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: 5 February 2020)

ABSTRACT: Cancers are routinely classified into subtypes according to various features, including histopathological characteristics and molecular markers. Previous genome-wide association studies have reported heterogeneous associations 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 to test for genetic associations using a mixed-effect two-stage polytomous model score test (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. 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. The proposed methods have been implemented in a user-friendly and high-speed R statistical package called TOP (https://github.com/andrewhaoyu/TOP).

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), providing tremendous opportunities to investigate for potential distinct etiological pathways between cancer subtypes. For example, a breast cancer ER-negative specific GWAS reported 20 SNPs that were more strongly associated with the risk of developing ER-negative than ER-positive disease (Milne *and others*, 2017). Previous studies also suggested traditional breast cancer risk factors, such as age, obesity, and hormone therapy use, were heterogeneously associated with the risk of breast cancer subtypes (Barnard *and others*, 2015).

The most common procedure for testing for associations between risk factors and cancer subtypes is by 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, leading to many cancer cases with no subtype definition, and often these cases are dropped from the model. Second, the tumor markers that defined the subtypes are commonly highly correlated with each other. Testing each subtype separately without modeling the correlation limits the power of the model. 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 according to 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 into a case-control odds ratio for a reference subtype, a case-case odds ratio for each tumor characteristic, and higher-order interactions between the 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 case-case odds ratios can be interpreted as the measures of etiological heterogeneity for tumor characteristics.

Although the two-stage model can improve the power compared to fitting standard logistic regressions for each subtype (Chatterjee, 2004; Zabor and Begg, 2017), the two-stage model does have notable limitations and has not been widely applied to analyze data on multiple tumor characteristics. First, similar to standard logistic regression, 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 the two-stage model can reduce the multiple testing burdens compared to traditional methods, as the number of tumor characteristics increases, the two-stage model 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 that incorporate 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 the 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 **??**, we describe the proposed three different hypothesis tests, the missing data algorithm, and the score tests. In Section III, we present the simulation results for type I error, power and computation time. In Section IV, the proposed methods are illustrated with applications using data from the Polish Breast Cancer

Study (PBCS). In Section V, we discuss the strengths and limitations of the methods and future research directions.

II. TWO-STAGE POLYTOMOUS LOGISTIC MODEL

The details of the two-stage polytomous logistic model have been described earlier (Chatterjee, 2004). We briefly summarize them for completeness. Suppose a disease can be classified using K disease characteristics, and each characteristic k can be classified into M_k categories; thus, 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 defined as either positive or negative.

Let D_i denote the disease status of subject i in the study such that $D_i \in \{0, 1, 2, \cdots, M\}$ and $i \in \{1, \cdots, N\}$. $D_i = 0$ 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, where P is the total number of other covariates. In the first stage model, a "saturated" polytomous logistic regression model is constructed as follows:

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

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

Because each cancer subtype is defined through a unique combination of the K tumor 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, the log odds ratios in the first stage can be represented as follows:

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

where $\theta^{(0)}$ represents the case-control log odds ratio for a reference disease subtype, $\theta^{(1)}_{k_1}$ represents the main effect of k_1th tumor characteristic, $\theta^{(2)}_{k_1 k_2}$ represents the second order interaction between k_1th and k_2th tumor characteristics , and so on. A reference level can be defined for each tumor characteristic, and the reference disease subtype is jointly 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, and $\boldsymbol{\theta}_{\mathrm{H}}$ 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 $\boldsymbol{\theta} = (\theta^{(0)}, \boldsymbol{\theta}_{\mathrm{H}}^T)^T$ as 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 models.

Up to now, we have only described second stage decomposition for the regression coefficients of $\mathbf{G}$. The second stage decomposition can also be applied to the other covariates, the details of which are 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.

