Published in final edited form as:

Biometrics. 2018 March ; 74(1): 176–184. doi:10.1111/biom.12745.

Multivariate Association Analysis with Somatic Mutation Data

Qianchuan He, Yang Liu, Ulrike Peters, and Li Hsu

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, U.S.A

Summary

Somatic mutations are the driving forces for tumor development, and recent advances in cancer genome sequencing have made it feasible to evaluate the association between somatic mutations and cancer-related traits in large sample sizes. However, despite increasingly large sample sizes, it remains challenging to conduct statistical analysis for somatic mutations, because the vast majority of somatic mutations occur at very low frequencies. Furthermore, cancer is a complex disease and it is often accompanied by multiple traits that reflect various aspects of cancer; how to combine the information of these traits to identify important somatic mutations poses additional challenges. In this article, we introduce a statistical approach, named as SOMAT, for detecting somatic mutations associated with multiple cancer-related traits. Our approach provides a flexible framework for analyzing continuous, binary, or a mixture of both types of traits, and is statistically powerful and computationally efficient. In addition, we propose a data-adaptive procedure, which is grid-search free, for effectively combining test statistics to enhance statistical power. We conduct an extensive study and show that the proposed approach maintains correct type I error and is more powerful than existing approaches under the scenarios considered. We also apply our approach to an exome-sequencing study of liver tumor for illustration.

Keywords

Association test; Mixture of traits; Multivariate traits; Somatic mutations; SOMAT

1. Introduction

Somatic mutations play a fundamental role in the development and progression of cancer (Luzzatto, 2011). Recent progress in high-throughput genome technologies has enabled sequencing of cancer genomes in hundreds of patients (Watson et al., 2013), making it possible to comprehensively evaluate the health impact of somatic mutations at a large scale. The study of somatic mutations with respect to cancer-related traits provides novel insight into tumor biology, offers new paths to drug development, and potentially leads to better clinical treatment (Yu, O'Toole, and Trent, 2015).

Correspondence to: Qianchuan He; Li Hsu.

6. Supplementary Materials

Web Appendix, Tables, and Figures referenced in Sections 2–4 are available with this paper at the Biometrics website on Wiley Online Library.

A major challenge in analyzing somatic mutation data lies in the low frequency of somatic mutations. The median non-silent mutation rate in protein-coding regions is estimated to be 1.5/Mb (Lawrence et al., 2013); for a typical gene with 1000bp of coding sequence, this translates to only 1.5×10^{-3} non-silent mutations per gene. Besides the low mutation rate, somatic mutations are often distributed across many different sites of the gene, hence the frequency of somatic mutation at each individual site can be extremely low; for this reason, it is nearly impossible to conduct statistical inference for a single mutation site. A common strategy to alleviate this problem is to summarize somatic mutations within a gene into a single variable by for example, presence or absence of mutations (Rachakonda et al., 2013), or the total number of mutations (Liu et al., 2015). These gene-level mutations are then further aggregated by gene-sets or pathways and evaluated together in association with outcomes. For a single trait, existing approaches that target set-based low-frequency variants association can be potentially used, such as the Sequence Kernel Association Test (SKAT) (Wu et al., 2011), the SKAT-O test (Lee, Wu, and Lin, 2012), and the Mixed effects Score Test (MiST) (Sun, Zheng, and Hsu, 2013). These approaches are closely connected with the random effects model. The SKAT assumes that the effect of each genetic variant follows an arbitrary distribution with mean 0 and variance τ^2 , while the SKAT-O seeks to combine burden test and SKAT for higher power. The MiST approach allows the mean of effects to incorporate functional features and shift from 0, which can gain statistical power when the genetic variants have a nonzero effect on average.

However, cancer is a complex disease and is often accompanied by many clinically relevant traits. For example, in liver cancer studies, α -fetoprotein and the prothrombin time are usually measured because they are associated with patients' prognosis (Ikeda et al., 1998; Lai et al., 1995). Leveraging these traits to identify disease-related somatic mutations can yield a more complete picture of the underlying genetic structure, and lead to better understanding of the etiology of cancer. Moreover, jointly analyzing multiple related traits can potentially increase statistical power, which is particularly important in somatic mutation studies because these mutations are typically very rare.

Statistical approaches for analyzing multivariate traits in genetic studies have been proposed in the literature (Ferreira and Purcell, 2009; Yang et al., 2010; He, Avery, and Lin, 2013), but these approaches are primarily designed to handle variants with higher allele frequencies. Maity, Sullivan, and Tzeng (2012) have developed an approach for analyzing less common variants. This approach accounts for the correlations among the traits by assuming a multivariate normal distribution for the residuals, and its p-values are obtained through a data resampling procedure which can be computation-intensive for large sample sizes. Recently, Sun et al. (2016) also introduced an approach for multivariate trait analysis. This approach assumes that for each genetic variant, the associated regression coefficients with K traits have a compound-symmetry correlation structure; that is, the pleiotropic effects of a given variant are equally correlated. The approaches of Maity et al. (2012) and Sun et al. (2016) are focused on continuous traits only, while in many cancer studies, the relevant traits may be binary or continuous. In addition, these approaches assume that the genetic variants have an average effect of 0, which may not always hold because many somatic mutations are believed to have non-null effects on cancer pathogenesis (Watson et al., 2013). In this article, we propose an approach for investigating the associations between somatic mutations and

multiple traits. The proposed approach is able to handle continuous, binary, or a mixture of both types of traits. We allow non-zero mean effects of somatic mutations to accommodate the potential concerted effect of somatic mutations within a pathway, and the proposed approach can provide component-wise p-values for the mean effect and the random effect separately. Furthermore, we develop a highly efficient procedure for combining two test statistics to enhance power; this procedure yields numerically tractable solutions and does not require grid-search to find the optimal combination. The rest of the paper is organized as follows. We describe the methods in the Methods section, and assess the finite sample performance of the proposed approach in the Simulation Study section. We illustrate the method in the Real Data Analysis section. Finally, the paper is concluded with a few final remarks.

2. Methods

In this section, we propose a statistical approach, SOMAT (somatic mutation association test), for evaluating genetic associations with somatic mutations for multiple traits. We consider both continuous and binary traits, as well as a mixture of both types.

2.1 Model Setup

Consider a study that consists of n independent subjects and that each subject has K clinically relevant traits. For the i th subject ($i = 1, \dots, n$), let Y_{ki} be the k th trait ($k = 1, \dots, K$), which can be either binary or continuous, and X_i be the d -dimensional covariates (including the intercept). Further, let G_i ($i = 1, \dots, n$) be the somatic mutations at p genes in a given gene set or pathway. G_i can be presence or absence of mutation, number of mutations, or some other codings.

