13	7.3×10^{-11}	1.4×10^{-7}	5.1×10^{-6}
rs11820646	11	129,461,171	.40	1.5×10^{-2}	8.6×10^{-1}	4.5×10^{-3}
rs2236007	14	37,132,769	.21	2.1×10^{-3}	1.9×10^{-1}	3.5×10^{-3}
rs1436904	18	24,570,667	.40	7.2×10^{-4}	6.6×10^{-2}	9.7×10^{-4}
rs1436904	22	29,121,087	.01	9.8×10^{-3}	1.6×10^{-1}	2.3×10^{-2}

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

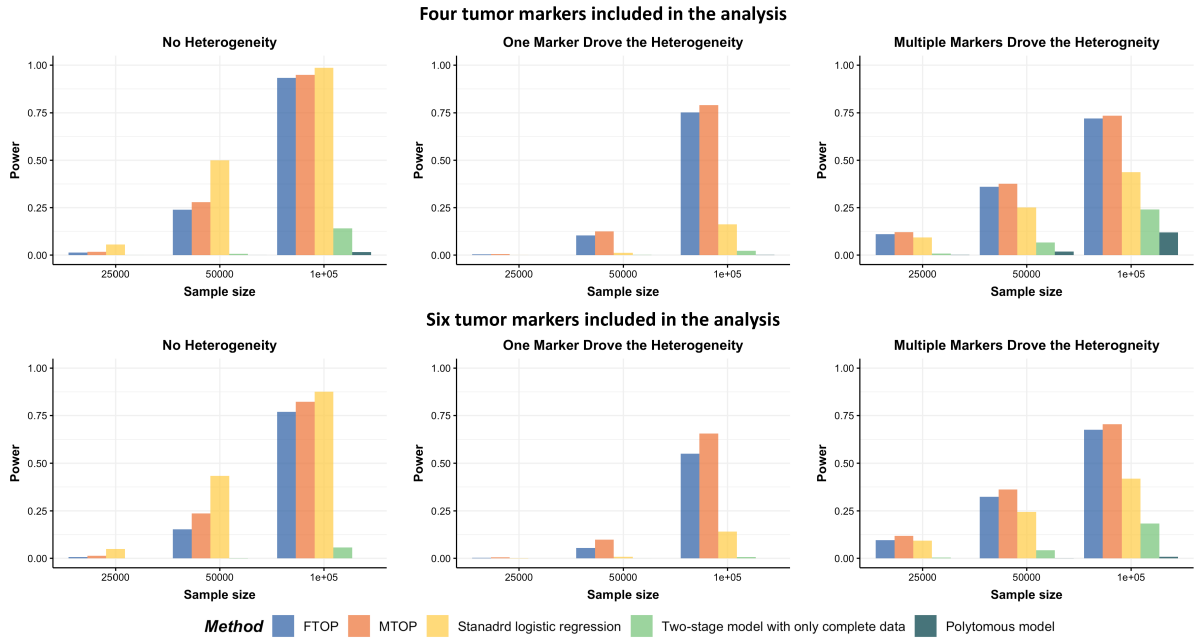

FIG. 1. Power comparison among MTOP, FTOP, standard logistic regression, two-stage model with only complete data and polytomous model with 2×10^5 random samples. For the three figures in the first row, four tumor markers were included in the analysis. Three binary tumor marker and one ordinal tumor marker defined 24 cancer subtypes. Around 70% cases would be incomplete. For the three figures in the second row, two extra binary tumor markers were included in the analysis. The six tumor markers defined 96 subtypes. Around 77% cases would be incomplete. The power was estimated by controlling the type one error $\alpha < 5.0 \times 10^{-8}$.