A. Hypothesis test under two-stage model

The first stage case-control log odds ratios of subtypes can be decomposed 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 for 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 $\mathrm{H}_1^{\mathrm{A}} : \boldsymbol{\theta} \neq \mathbf{0}$, which 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 can 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 simultaneously evaluates the etiologic heterogeneity with respect to a SNP and all the tumor characteristics. Rejecting this null hypothesis indicates that the first

stage case-control log odds ratios are significantly different between 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, H_0^{IH} : $\theta_{H(k)} = 0$ versus H_1^{IH} : $\theta_{H(k)} \neq 0$, where $\theta_{H(k)}$ is one of the case-case parameters of $\boldsymbol{\theta}_H$. The case-case parameter ($\theta_{H(k)}$) provides a measurement of etiological heterogeneity according to a specific tumor characteristic (Begg and Zhang, 1994). In the breast cancer example, we can directly test $H_0^{IH} : \theta_{ER}^{(1)} = 0$ versus $H_1^{IH} : \theta_{ER}^{(1)} \neq 0$. Rejecting the null hypothesis provides evidence that the case-control log odds ratios of ER+ and ER- subtypes are significantly different.

B. EM algorithm accounting for cases with incomplete tumor characteristics

In the previous sections, all the tumor characteristics were assumed to have no missing data. However, 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, while 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 the ith subject is disease subtype m. Given $\mathbf{T}_{io}$, the possible subtypes for subject i, denoted

as $\mathcal{Y}_{io} = \{Y_{im} : Y_{im}$ that is consistent with $\mathbf{T}_{io}\}$, are within a limited subset of all possible tumor subtypes. 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 (MAR). 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, the E step of them EM algorithm at the vth iteration is

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

where Y_{im}^{E} is the probability of the ith person to be the mth subtype given his observed tumor characteristics $(\mathbf{T}_{io})$, genotype (G_i), and other covariates $(\mathbf{X}_i)$. $I(Y_{im} \in \mathcal{Y}_{io})$ denotes whether the mth subtype for the ith subject belong to the subsets of possible subtypes given the observed tumor characteristics. The M step at the vth iteration is

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

The M step can 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 $\mathbf{W} = \mathbf{D} - \mathbf{A}\mathbf{A}^T$, $\mathbf{D} = \mathrm{diag}(\mathbf{P})$, $\mathbf{P} = E(\mathbf{Y}|\mathbf{C}; \delta)$, 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 $\mathbf{W}^{(t)}$ are respectively defined as $\mathbf{P}$ and $\mathbf{W}$ evaluated at the $\boldsymbol{\delta}^{(t)}$. The weighted least square update is $\boldsymbol{\delta}^{(t+1)} = (\mathbf{Z}^T \mathbf{C}_M^T \mathbf{W}^{(t)} \mathbf{C}_M \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{C}_M^T \mathbf{Y}^{*(t)}$. As $t \to \infty$, the weighted least square interaction converges to $\hat{\boldsymbol{\delta}}^{(v+1)}$, which will be used in next iteration. The EM algorithm will converge to the MLE of the second stage parameters (denoted as $\hat{\boldsymbol{\delta}}$), and the observed information matrix $\mathbf{I}$ is $\mathbf{I} = \mathbf{Z}^T \mathbf{C}_M^T (\mathbf{W} - \mathbf{W}_{\mathrm{mis}}) \mathbf{C}_M^T \mathbf{Z}$, where $\mathbf{W}_{\mathrm{mis}} = \mathbf{D}_{\mathrm{mis}} - \mathbf{A}_{\mathrm{mis}} \mathbf{A}_{\mathrm{mis}}^T$, $\mathbf{D}_{\mathrm{mis}} = \mathrm{diag}(\mathbf{P}_{\mathrm{mis}})$,

$\mathbf{P}_{\text{mis}} = E(\mathbf{Y}|\mathbf{C}, \mathbf{T}_o; \delta)$, and $\mathbf{A}_{\text{mis}} = \mathbf{D}_{\text{mis}}(\mathbf{1}_M \otimes \mathbf{I}_N)$ (Louis, 1982). More details of the EM algorithm are in Supplementary Section 2.

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

C. 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 to be fixed. The score test only needs to estimate the second stage parameters of $\mathbf{X}$ under the null hypothesis once, making it much more computationally efficient than the Wald test. Moreover, the EM algorithm only needs to be implemented once under the null hypothesis. Since we don't perform any second stage decomposition on the intercept parameters in the 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 is due to the high correlation between the tumor markers. In the breast cancer example, the correlation between ER and PR is 0.63, between ER and HER2 is -0.16, and between PR and HER2 is -0.17 (Supplementary Table 1). 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

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

FTOP has the same degrees of freedoms and similar asymptotic power (Yi and Wang, 2011) as the Wald test. In GWAS which needs to perform millions of tests, FTOP can be first used to scan the whole genome with global association test, and then select the potential risk regions. In the selected risk regions, each SNP can be tested for global heterogeneity and individual tumor characteristic heterogeneity using Wald test.