It is natural to assume a generalized linear model (GLM) for an individual trait Y_{ki} with mean $\mu_k(\alpha_k^T X_i + \beta_k^T G_i)$ and dispersion parameter ϕ_k ; here, $\mu_k(\cdot)$ is a mean function, and $\alpha_k \equiv (\alpha_{k1}, \dots, \alpha_{kd})^T$ and $\beta_k \equiv (\beta_{k1}, \dots, \beta_{kp})^T$ are regression parameters. For continuous traits, we use the identity function $\mu_k(x) = x$, and for dichotomous traits, we adopt the expit function $e^x/(1 + e^x)$. To test the association of somatic mutations with multivariate traits is to test $H_0: \beta_k = 0, k = 1, \dots, K$. This amounts to test a total of $p \times K$ parameters. Hence, even if a gene set contains only a moderate number of genes, the number of parameters can be very large. To reduce the number of parameters and enhance power for detecting the association, we propose to leverage the functional information about these genes by further specifying β_{kj} as

$$\beta_{kj} = \theta_k^T W_j + \xi_{kj}, \quad (1)$$

for $j = 1, \dots, p$, where θ_k is a q -vector of regression coefficients, W_j is a q -vector of known characteristics for the j th somatic mutation, and ξ_{kj} is the mutation-specific effect. We assume that ξ_{kj} is a random effect following a distribution with mean 0 and variance τ_k^2 . W_j can be an indicator of whether the mutation is synonymous or not, or simply just an

intercept. If W_j is one for all variants, then θ_k quantifies the average effect of mutations in the gene set. Under this scenario, the number of parameters to be tested for each trait is only 2, one for the average effect θ_k and the other for the variance component τ_k^2 , which is typically fewer than p parameters associated with p genes.

Let $W = (W_1, \dots, W_p)$ be the $q \times p$ characteristics matrix, and $\xi_k = (\xi_{k1}, \dots, \xi_{kp})^T$. Plugging in (1) into the mean function yields

$$\mu_k^{-1}\{E(Y_{ki})\} = \alpha_k^T X_i + \theta_k^T (WG_i) + \xi_k^T G_i,$$

for which our main interest is to test the nullity of θ_k and τ_k^2 , for $k = 1, \dots, K$. It is worth noting that we do not impose a parametric modeling for the correlation among the traits, and will estimate their correlation empirically to account for the dependence in our testing procedure (to be shown later). This framework is broad and includes existing methods for genetic association testing. For example, when $\theta_k^T W$ is replaced by a $p \times 1$ regression parameter vector γ_k and additionally $\xi_k = 0$, this model becomes a multi-variant version of the model by He et al. (2013). When $K = 1$, this model reduces to the model of Sun et al. (2013).

For a single trait, say the k th trait, under the null hypothesis that $H_{01} : \theta_k = 0$ and $\tau_k^2 = 0$, the score statistic for θ_k can be written as

$$U_k^{(\theta)} = \tilde{\phi}_k^{-1} \sum_{i=1}^n \{Y_{ki} - \mu_k(\tilde{\alpha}_k^T X_i)\} WG_i,$$

where $\tilde{\alpha}_k$ and $\tilde{\phi}_k$ are the GLM estimators of α_k and ϕ_k under H_{01} , respectively. Here, $\tilde{\phi}_k = \sum_{i=1}^n (Y_{ki} - \tilde{\alpha}_k^T X_i)^2 / (n - d)$ for a continuous trait, and $\tilde{\phi}_k = 1$ for a binary trait. We can also derive score test statistic for τ_k^2 under H_{01} ; however, under this situation, the score statistics for θ_k and τ_k^2 are correlated, and there is no explicit form for the covariance due to different distributions for these two score statistics. As a result, it is difficult to derive an efficient combination of the two score statistics for θ_k and τ_k^2 that may improve the power. Instead, we derive the score statistic for τ_k^2 under $H_{02} : \tau_k^2 = 0$, for which the score statistic can be written as

$$U_k^{(\tau)} = \hat{\phi}_k^{-2} \sum_{i=1}^n \{Y_{ki} - \mu_k(\hat{\alpha}_k^T X_i + \hat{\theta}_k^T WG_i)\} G_i^T \times \sum_{i=1}^n \{Y_{ki} - \mu_k(\hat{\alpha}_k^T X_i + \hat{\theta}_k^T WG_i)\} G_i,$$

where $\hat{\phi}_k$, $\hat{\alpha}_k$ and $\hat{\theta}_k$ are the estimators of ϕ_k , α_k and θ_k under H_{02} , respectively. It has been shown that $U_k^{(\theta)}$ and $U_k^{(\tau)}$ are independent for a given k (Sun et al., 2013), hence one can harness existing well-known procedures, such as Fisher's or Tippett's combinations, for combining p-values from two independent statistics; we also develop an optimal linear combination procedure which relies upon independent statistics (to be shown). Furthermore, the two statistics allow the dissection of the random effects from the fixed effects, which provides valuable biological insight for understanding source of the association for the genetic effects of the mutations.

For multiple traits, we wish to test the hypothesis that $H_0: \theta_k = 0, \tau_k^2 = 0$ for all K traits simultaneously. In the next section, we discuss how to conduct the joint test for the situation where only continuous traits are present and the situation where binary traits are also involved.

2.2 Multivariate Continuous Traits

To conduct the multivariate testing, we first need to characterize the covariance between the two score statistics for all K traits. Let $R_k^{(\tau)} \equiv \hat{\phi}_k^{-1} \sum_{i=1}^n \{Y_{ki} - \mu_k(\hat{\alpha}_k^T X_i + \hat{\theta}_k^T W G_i)\} G_i$, the right half of the variance component test $U_k^{(\tau)} \equiv (R_k^{(\tau)})^T R_k^{(\tau)}$. The covariance between the two score statistics can be characterized through assessing the covariance matrix between $U^{(\theta)} \equiv (U_1^{(\theta)T}, \dots, U_K^{(\theta)T})^T$ and $R^{(\tau)} = (R_1^{(\tau)}, \dots, R_K^{(\tau)})^T$. As shown in Sun et al. (2013), $U_k^{(\theta)}$ and $R_k^{(\tau)}$ are independent for a given k . For the covariance between $U_k^{(\theta)}$ and $R_l^{(\tau)}$ for $k \neq l$, $k, l = 1, \dots, K$, perhaps surprisingly, we show that the covariance is equal to 0 for all $k \neq l$. The details of the proof are provided in Web Appendix. It can be shown that $U^{(\theta)}$ and $R^{(\tau)}$ have asymptotically a joint normal distribution, hence the two statistics are asymptotically independent. Subsequently, we can test $\theta \equiv (\theta_1^T, \dots, \theta_K^T)^T$ and $\tau \equiv (\tau_1, \dots, \tau_K)^T$ separately.