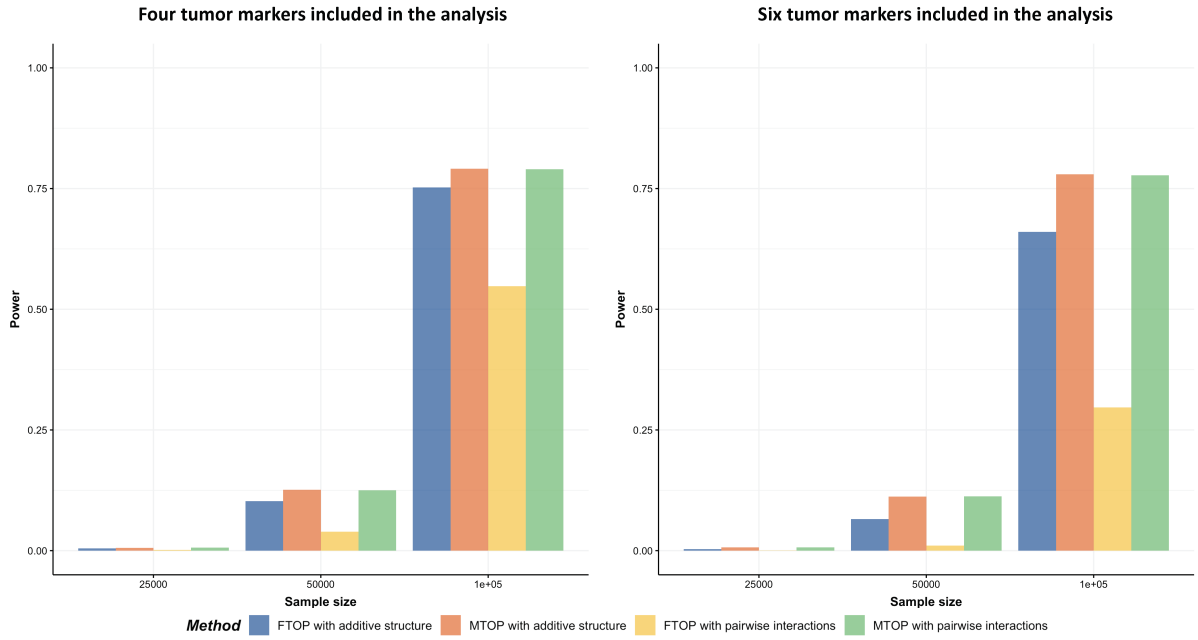

FIG. 2. Power comparison of global association test with pairwise interactions. Four methods were evaluated, including FTOP with additive structure, MTOP with additive structure (ER fixed), FTOP with pairwise interactions and MTOP with pairwise interactions (ER fixed). For the three figures in the first row, four tumor markers were included in the analysis. Three binary tumor marker and one ordinal tumor marker defined 24 cancer subtypes. Around 70% cases were incomplete. For the three figures in the second row, two extra binary tumor markers were included in the analysis. The six tumor markers defined 96 subtypes. Around 77% cases were incomplete. The total sample size was 25,000, 50,000 and 100,000. We generated 2×10^5 random replicates. The power was estimated by controlling the type one error $\alpha < 5.0 \times 10^{-8}$.

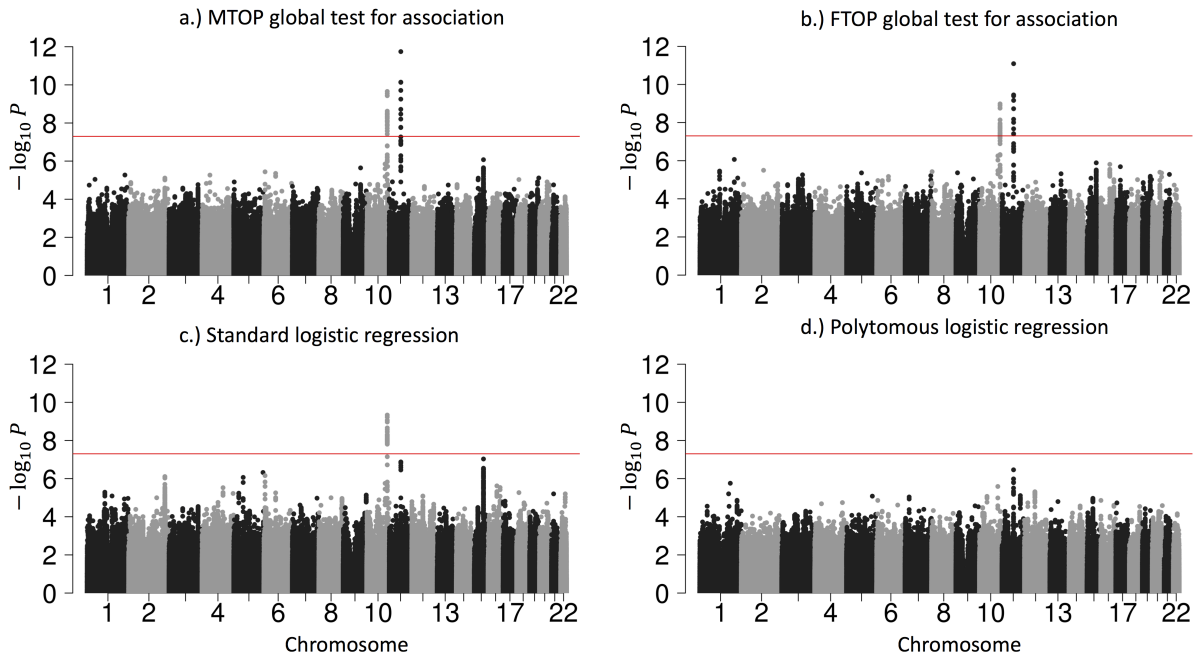

FIG. 3. Manhattan plot of genome-wide association analysis with PBCS using four different methods. PBCS have 2,078 invasive breast cancer and 2,219 controls. In total, 7,017,694 SNPs on 22 auto chromosomes with MAF more than 5% were included in the analysis. ER, PR, HER2 and grade were used to define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