D. 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 random effects. 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 as follows:

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

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-effect parameters $\boldsymbol{\theta}_{\mathrm{fH}}$ can be used for tumor characters with prior information suggesting that they are a source of heterogeneity, and the random-effect parameters $\mathbf{u}$ can model tumor characteristics with little or no prior information. In 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 of breast cancer heterogeneity (García-Closas *and others*, 2013; Milne *and others*, 2017). The main effects of PR and HER2 and other potential interactions effects can be modeled as random effects ($\mathbf{u}$). In 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 two independent scores can be independent. First for $\boldsymbol{\theta}_{\mathrm{f}}$, the score test statistic $Q_{\boldsymbol{\theta}_{\mathrm{f}}}$ is derived under the global null hypothesis $\mathrm{H}_0^{\mathrm{A}} : \boldsymbol{\theta}_{\mathrm{f}} = \mathbf{0}$, $\sigma^2 = 0$ as usual. But for σ^2, the the score statistic Q_{σ^2} is derived 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 to be independent (Supplementary Section 4), and the Fisher's procedure (Koziol and Perlman, 1978) can be used to combine the p-value generated from the two independent tests. Similarly to 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 the 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 $\mathbf{Z}_{\mathbf{G}}$ with $\mathbf{Z}_{\mathrm{f}}$. Under the null hypothesis, $Q_{\boldsymbol{\theta}_{\mathrm{f}}}$ follows a χ^2 distribution with the degrees of freedom l_{f} the same as the length of $\boldsymbol{\theta}_{\mathrm{f}}$.

To explicitly express 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 as follows:

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

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 (Supplementary Section 3), where $\chi^2_{i,1}$ i.i.d. follows χ^2_1. $(\rho_1, \ldots, \rho_s)$ are the eigenvalues of $\tilde{\mathbf{I}}_\mathrm{r} = \mathbf{I}_{\mathbf{uu}} - \mathbf{I}^T_{\mathbf{u}\boldsymbol{\tau}} \mathbf{I}^{-1}_{\boldsymbol{\tau}\boldsymbol{\tau}} \mathbf{I}_{\boldsymbol{\tau}\mathbf{u}}$, 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 $\mathbf{I}_{\boldsymbol{\tau}u} = \mathbf{I}^T_{u\boldsymbol{\tau}} = \mathbf{Z}_{\boldsymbol{\tau}}^T \mathbf{C}_M^T \mathbf{W}_\mathrm{r} \mathbf{G}_M \mathbf{Z}_r$, 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; thus, 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 of the global etiology heterogeneity test, $\mathrm{H}_0^\mathrm{EH} : \boldsymbol{\theta}_\mathrm{fH} = \mathbf{0}$, $\sigma^2 = 0$, can be computed following a similar procedure as the global association test.

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 computation time (Supplementary Section 5) of the fixed-effect and mixed-effect two-stage models. Data

were simulated to mimic the PBCS. We simulated four tumor characteristics: ER (positive vs. negative), PR (positive vs. negative), HER2 (positive vs. negative), and grade (ordinal 1, 2, 3), which collectively defined $2^3 \times 3 = 24$ breast cancer subtypes.

In each simulation, genotype data $\mathbf{G}$ was simulated under the Hardy-Weinberg equilibrium with minor allele frequency (MAF) as 0.25. An additional covariate ($\mathbf{X}$) was simulated following a standard normal distribution independent of $\mathbf{G}$. We simulated a multinomial outcome with 25 groups, one for the control group, and the other 24 for different cancer subtypes, using the polytomous logistic regression model as follows:

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

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 2) (Michailidou *and others*, 2017), we computed the corresponding polytomous logistic regression intercept parameters α_m. The case-control ratio was set around 1:1, and the proportions of ER+, PR+ and HER2+ were 0.81, 0.68, and 0.17, respectively. 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, approximately 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. The data were

generated by setting $\beta_m = 0$ in Equation 11, where none of the subtypes was 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 $\alpha = 1.0 \times 10^{-4}, 1.0 \times \times 10^{-5}$, and 1.0×10^{-6} level.