Specifically, we can construct a multivariate score test statistics for θ as $T_\theta \equiv U^{(\theta)T} \{V^{(\theta)}\}^{-1} U^{(\theta)}$, which follows a chi-square distribution with $q \times K$ degree of freedom. Here, $V^{(\theta)}$ is the covariance estimate for $U^{(\theta)}$ and can be calculated as follows. It can be shown that $n^{-1/2} U_k^{(\theta)}$ is asymptotically equivalent to the sum of n independent terms:

$$n^{-1/2} \phi_k^{-1} \sum_{i=1}^n \{Y_{ki} - \mu_k(\alpha_k^T X_i)\} (W G_i - E_k^T F_k^{-1} X_i),$$

where E_k and F_k are the limits of $n^{-1} \sum_{i=1}^n \mu'_k(\alpha_k^T X_i) X_i (W G_i)^T$ and $n^{-1} \sum_{i=1}^n \mu'_k(\alpha_k^T X_i) X_i X_i^T$, respectively, and $\mu'_k(x)$ is the first derivative of $\mu_k(x)$. Let $\tilde{E}_k = \sum_{i=1}^n \mu'_k(\tilde{\alpha}_k^T X_i) X_i (W G_i)^T$, and $\tilde{F}_k = \sum_{i=1}^n \mu'_k(\tilde{\alpha}_k^T X_i) X_i X_i^T$. Since $n^{-1/2} U_k^{(\theta)}$ can be approximated by the sum of n independent terms, the covariance matrix for $n^{-1/2} U^{(\theta)}$ can be estimated by

$n^{-1}V^{(\theta)} = n^{-1}\left(V_{kl}^{(\theta)}\right)_{k=1,\dots,K;l=1,\dots,K}$, where $V_{kl}^{(\theta)} = \sum_{i=1}^n U_{ki}^{(\theta)}U_{li}^{(\theta)T}$, and $U_{ki}^{(\theta)} = \tilde{\phi}_k^{-1}\{Y_{ki} - \mu_k(\tilde{\alpha}_k^T X_i)\}(WG_i - \tilde{E}_k^T \tilde{F}_k^{-1} X_i)$, $k=1, \dots, K$; $i=1, \dots, n$. There are potentially other ways for testing θ , such as the T' statistic in He et al. (2013), which is $e^T C^{-1} Z^T (e^T C^{-1} e)^{1/2}$, where $e = (1, \dots, 1)^T$, C is a $K \times K$ matrix with $C_{kl} = V_{kl}^{(\theta)} / (V_{kk}^{(\theta)} V_{ll}^{(\theta)})$, and Z is a K -length vector with $Z_k = U_k^{(\theta)} / V_{kk}^{(\theta)}$. T' tends to be more powerful when the fixed effects are similar to each other across different traits.

Next, we show how to test τ_k^2 , $k=1, \dots, K$, based on $U^{(\tau)}$. Let $M = (X, GW^T)$ and $P_M = (I - M(M^T M)^{-1} M^T)$. Let Ω denote the $K \times K$ covariance matrix among the traits for a subject conditional on M under H_{02} . Let $Y_k = (Y_{k1}, \dots, Y_{kn})^T$ and $\hat{\mu}_k = \hat{X} \alpha_k + GW^T \hat{\theta}_k$ be $n \times 1$ vectors. Then,

$$(Y_1^T - \hat{\mu}_1^T, \dots, Y_K^T - \hat{\mu}_K^T) \Sigma^{-1/2}$$

follows a standard multivariate normal distribution, where $\Sigma = \Omega \otimes P_M$ and $\otimes$ represents the Kronecker product. We propose a statistics

$$T_\tau = \sum_{k=1}^K U_k^{(\tau)},$$

where $U_k^{(\tau)}$ has been defined before, for testing $\tau_k^2 = 0$, $k=1, \dots, K$. For ease of presentation, we ignore the dispersion parameter $\hat{\phi}_k$ from $U_k^{(\tau)}$ for a moment. Then T_τ can be written in the form

$$(Y_1^T - \hat{\mu}_1^T, \dots, Y_K^T - \hat{\mu}_K^T) \Sigma^{-1/2} \Sigma^{1/2} A \Sigma^{1/2} \Sigma^{-1/2} (Y_1^T - \hat{\mu}_1^T, \dots, Y_K^T - \hat{\mu}_K^T)^T,$$

where $A = I_K \otimes GG^T$. Then, T_τ follows a mixture of weighted chi-square distributions with weights being the eigenvalues of $\Sigma^{1/2} A \Sigma^{1/2}$, or equivalently, ΣA . In practice Σ is unknown, so we use the empirical estimator $\hat{\Sigma}$ to replace Σ .

The dimension of $(\hat{\Sigma} A)$ is $nK \times nK$ which can be large for cancer genomic studies; hence directly calculating its eigenvalues can be time-consuming and numerically unstable. Note that

$$\hat{\Sigma} A = (\hat{\Omega} \otimes P_M)(I_K \otimes GG^T) = \hat{\Omega} \otimes P_M GG^T,$$

whose spectrum can be computed by $\lambda_j \times \vartheta_j$, ($i = 1, \dots, K, j = 1, \dots, n$), where λ_j and ϑ_j are the eigenvalues of $\hat{\Omega}$ and $P_M G G^T$, respectively. Hence, one only needs to calculate the eigenvalues from a $K \times K$ matrix and a $p \times p$ matrix separately.

In the above presentation, $\hat{\phi}_k$ is ignored in the expression of $U_k^{(\tau)}$. To account for $\hat{\phi}_k$, we only need to replace $\Omega = (\Omega_{kl})_{K \times K}$ by another matrix $\Omega^* = (\Omega_{kl} \phi_k^{-1} \phi_l^{-1})_{K \times K}$. This is because

$$\text{Cov}(\phi_k^{-1}(Y_k - \hat{\mu}_k), \phi_l^{-1}(Y_l - \hat{\mu}_l)) = \text{Cov}(\phi_k^{-1} Y_k P_M, \phi_l^{-1} Y_l P_M),$$

which is

$$\phi_k^{-1} \phi_l^{-1} P_M \text{Cov}(Y_k, Y_l) P_M = (\phi_k^{-1} \phi_l^{-1} \Omega_{kl}) \times P_M.$$

In matrix form, this is just $\Omega^* \otimes P_M$, and Ω^* can be empirically estimated from the data, i.e., $\hat{\Omega}^* = (\hat{\Omega}_{kl} \hat{\phi}_k^{-1} \hat{\phi}_l^{-1})_{K \times K}$. We did not use Ω when testing θ , as the correlations among the traits were accounted for by $V^{(\theta)}$ in an empirical manner, similar to He et al. (2013). For testing τ^2 , Ω is used as it can help to reduce the eigen-decomposition from a large matrix to smaller matrices, which can improve computational efficiency.

To combine the independent score tests of θ and τ , a number of nonparametric procedures can be used. One of the most widely used procedures is Fisher's combination. Thus, we adopt Fisher's procedure to benchmark the performance of our approach. Note that Sun et al. (2013) also used Fisher's or Tippett's procedure for combining their score statistics. Specifically, given the p values of the score tests θ and τ as p_θ and p_τ , we use $-2 \log(p_\theta) - 2 \log(p_\tau) \sim \chi_4^2$ to derive the overall p value. We name this procedure as fSOMAT. Later in Section 2.4, we further propose a new procedure for optimal linear combination of independent test statistics, T_θ and T_τ .

2.3 Binary Traits Or A Mixture Of Traits

When there exist binary traits, the score statistic for the fixed effects θ_k as well as the covariance matrix for θ can be estimated as before. However, the covariance matrix for the score statistic of the variance components τ^2 no longer has a closed form. Without loss of generality, assume that the k th trait is binary. To obtain the covariance matrix for the score of τ^2 , we note that $n^{-1/2} R_k^{(\tau)}$, which is in the form of

$n^{-1/2} \sum_{i=1}^n \hat{\phi}_k^{-1} \{Y_{ki} - \mu_k(\hat{\alpha}_k^T X_i + \hat{\theta}_k^T W G_i)\} G_i$, can be shown to be asymptotically equivalent to

$$n^{-1/2} \sum_{i=1}^n \phi_k^{-1} \{Y_{ki} - \mu_k(\alpha_k^T X_i + \theta_k^T W G_i)\} (G_i - E_k^* F_k^{*-1} M_i),$$

where each summand is independent, and E_k^* and F_k^* are the limits of $n^{-1} \sum_{i=1}^n \mu_k'(\alpha_k^T X_i + \theta_k^T W G_i) M_i G_i^T$ and $n^{-1} \sum_{i=1}^n \mu_k'(\alpha_k^T X_i + \theta_k^T W G_i) M_i M_i^T$, respectively. Let $R_{ki}^{(\tau)}$ be the estimate of $\phi_k^{-1} \{Y_{ki} - \mu_k(\alpha_k^T X_i + \theta_k^T W G_i)\} (G_i - E_k^{*T} F_k^{*-1} M_i)$ (i.e., with the parameters $\phi_k, \alpha_k, \theta_k$ replaced by their estimates). Then, the covariance matrix between $n^{-1/2} R_k^{(\tau)}$ and $n^{-1/2} R_l^{(\tau)}$ can be estimated by

$$n^{-1} V_{kl}^{(\tau)} = n^{-1} \sum_{i=1}^n R_{ki}^{(\tau)} R_{li}^{(\tau)}.$$

Thus we can estimate the covariance matrix for $n^{-1/2} R^{(\tau)}$ by

$$n^{-1} V^{(\tau)} = n^{-1} \left(V_{kl}^{(\tau)} \right)_{k=1, \dots, K; l=1, \dots, K}. \text{ Subsequently, the statistic } T_\tau = (R^{(\tau)})^T R^{(\tau)}$$

approximately follows a mixed chi-square distribution with weights being the eigenvalues of $V^{(\tau)}$. Here, Ω is not used for testing τ^2 , because it is not possible to write Σ as $\Omega \otimes P_M$ when binary traits are involved; the reason is that each binary trait (say k th trait) has its own $P_{M,k}$, where $P_{M,k} = I - D_k M (M^T D_k M)^{-1} M^T$, and $D_k = \text{Cov}(Y_k)$.

In addition to obtaining $V^{(\tau)}$, one needs to derive the covariance between $U^{(\theta)}$ and $R^{(\tau)}$ in order to test the nullity of θ and τ^2 jointly. When there are no confounding covariates, we show in the Web Appendix that $U^{(\theta)}$ and $R^{(\tau)}$ are statistically independent, i.e., $\text{Cov}(U^{(\theta)}, R^{(\tau)}) = 0$. In fact, as long as $\text{Cov}(Y_k)$ can be expressed as a scalar matrix, i.e., $c \times I_{n \times n}$ for some constant c , then it always holds that $\text{Cov}(U^{(\theta)}, R^{(\tau)}) = 0$ (see Web Appendix for more details). When there exist confounding covariates in the model, $\text{Cov}(U^{(\theta)}, R^{(\tau)})$ in general is not equal to 0. However, as shown by Hu et al. (2013), when the proportion of cases (POC) for the binary trait is not close to 0 or 1, the covariance of $Y_{ki}, \mu_k(1 - \mu_{ki})$, does not strongly depend on the covariate values, where $\mu_{ki} = \exp(\alpha_k^T X_i) / (1 + \exp(\alpha_k^T X_i))$. Hence, $\text{Cov}(Y_k)$ can be approximated by a scalar matrix, $\mu_{k0}(1 - \mu_{k0}) \times I$, where $\mu_{k0} = \exp(\alpha_{k1}) / (1 + \exp(\alpha_{k1}))$ and α_{k1} is the intercept component of α_k . Therefore, we can still treat the scores of θ and τ^2 as independent terms, and utilize Fisher's procedure to combine the p values for θ and τ^2 .

2.4 A New Procedure for Combining the Test Statistics

While Fisher's procedure is widely used for combining two independent test statistics, it may not be optimal for combining the evidence from two test statistics. We consider a linear combination of these statistics, $Q(\rho) = (1 - \rho)T_\theta + \rho T_\tau$, where $\rho \in [0, 1]$. When $\rho = 0$, the test statistic is the score statistic for fixed effects, and when $\rho = 1$ the test statistic is the random effect score statistic. Note that (1) T_θ asymptotically follows a chi-square distribution with degree of freedom qK , (2) T_τ asymptotically follows a mixed chi-square distribution with weights being the eigen values of $\hat{\Sigma}A$, $\lambda_1, \dots, \lambda_{pK}$, and (3) T_θ is independent of T_τ . Hence the mean and standard deviation of $Q(\rho)$ are $\mu_{Q(\rho)} = (1 - \rho)qK + \rho\Lambda_1$ and $\sigma_{Q(\rho)} = \sqrt{2(1 - \rho)^2 qK + 2\rho^2 \Lambda_2}$, respectively, where $\Lambda_1 = \sum_{j=1}^{pK} \lambda_j$ and $\Lambda_2 = \sum_{j=1}^{pK} \lambda_j^2$. It is tempting to use $\sup_{\rho \in [0, 1]} Q(\rho)$ as the test statistic; however, because $\mu_{Q(\rho)}$ and $\sigma_{Q(\rho)}$

vary with ρ , $Q(\rho)$ is not directly comparable across different values of ρ . Thus, we propose to first standardize $Q(\rho)$ and then use

$$T^* = \sup_{0 \leq \rho \leq 1} T(\rho) = \sup_{0 \leq \rho \leq 1} \left\{ \frac{Q(\rho) - \mu_{Q(\rho)}}{\sigma_{Q(\rho)}} \right\}$$

as the test statistic, which we name as optimal SOMAT (oSOMAT). This standardization idea shares the same spirit as Davies (1987), where a standardized test statistic was considered under a series of values of a tuning parameter and the maximal statistic was chosen as the test statistic.