Both MTOP and FTOP were applied with an additive two-stage model by constraining all the interaction terms as 0 in Equation 2. 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. Furthermore, the MTOP assumed the baseline and ER case-case parameter as fixed effects and the other 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. For MTOP and FTOP, we removed all the subtypes with fewer than 10 cases to avoid potential nonconvergence of the model.

Table I presents the estimated type I errors under the global null hypothesis. Both MTOP and FTOP correctly control the type I error, especially for the larger sample sizes. FTOP is conservative with 5,000 subjects, especially for $\alpha = 1.0 \times 10^{-6}$, however, the method is still valid. The well-controlled type I error also shows that removing rare subtypes doesn't bias the estimate, as further demonstrated by additional simulations that are presented in Supplementary Section 6. In the later sections, we generally used the additive second stage structure for both MTOP and FTOP unless otherwise specified.

B. Statistical power

We assessed the statistical power of the proposed methods using various simulation settings with sample sizes as 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

The data were simulated 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. The disease subtypes were generated 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 two-stage model. Under scenarios II, datasets were simulated with only ER heterogeneity by setting the case-case parameter for ER 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 to be 0, the ER case-case parameter to be 0.08, and all the other case-case parameters following a normal distribution with mean 0 and variance 4.0×10^{-4}. Under this scenario, all tumor characteristics contributed to the subtype-specific heterogeneity. Moreover, to evaluate different methods under a larger number of tumor characteristics, additional simulations were conducted by adding two additional binary tumor characteristics to the previous four tumor characteristic setting. This defined $2^5 \times 3 = 96$ 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 characteristic missing. We compared the statistical power to detect the overall association using 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 the degrees of freedom between MTOP and FTOP was small. However, with six tumor markers, the power gain of MTOP was more apparent owing to the larger difference in the degrees of freedom between the models. FTOP was the least efficient in scenarios with no or little heterogeneity, such as scenarios I and II, but with increasing heterogeneity, such-as scenario III, the power of MTOP and FTOP were more similar.

The simulation study also showed that the incorporation of cases with missing tumor characteristics significantly increased the power of the methods (Figure 1). Under the four tumor markers setting with around 70% incomplete cases, the power gain of FTOP incorporating the missing data algorithm was at least 200% compared to FTOP with only complete data. As expected, under the six tumor markers setting, which resulted in more missing

tumor marker data, the power of FTOP with the missing data algorithm was once again significantly higher than FTOP with only complete data. MTOP was the most powerful method when heterogeneity across cancer subtypes was present. Additional power simulations with 5,000 subjects are described in Supplementary Section 7.

The previous simulations mainly focused on the 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 that included main effects and pairwise interactions as shown in Equation 2 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 and 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 because of the incorporation of the interaction terms as fixed effects. On the other hand, MTOP with/without pairwise interactions maintained a high power even when there were underlying interaction effects.

2. Global heterogeneity test

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

3. Individual tumor marker heterogeneity test

We further evaluated the power of the individual tumor marker heterogeneity test. The data were 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 a rate of 0.17, 0.30 and 0.50. We compared two different methods, FTOP with all four tumor characteristics and the polytomous model. The polytomous model was set up to test each marker at a time. In the polytomous model, we removed cases with missing data only on the relevant tumor marker to avoid penalizing the power of the model by removing cases that were missing tumor marker data on the other tumor markers. FTOP with all four tumor characteristics had smaller power compared to the polytomous model in testing the effect of ER (Supplementary Figure 4). Since FTOP included all four tumor characteristics, and the tumor markers were highly correlated, the variability of underlying parameters was larger. However, the type I errors of the polytomous model in testing PR, HER2 and grade were inflated under this case (Supplementary Figure 5). Under this simulation, these three markers had no effect. On the other hand, FTOP controlled the type I error of all the tests.

Overall, for the global test for association and the 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, and MTOP had stable power even with a large number of pairwise interactions terms included.

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 genotyping data to compare MTOP, FTOP, standard logistic regression, and polytomous logistic regression to detect SNPs associated with breast cancer risk.