Denote the observed values of T_θ and T_τ as t_θ and t_τ . We wish to find

$$t^* = \sup_{0 \leq \rho \leq 1} \frac{(1-\rho)t_\theta + \rho t_\tau - \mu_{Q(\rho)}}{\sigma_{Q(\rho)}}. \text{ It can be shown that } t^* \text{ has a closed form,}$$

$$t^* = \begin{cases} \sqrt{\frac{(t_\theta - qK)^2}{2qK} + \frac{(t_\tau - \Lambda_1)^2}{2\Lambda_2}}, & \text{if } t_\theta > qK \text{ and } t_\tau > \Lambda_1, \\ \max \{(t_\theta - qK)/\sqrt{2qK}, (t_\tau - \Lambda_1)/\sqrt{2\Lambda_2}\}, & \text{otherwise.} \end{cases}$$

The following theorem describes how to calculate the p-value of the test statistics T^* .

Theorem 1—Let $F_\theta(\cdot)$ and $f_\theta(\cdot)$ be the distribution and the density functions of T_θ which follows a chi-square distribution with degree of freedom qK . Let $F_\tau(\cdot)$ be the distribution function of T_τ which follows a mixed chi-square distribution with weights being $\lambda_1, \dots, \lambda_{pK}$. If T_θ and T_τ are independent, then

$$P(T^* > t^*) = \begin{cases} 1 - F_\tau(t_\tau^*)F_\theta(qK) - \int_{qK}^{t_\theta^*} F_\tau(\delta(u))f_\theta(u)du, & \text{if } t^* > 0 \\ 1 - F_\tau(t_\tau^*)F_\theta(t_\theta^*), & \text{if } t^* \leq 0, \end{cases}$$

where $t_\theta^* = t^* \sqrt{2qK} + qK$, $t_\tau^* = t^* \sqrt{2\Lambda_2} + \Lambda_1$, and $\delta(u) = \sqrt{(t^*)^2 - \left(\frac{u - qK}{\sqrt{2qK}}\right)^2} \sqrt{2\Lambda_2} + \Lambda_1$.

In Theorem 1, $F_\tau(\cdot)$ can be well approximated by the moment matching method of Liu, Tang, and Zhang (2009). The integral part can be evaluated by univariate numerical integration, hence the p-value of T^* can be easily calculated. Note that Lee et al. (2012) also proposed an optimal test (SKAT-O) for combining SKAT and the burden test, $\inf_{0 \leq \rho \leq 1} p_\rho$, where p_ρ is the p value calculated for a given ρ . Their procedure seeks to directly minimize p_ρ , which involves complicated quantile functions. Hence, SKAT-O does not have a closed form for the optimal ρ , and requires a grid search to find such ρ . The number and placement of grid points could all affect whether an optimal ρ can be more accurately identified. Our proposed procedure instead aims to maximize the standardized $Q(\rho)$, which yields

numerically tractable solution to identifying optimal ρ . It is not only computationally efficient but also guarantees the optimal ρ is identified.

3. Simulation Study

We conducted simulations to evaluate the performance of the proposed approach in finite sample sizes, and compared it with several other approaches. Specifically, we included three univariate test statistics, as it is common in practice to analyze each trait one by one and then apply for the Bonferroni correction to account for testing for multiple traits. These three univariate test statistics are: (1) Score test under the generalized linear model (uni-GLM), (2) SKAT (uni-SKAT), and (3) MiST (uni-MiST). For each of these tests, the multivariate trait p-value is calculated by $K \times \min_{k=1, \dots, K} p^{(k)}$, where $p^{(k)}$ is the univariate p-value for the k th trait. In addition, we also included two multivariate test statistics: (4) GEE-based multivariate score test (gScore) (He et al., 2013), where the effects of somatic mutations are treated as fixed, and (5) multivariate linear kernel machine test (MV-KM) (Maity et al., 2012) where the effects of somatic mutations are assumed to be random. For the uni-SKAT, we used the default weights, i.e., $w_j = \text{dbeta}(\text{MAF}_j, 1, 25)$, where MAF_j is the minor allele frequency of the j th mutation, and dbeta is the density function of the beta-distribution. For the MV-KM, we used the default linear kernel which contains no weights. Note that the MV-KM can be applied only to continuous traits, and hence this approach is not included when binary traits are involved.

To mimic real applications, we simulated a pathway that includes p genes. Here we assume that the somatic mutations are consolidated at the gene-level. Specifically, let G_i be the presence/absence of non-silent somatic mutations for p genes on the i th subject ($i = 1, \dots, n$), i.e., $G_{ij} = 1$ if the j th gene in the i th sample contains mutations, and 0 otherwise. The mutations were generated from a Bernoulli distribution with probability p_j where $p_j, j = 1, \dots, p$, are set to be between 1% and 2%. We simulated one confounding covariate X_i from a normal variable with mean $0.1 \times G_{i1}$ and variance 1, where G_{i1} is the first component of G_i . We considered the situation in which only continuous traits were present in the model.

Two traits were generated as $Y_{1i} = 1 + X_i + \beta_1^T G_i + \varepsilon_{1i}$ and $Y_{2i} = 1 + X_i + \beta_2^T G_i + \varepsilon_{2i}$, where ε_{1i} and ε_{2i} follow the standard bi-variate normal distribution with correlation equal to 0.5. We also considered the situation in which both binary and continuous traits were included in the model. The binary trait was generated by $\text{logit}\{P(Y_{1i} = 1)\} = -1 + 0.5X_i + \beta_1^T G_i$, and to introduce correlation among the two traits, the continuous trait was generated by $Y_{2i} = 1 + X_i + \beta_2^T G_i + \varepsilon_{2i}$, where $\varepsilon_{2i} \sim N(Y_{1i}, 1)$. Under each simulation setting, we considered sample size $n = 500$ and $n = 1000$, and the number of genes $p = 10$ and $p = 20$. For each setting, a total of 10,000 datasets were generated.

We assessed the type I error by generating the bivariate traits with $\beta_1 = \beta_2 = 0$ (Table 1). All methods have well-controlled type I error. To compare power, we specified the genetic effects as follows, $\beta_{1j} = \theta_1 W_j + \xi_{1j}$, $\beta_{2j} = \theta_2 W_j + \xi_{2j}$, $j = 1, \dots, p$, where ξ_{1j} follows the normal distribution with mean 0 and variance τ_1^2 , and ξ_{2j} follows the normal distribution

with mean 0 and variance τ_2^2 . We set $W_j = 1$ for $j = 1, \dots, p$, $\theta_1 = 0.2$, and $\theta_2 = -0.2$.

Further, we set $\tau_1^2 = \tau_2^2 = 0.2^2$ to keep β_{1j} and β_{2j} at reasonable sizes. We named this model-setting as H_s . Table 2 summarizes the power comparison. The uni-GLM and gScore both have low power for detecting the genetic association, mainly because these methods are designed to capture mutations with fixed effects and do not accommodate the random effects in the model. Similarly, the SKAT and MV-KM that only account for the random effects and fail to include the fixed effects also leads to power loss in this setting. Compared to the uni-MiST, fSOMAT and oSOMAT showed superior performance because they are able to garner the genetic effects from both traits and conduct the association test in a joint manner. The oSOMAT has a moderate but noticeable gain of power over fSOMAT. In the situation of mixed traits (lower panel), all methods tend to have lower statistical power due to less information carried by the binary trait, but the relative performance is similar to continuous traits (upper panel). We also varied the correlation among the two continuous traits to 0.2 and 0.8, and our approaches remain to be more powerful than existing approaches (see Web Tables 1–2).

In the previous experiment, we let the effects of all the somatic mutations to be nonzero (i.e., all the mutations contribute to the traits of interest). We wish to test if our proposed approach is sensitive to this assumption. To this end, we let 10% of the components of β_1 and β_2 to have effect of 0, that is, 10% of the mutations do not have any influence on the traits. We named this model as H_{a1} . The results are shown in Table 3. It can be seen that all methods have visible loss of power, likely due to weakened signals from the set of somatic mutations. However, fSOMAT and oSOMAT compete favorably with the other methods, suggesting that our approaches are well suited to handle this situation. To let more somatic mutations to have a null effect, we set 20% of the components of β_1 and β_2 to be zero, and named this model as H_{a2} . Table 4 shows that further power loss occurred under H_{a2} for all methods, but fSOMAT and oSOMAT remain to be competitive and the pattern of statistical power is largely consistent with that of Table 3.

We also conducted experiments on type I error and power analysis with four continuous traits, where the correlations among the four traits were set to be 0.5. To reduce the overall genetic signals, we let $\theta_1 = 0.1$, $\theta_2 = 0.1$, $\theta_3 = -0.1$, $\theta_4 = -0.1$, and $\tau_1^2 = \dots = \tau_4^2 = 0.1^2$. Web Table 3 shows that the SOMAT continues to perform well and can sometimes have substantial gain of power over the compared methods. We conducted simulation studies to evaluate our approach under a smaller sample size (Web Tables 4–5), weaker correlations among the traits (Web Table 8), higher values of q and K (Web Tables 9–11 and Web Figure 1), and non-symmetrical effects for θ_k (Web Tables 17–18); the results show that our approach remains to be competitive under these situations. We tested our approach under a series of values of the POC, and found that the proposed approach performs well even when the POC is as low as 0.05 or as high as 0.95 (Web Tables 6–7). When there are only random effects in the model, we found that MV-KM has the highest power (Web Table 12); however, our proposed approach is competitive. When there exist only fixed effects, our approach has the highest power (Web Table 13); the gScore approach has low power under such a situation due to its high degree of freedom for testing the fixed effects. We also investigated

the situations in which only a subset of the traits are associated with mutations, and our approach competes favorably with existing approaches (Web Tables 14–16).

4. Real Data Analysis

For illustration we applied the proposed SOMAT approach as well as other methods to the hepatocellular carcinoma (HCC) somatic mutation data from the cancer genome atlas (TCGA) consortium to evaluate the association of somatic mutations with HCC-related traits. Briefly, the HCC data was curated by the Baylor College of Medicine Human Genome Sequencing Center. The data include 373 tumor samples with 39,422 gene transcripts, where the presence/absence of non-silent somatic mutations is recorded for each gene (i.e., $G_{ij} = 1$ if the j th gene in the i th sample contains mutations, and 0 otherwise).

Our goal is to identify somatic mutations that have significant health impact on HCC cancer patients. Previous studies have shown that α -fetoprotein (AFP) and prothrombin time (PT) are important clinical traits associated with the survival of HCC patients; see Ikeda et al. (1998) and Lai et al. (1995) for further details. In particular, AFP is treated as a key clinical trait for post-treatment monitoring of HCC patients, and an out-of-range high AFP-value is considered as an alarming sign for patients' survival. The second trait, PT, is related to liver damage or cirrhosis and also commonly measured for HCC patients. In literature, the AFP is usually transformed into a binary factor for analysis (Ikeda et al., 1998), and PT can be treated as a continuous trait in the regression analysis of HCC (Tang et al., 2012). Hence we dichotomized the AFP based on whether or not a patient's AFP values is above his/her upper limit of normal values. Thus, our joint analysis includes two traits, the binary AFP-level and the continuous PT. After removing observations with missing values, we obtained $n = 252$ samples for analysis. We included age, gender as the confounding covariates, and removed genes with singleton or doubleton somatic mutations. Web Table 19 presents the frequency of somatic mutations in all of the 39,422 transcripts among the 252 samples. It can be seen that 27,153 (69%) of the transcripts do not carry any somatic mutations, and 8,498 (22%) transcripts harbor mutations in less than 1% of the samples. Only 76 (0.2%) transcripts have mutations in more than 5% of the samples, showing that somatic mutations in general occur at very low frequencies. The pattern of these observed frequencies is consistent with that in the European HCC data (Schulze et al., 2015).

We assigned the genes with mutations into different pathways according to the Broad Institute's hallmark pathway database. This yielded 50 genetic pathways, with each pathway containing 4~82 mutated genes. For each of these pathways, we applied SOMAT along with the other considered methods for comparison (MV-KM is not included because it does not accommodate binary traits). For controlling the family-wise error rate (FWER) at the level 0.05, the threshold for declaring statistical significance was set to be $0.05/50 = 0.001$. At this significance level, both fSOMAT and oSOMAT identified one (identical) significant pathway, the wnt- β -catenin signaling pathway (see Table 5). The component-wise p-values for the fixed effects and random effects of SOMAT can be found in Web Table 20. Figure 1 shows a Manhattan plot of oSOMAT p-values for all 50 pathways. The gScore did not detect any significant pathway, likely because it only considered fixed effects. The uni-SKAT (under its default kernel, i.e., with weights based on the minor allele frequency of mutations)

yielded a p-value of 0.4933. We also applied the uni-SKAT with no weights (called uni-rSKAT) and it gave a p-value of 0.0006, which is higher than the oSOMAT's p-value (0.0002). The uni MiST accommodates both effects but does not consider the two traits jointly. The univariate-trait p-values (without Bonferroni correction) for AFP and PT in uni-MiST are $1.1E-3$ and 0.1626, respectively, neither of which reaches the univariate significance level of 0.001/2. The wnt- β -catenin pathway includes 12 genes with somatic mutations (see Web Table 21), and these genes involve various cellular functions, such as chromatin modification (HDAC5), and cell growth and adhesion (CTNNB1). Previous work has indicated that the wnt- β -catenin pathway is involved in liver metabolism and cancer (Monga, 2011), and our finding further suggests that somatic mutations in this pathway can have significant impact on the health outcomes of HCC.