Supplementary Table 4 presents the sample size 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 the analysis. Both MTOP and FTOP used the additive second stage design. Besides, we modeled the baseline and ER case-case parameters 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). 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 was 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 detected FGFR2 and CCND1 with the 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, $P = 1.0 \times 10^{-9}$ vs. $P = 2.2 \times 10^{-10}$; for rs78540526, $P = 1.7 \times 10^{-11}$ vs. $P = 1.8 \times 10^{-12}$). In both scenarios with pairwise interactions parameter included, 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, correlations between markers, and missing tumor data. Under the model framework, we develop two computationally efficient score tests, FTOP and MTOP, which model 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 models haven't accounted for

missing tumor markers, which is 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. In a recent large breast cancer GWAS analysis with 106,571 cases and 95,762 controls, the newly developed methods MTOP and FTOP have identified 16 novel loci (Zhang *and others*, 2019).

Incorporating missing tumor characteristics based on the proposed EM algorithm requires the assumption of MAR, i.e. the mechanism of missing of the individual tumor characteristics can depend only on other observed tumor characteristics and covariates, but not on the unobserved missing value themselves. For the analysis of tumor heterogeneity, information on aggressive types of tumors may be systematically missing. If the missing tumor characteristics are important determinants of aggressiveness, then the underlying assumption is violated. In general, dealing with non-ignorable missing data is a complex problem and certain sensitivity analyses can 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 the missingness mechanism itself is related to the genetic variant. 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 $\mathbf{X}$. For MTOP, the computational complexity is $O(NM^2P^2lk)$, where l and k are respectively the numbers of iteration required for weighted least square and EM algorithm to converge.

Currently, we only implement the linear kernel in MTOP, but other common kernels that capture the similarity between tumor characteristics can be used in the future. If there is prior knowledge about the overlapping genetic architecture across different tumor subtypes, this will 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, the two-stage model is generalized to multivariates. In Supplementary Section 2-3, the details of the EM algorithm and the variance component score statistic are respectively presented. In Supplementary Section 4, $Q_{\boldsymbol{\theta}_f}$ and Q_{σ^2} are proved to be independent. In Supplementary Section 5, computation time simulations are presented. In Supplementary Section 6, the simulations to evaluate the bias of the estimates are shown. In Supplementary Section 7, simulations with 5,000 subjects are presented. Supplementary Table 1 shows the correlations of ER, PR, HER2, and grade. Supplementary Table 2 presents the frequencies of the joint distribution of ER, PR, HER2, and grade. Supplementary Table 3 shows the simulation results to evaluate bias. Supplementary Table 4 presents the sample size of tumor characteristics in PBCS. Supplementary Figure 1 shows the computation

time simulations results. Supplementary Figure 2 presents the power analysis of the global association test with 5,000 subjects. Supplementary Figure 3 presents global heterogeneity test simulation results. Supplementary 4-5 respectively present the power and type I error simulations results of individual tumor marker heterogeneity test. Supplementary Figure 6 is the QQ plot of GWAS with PBCS. Supplementary Figure 7 shows the GWAS with PBCS using MTOP with ER and grade as fixed effects. Supplementary Figure 8-9 respectively present the GWAS with PBCS using MTOP/FTOP with pairwise interactions.

ACKNOWLEDGEMENTS

This work was supported by funds from the NCI Intramural Research Program, Bloomberg Distinguished Professorship endowment, and NHGRI (1R01 HG010480-01). 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.

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

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

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

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

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

565 Network, Cancer Genome Atlas. (2012). Comprehensive molecular portraits of hu-
566 man breast tumours. *Nature* **490**(7418), 61–70.

567 Perou, C. M. *and others*. (2000). Molecular portraits of human breast tumours. *Na-*
568 *ture* **406**(6797), 747–52.

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

571 Rosner, B. *and others*. (2013). Breast cancer risk prediction with heterogeneous risk pro-
572 files according to breast cancer tumor markers. *American Journal of Epidemiology* **178**(2),
573 296–308.

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

576 Wang, M., Kuchiba, A. and Ogino, S. (2015). A meta-regression method for studying
577 etiological heterogeneity across disease subtypes classified by multiple biomarkers. *Amer-*
578 *ican journal of epidemiology* **182**(3), 263–270.

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

581 Yang, X. R. *and others*. (2007). Differences in risk factors for breast cancer molecular sub-
582 types in a population-based study. *Cancer Epidemiology and Prevention Biomarkers* **16**(3),
583 439–443.

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

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

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

590 Zhang, H. *and others*. (2019). Genome-wide association study identifies 32 novel breast
591 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.

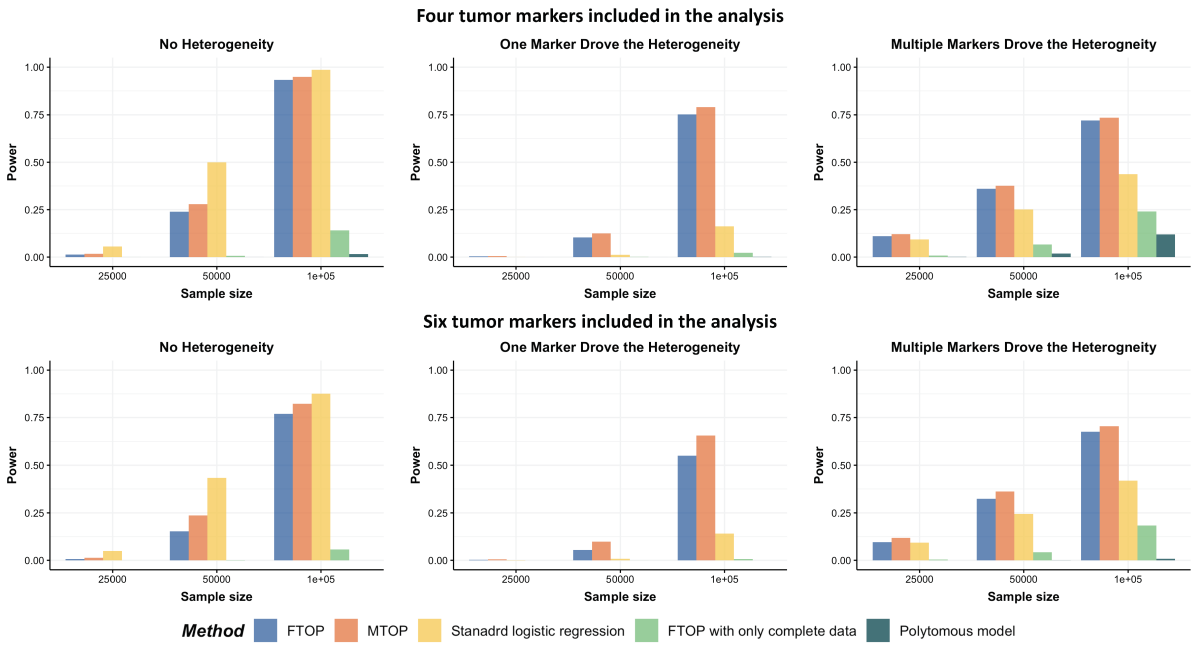

FIG. 1. Power comparison among MTOP, FTOP, standard logistic regression, two-stage model with only complete data and polytomous model with 2×10^5 random samples. For the three figures in the first row, four tumor markers were included in the analysis. Three binary tumor marker and one ordinal tumor marker 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 $\alpha < 5.0 \times 10^{-8}$.