To gain more insight into the wnt- β -catenin pathway, we examined the biological connections of the gene components in this pathway. We used the BisoGenet, a genetic network visualization tool (Martin et al., 2010), to plot the wnt- β -catenin pathway. This tool utilizes existing biological function databases to characterize the connections between different genes. Web Figure 2 shows the biological interactions among these genes, including self-interactions. (One gene was omitted due to insufficient information on this gene.) It can be seen that several genes, TP53, AXIN1, CTNNB1, and NOTCH1, are located in the central part of the pathway, i.e., they tend to have a higher number of biological connections than the other genes. Interestingly, TP53, AXIN1, and CTNNB1 are observed to have higher mutation rates within this pathway, suggesting that genes in the center of a pathway may be subject to higher mutagenesis pressure. This indicates that drug development may be first focused on these priority genes. Further functional studies, such as Zhuang et al. (2012), will help to elucidate how the genes (and their mutations) interact with each other to influence the clinical traits that are associated with HCC.

5. Discussion

Understanding the association between somatic mutations and cancer-related traits is a fundamental topic in cancer research. The extremely low frequency of somatic mutations poses tremendous challenges for statistical analysis of these mutations. Furthermore, because cancer is a highly complex disease, analyzing a single trait only often lacks power to detect the association of somatic mutations. In this article, we have proposed an effective approach for analyzing somatic mutations under multiple traits. Our approach accounts for both fixed and random effects, and can accommodate the mixture of binary and continuous traits. These features allow the application of SOMAT in many practical settings. Moreover, oSOMAT that combines the two independent test statistics linearly is numerically tractable and computationally fast. In our experiments, we have observed that for two traits with $n = 1000$ and $p = 10$, the typical computation time for MV-KM is 114 seconds, while oSOMAT only needs 0.03 second to complete the calculation. This feature of oSOMAT is useful for handling datasets with large sample sizes, which has become increasingly common for cancer genomic studies. Simulation studies and real data analysis demonstrate that the proposed SOMAT approach has excellent performance in identifying disease-associated somatic mutations.

In our analysis, we used the presence/absence of mutations in each gene for defining the genotype. Alternatively, one may weigh different mutations by their functional impact score (Reva, Antipin, and Sander, 2011) and then aggregate them at the gene level by linear combination. To prevent assigning undue scores to large genes that tend to have more somatic mutations, one may standardize the score by the size of the gene. For cancers that tend to have relatively high mutation frequencies due to life style or environmental exposure, such as melanoma and lung cancer, it may be sensible to examine mutations at higher resolution, that is, study different mutations separately, such as nonsense, missense, and splice-site mutations. This may help to finesse the quantification of the biological effects of various mutations (Petitjean et al., 2007). Furthermore, prior knowledge on biological pathway/network provides valuable information for understanding the interactions among the studied genes. How to utilize this network information for association analysis of somatic mutations will be an interesting topic for future research.

Somatic mutagenesis, by nature, is a dynamic process depending on both space (at cellular level) and time (Swanton, 2012). For example, it has been shown that tumor tissues usually consist of highly heterogeneous subclones, and different subclones may have distinct mutation profiles (Burrell and Swanton, 2014). Current technology is still primarily focused on the average mutation profile of a tumor, but the emerging single-cell-sequencing technology can provide more in-depth characterization of the subpopulations of the tumors. This spatial dependence is further complicated by metastasis, which involves the movement and adaption of tumor cells to a new site (Michor and Polyak, 2010). At the same time, tumor is evolving with time and constrained by chemo or radio-therapy, and hence it would be meaningful to trace the longitudinal changes of mutations to better understand the etiology of cancer. It will be interesting to extend our methods in future to incorporate such spatial and temporal correlations in the model.

In summary, we have developed an approach that is statistically powerful and computationally efficient for studying the association between somatic mutations and cancer-related traits. The analysis of somatic mutations will provide new insights into tumor development and lead to better understanding of cancer etiology.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The results shown here are in part based upon data generated by the TCGA Research Network. The authors acknowledge support from Fred Hutchinson Cancer Research Center's Institutional Support, R01CA189532, R01CA195789, and U01CA137088. QH and YL equally contributed to this article.

References

- Burrell RA, Swanton C. Tumour heterogeneity and the evolution of polyclonal drug resistance. *Molecular oncology*. 2014; 8:1095–1111. [PubMed: 25087573]
- Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*. 1987; 74:33–43.

- Ferreira MA, Purcell SM. A multivariate test of association. *Bioinformatics*. 2009; 25:132–133. [PubMed: 19019849]
- He Q, Avery CL, Lin DY. A general framework for association tests with multivariate traits in large-scale genomics studies. *Genetic epidemiology*. 2013; 37:759–767. [PubMed: 24227293]
- Hu YJ, Berndt SI, Gustafsson S, Ganna A, Hirschhorn J, North KE, et al. Meta-analysis of gene-level associations for rare variants based on single-variant statistics. *The American Journal of Human Genetics*. 2013; 93:236–248. [PubMed: 23891470]
- Ikeda Y, Shimada M, Hasegawa H, Gion T, Kajiyama K, Shirabe K, et al. Prognosis of hepatocellular carcinoma with diabetes mellitus after hepatic resection. *Hepatology*. 1998; 27:1567–1571. [PubMed: 9620328]
- Lai E, Fan ST, Lo CM, Chu KM, Liu CL, Wong J. Hepatic resection for hepatocellular carcinoma. an audit of 343 patients. *Annals of surgery*. 1995; 221:291–298. [PubMed: 7717783]
- Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013; 499:214–218. [PubMed: 23770567]
- Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012; 13:762–775. [PubMed: 22699862]
- Liu H, Tang Y, Zhang HH. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*. 2009; 53:853–856.
- Liu Y, Yasukawa M, Chen K, Hu L, Broaddus RR, Ding L, et al. Association of somatic mutations of adams genes with chemotherapy sensitivity and survival in high-grade serous ovarian carcinoma. *JAMA oncology*. 2015; 1:486–494. [PubMed: 26181259]
- Luzzatto L. Somatic mutations in cancer development. *Environmental Health*. 2011; 10:S12. [PubMed: 21489208]
- Maity A, Sullivan PF, Tzeng JY. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic epidemiology*. 2012; 36:686–695. [PubMed: 22899176]
- Martin A, Ochagavia ME, Rabasa LC, Miranda J, Fernandez-de Cossio J, Bringas R. Bisogenet: a new tool for gene network building, visualization and analysis. *BMC bioinformatics*. 2010; 11:91. [PubMed: 20163717]
- Michor F, Polyak K. The origins and implications of intratumor heterogeneity. *Cancer prevention research*. 2010; 3:1361–1364. [PubMed: 20959519]
- Monga SPS. Role of wnt/ β -catenin signaling in liver metabolism and cancer. *The international journal of biochemistry & cell biology*. 2011; 43:1021–1029. [PubMed: 19747566]
- Petitjean A, Achatz M, Borresen-Dale A, Hainaut P, Olivier M. Tp53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene*. 2007; 26:2157–2165. [PubMed: 17401424]
- Rachakonda PS, Bauer AS, Xie H, Campa D, Rizzato C, Canzian F, et al. Somatic mutations in exocrine pancreatic tumors: association with patient survival. *PloS one*. 2013; 8:e60870. [PubMed: 23565280]
- Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*. 2011; 39:e118. [PubMed: 21727090]
- Schulze K, Imbeaud S, Letouzé E, Alexandrov LB, Calderaro J, Rebouissou S, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nature genetics*. 2015; 47:505–511. [PubMed: 25822088]
- Sun J, Ouakacha K, Forgetta V, Zheng HF, Richards JB, Ciampi A, et al. A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects. *European Journal of Human Genetics*. 2016; 24:1344–51. [PubMed: 26860061]
- Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic epidemiology*. 2013; 37:334–344. [PubMed: 23483651]
- Swanton C. Intratumor heterogeneity: evolution through space and time. *Cancer research*. 2012; 72:4875–4882. [PubMed: 23002210]

- Tang W, Schwienbacher C, Lopez LM, Ben-Shlomo Y, Oudot-Mellakh T, Johnson AD, et al. Genetic associations for activated partial thromboplastin time and prothrombin time, their gene expression profiles, and risk of coronary artery disease. *The American Journal of Human Genetics*. 2012; 91:152–162. [PubMed: 22703881]
- TCGA. [accessed April 20, 2016] <https://genome-cancer.soe.ucsc.edu/proj/site/xena/datapages/>
- Watson IR, Takahashi K, Futreal PA, Chin L. Emerging patterns of somatic mutations in cancer. *Nature reviews Genetics*. 2013; 14:703–718.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*. 2011; 89:82–93. [PubMed: 21737059]
- Yang Q, Wu H, Guo CY, Fox CS. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic epidemiology*. 2010; 34:444–454. [PubMed: 20583287]
- Yu B, O'Toole SA, Trent RJ. Somatic dna mutation analysis in targeted therapy of solid tumours. *Translational pediatrics*. 2015; 4:125–138. [PubMed: 26835368]
- Zhuang Z, Yang C, Lorenzo F, Merino M, Fojo T, Kebebew E, et al. Somatic hif2a gain-of-function mutations in paraganglioma with polycythemia. *New England Journal of Medicine*. 2012; 367:922–930. [PubMed: 22931260]

Figure 1.
Analysis of somatic mutations in the hallmark pathways by the oSOMAT approach

Table 1

Type I error for continuous traits (Cont.) and mixed traits (Mixed)

	<i>n</i>	<i>p</i>	uni-GLM	uni-SKAT	uni-MIST	gScore	MV-KM	ISOMAT	oSOMAT
Cont.	500	10	0.048	0.046	0.047	0.027	0.051	0.048	0.049
		20	0.045	0.044	0.047	0.023	0.048	0.046	0.047
	1000	10	0.049	0.048	0.049	0.039	0.050	0.054	0.052
		20	0.044	0.045	0.043	0.035	0.049	0.050	0.054
Mixed	500	10	0.041	0.040	0.042	0.028	—	0.036	0.034
		20	0.043	0.044	0.043	0.027	—	0.036	0.036
	1000	10	0.045	0.044	0.048	0.039	—	0.046	0.043
		20	0.042	0.041	0.045	0.033	—	0.041	0.041

Table 2

Power for continuous traits (Cont.) and mixed traits (Mixed) under H_s

	<i>n</i>	<i>p</i>	uni-GLM	uni-SKAT	uni-MiST	gScore	MV-KM	fSOMAT	oSOMAT
Cont.	500	10	0.183	0.193	0.391	0.105	0.577	0.705	0.755
		20	0.261	0.282	0.761	0.164	0.831	0.973	0.985
	1000	10	0.389	0.399	0.740	0.325	0.926	0.970	0.982
		20	0.610	0.634	0.989	0.585	0.996	0.999	1.000
Mixed	500	10	0.110	0.112	0.162	0.058	—	0.216	0.226
		20	0.132	0.142	0.316	0.068	—	0.489	0.527
	1000	10	0.162	0.166	0.292	0.116	—	0.457	0.497
		20	0.234	0.241	0.586	0.167	—	0.845	0.877

Table 3

Power for continuous traits (Cont.) and mixed traits (Mixed) for H_{a1}

	<i>n</i>	<i>p</i>	uni-GLM	uni-SKAT	uni-MIST	gScore	MV-KM	ISOMAT	oSOMAT
Cont.	500	10	0.172	0.185	0.351	0.100	0.553	0.637	0.687
		20	0.245	0.266	0.695	0.156	0.806	0.950	0.967
	1000	10	0.361	0.381	0.678	0.309	0.913	0.946	0.963
		20	0.574	0.603	0.973	0.554	0.992	0.999	0.999
Mixed	500	10	0.105	0.107	0.147	0.057	—	0.189	0.194
		20	0.121	0.133	0.278	0.064	—	0.427	0.464
	1000	10	0.155	0.161	0.261	0.114	—	0.404	0.433
		20	0.217	0.227	0.525	0.158	—	0.785	0.824

Table 4

Power for continuous traits (Cont.) and mixed traits (Mixed) for H_{a2}

	<i>n</i>	<i>p</i>	uni-GLM	uni-SKAT	uni-MIST	gScore	MV-KM	ISOMAT	oSOMAT
Cont.	500	10	0.159	0.176	0.305	0.093	0.519	0.559	0.598
		20	0.221	0.243	0.614	0.143	0.776	0.909	0.935
	1000	10	0.324	0.354	0.601	0.284	0.888	0.905	0.924
		20	0.521	0.562	0.939	0.515	0.987	0.999	0.999
Mixed	500	10	0.100	0.101	0.131	0.055	—	0.160	0.163
		20	0.115	0.125	0.244	0.063	—	0.365	0.397
	1000	10	0.144	0.152	0.230	0.108	—	0.346	0.369
		20	0.202	0.213	0.460	0.153	—	0.712	0.751

Table 5
Results of genetic pathway analysis for somatic mutations with respect to the AFP and PT traits

Pathway	#genes	uni-GLM	uni-SKAT	uni-MIST	gScore	fSOMAT	oSOMAT
wnt- β -catenin	12	0.0237	0.4933	0.0023	0.0119	0.0006	0.0002