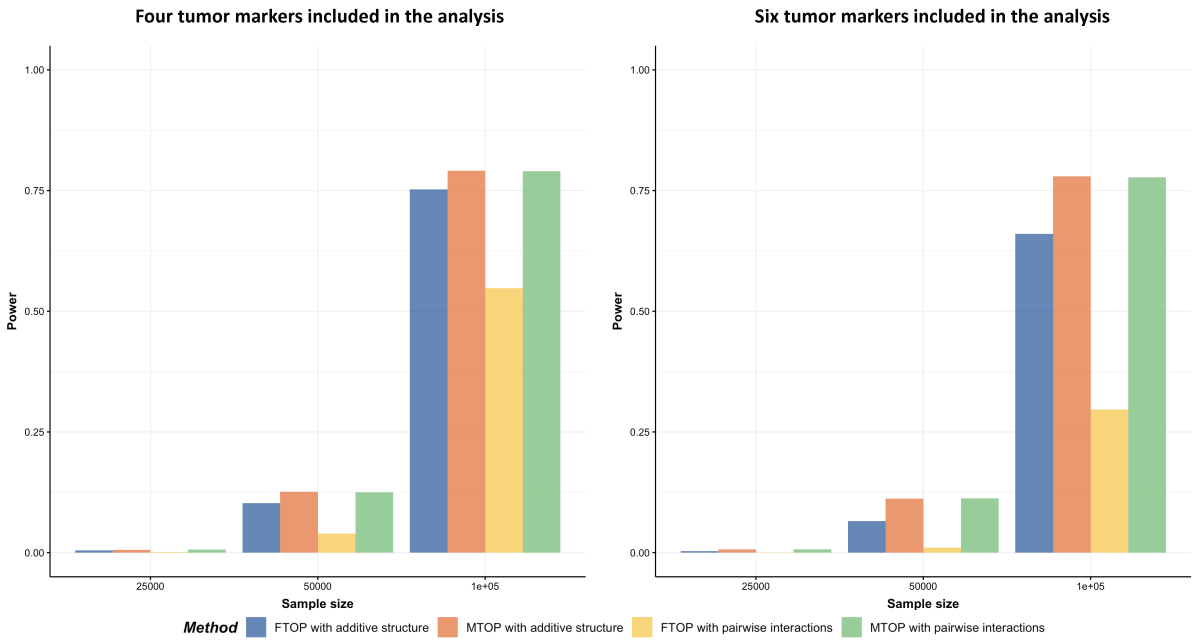

FIG. 2. Power comparison of global association test with pairwise interactions. Four methods were evaluated, including FTOP with additive structure, MTOP with additive structure (ER fixed), FTOP with pairwise interactions and MTOP with pairwise interactions (ER fixed). For the three figures in the first 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 $\alpha < 5.0 \times 10^{-8}$.

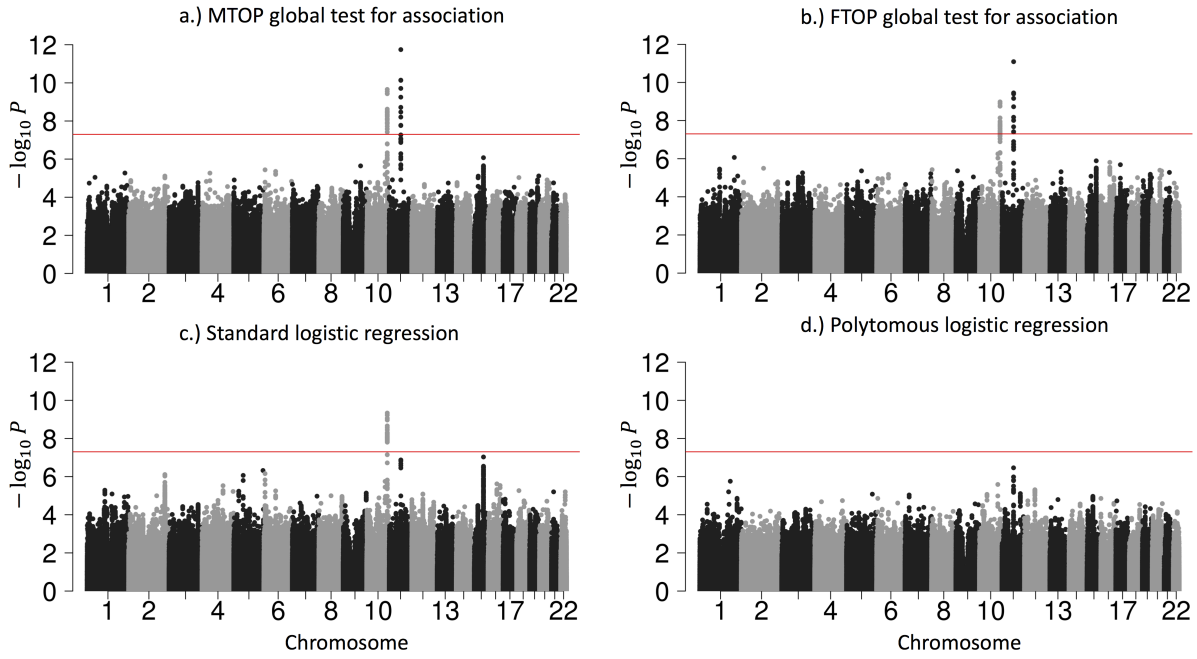

FIG. 3. Manhattan plot of genome-wide association analysis with PBCS using four different methods. PBCS have 2,078 invasive breast cancer and 2,219 controls. In total, 7,017,694 SNPs on 22 auto chromosomes with MAF more than 5% were included in the analysis. ER, PR, HER2 and grade were used to define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