PLOS | GENETICS

RESEARCH ARTICLE

Association analysis using somatic mutations

Yang Liu[1], Qianchan He[2], Wei Sun[2]*

1 Department of Mathematics and Statistics, Wright State University, Dayton, Ohio, United States of America, **2** Biostatistics Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America

* wsun@fredhutch.org

Abstract

Somatic mutations drive the growth of tumor cells and are pivotal biomarkers for many cancer treatments. Genetic association analysis using somatic mutations is an effective approach to study the functional impact of somatic mutations. However, standard regression methods are not appropriate for somatic mutation association studies because somatic mutation calls often have non-ignorable false positive rate and/or false negative rate. While large scale association analysis using somatic mutations becomes feasible recently—thanks for the improvement of sequencing techniques and the reduction of sequencing cost—there is an urgent need for a new statistical method designed for somatic mutation association analysis. We propose such a method with computationally efficient software implementation: Somatic mutation Association test with Measurement Errors (SAME). SAME accounts for somatic mutation calling uncertainty using a likelihood based approach. It can be used to assess the associations between continuous/dichotomous outcomes and individual mutations or gene-level mutations. Through simulation studies across a wide range of realistic scenarios, we show that SAME can significantly improve statistical power than the naive generalized linear model that ignores mutation calling uncertainty. Finally, using the data collected from The Cancer Genome Atlas (TCGA) project, we apply SAME to study the associations between somatic mutations and gene expression in 12 cancer types, as well as the associations between somatic mutations and colon cancer subtype defined by DNA methylation data. SAME recovered some interesting findings that were missed by the generalized linear model. In addition, we demonstrated that mutation-level and gene-level analyses are often more appropriate for oncogene and tumor-suppressor gene, respectively.

Author summary

Cancer is a genetic disease that is driven by the accumulation of somatic mutations. Association studies using somatic mutations is a powerful approach to identify the potential impact of somatic mutations on molecular or clinical features. One challenge for such tasks is the non-ignorable somatic mutation calling errors. We have developed a statistical method to address this challenge and applied our method to study the gene expression traits associated with somatic mutations in 12 cancer types. Our results show that some

Citation: Liu Y, He Q, Sun W (2018) Association analysis using somatic mutations. PLoS Genet 14(11): e1007746. https://doi.org/10.1371/journal.pgen.1007746

Editor: Xiaofeng Zhu, Case Western Reserve University, UNITED STATES

Received: February 7, 2018

Accepted: October 7, 2018

Published: November 2, 2018

Data Availability Statement: We have used the genomic and clinical data from The Cancer Genome Atlas, and these data are publicly available from the NCI GDC data portal (https://portal.gdc.cancer.gov/).

Funding: WS and YL's research is supported in part by National Institute of General Medical Sciences Grants GM105785 and GM126550. QH's research is supported in part by National Cancer Institute NCI-R01CA223498 and NCI-P30CA015704. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We have used the cloud service from Seven Bridges Cancer Genomics Cloud to extract read count information from bam files. The Seven Bridges Cancer Genomics Cloud has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I.

Competing interests: The authors have declared that no competing interests exist.

somatic mutations affect gene expression in several cancer types. In particular, we show that the associations between gene expression traits and TP53 gene level mutation reveal some similarities across a few cancer types.

Introduction

Somatic mutations play a central role in the development and progression of cancer. Associations between somatic mutations and molecular/clinical outcomes can provide important insights into cancer etiology or the mechanism of tumor growth, and potentially contribute to precision cancer therapy. Despite the functional importance of somatic mutations, few computational methods have been developed for association studies using somatic mutations. There are probably two reasons for this. First, since somatic mutation data are relatively new, most efforts were spent on bioinformatic challenges such as somatic mutation calling and functional annotations, e.g., inference of driver mutations [1–3], or estimation of cancer subtypes using somatic mutations [4, 5]. Second, systematic studies of somatic mutations in large observational studies are not feasible until recently, thanks for the drop of sequencing cost and the improved capability to handle formalin-fixed paraffin-embedded (FFPE) tissue samples. While these challenges on sequencing tumor samples and calling mutations have been addressed, a limiting factor to harvest the rich information of somatic mutation associations is appropriate statistical methods for data analysis.

A unique feature of somatic mutations, in contrast to germline mutations, is the difficulty to confidently call mutations from sequencing data. A major factor that contributes to this challenge is that a tumor sample is often a mixture of tumor cells and non-tumor cells (e.g., infiltrating immune cells) and a somatic mutation may only occur in a subset of the tumor cells, known as intra-tumor heterogeneity [6]. Therefore the signals of a somatic mutation may be visible only in a small proportion of sequence reads, and it is challenging to separate such weak signals from sequencing errors or DNA damages caused by FFPE [7]. Another factor that limits mutation call availability/accuracy is low coverage of sequencing reads, particularly in whole genome sequencing data. Although many methods have been developed for somatic mutation calling [8–11], there is no consensus on the best variant calling algorithm. The general recommendation is to take the intersection of the mutations called by a few methods, followed by additional filters [12, 13]. Such a strategy reduces false positive rate, but at the price of inflated false negative rate. Therefore it is important to account for somatic mutation calling uncertainty in association studies.

Such uncertainty of somatic mutation calling renders association methods for germline genetic variants inappropriate for somatic mutation associations. Generalized linear models are the most commonly used tools to assess germline genetic associations, for example, linear model for continuous traits and logistic regression for binary traits. Such methods do not account for mutation calling uncertainty. A few germline genetic association methods have been developed when the germline genomic features have inherent uncertainty, for example, for haplotype association [14, 15] or for case-control associations with systematic difference between cases and controls [16]. However, these methods are designed for specific problems and are not applicable to somatic mutation association studies.

A few earlier works have studied the associations between somatic mutations and gene expression using gene-level mutation [17], by integrating gene-gene interaction networks [18], or by a meta-analysis across multiple cancer types [19]. However, none of these works has considered the uncertainty of somatic mutation calling. In this paper, we propose a Somatic mutation Association test with Measurement Errors (SAME), which accounts for somatic mutation calling uncertainty by modeling the true somatic mutation status as a latent variable and exploiting read count data to augment the mutation calls. We develop two versions of this test, one for mutation-level analysis using a single somatic mutation (mSAME) and the other one for gene-level analysis using multiple mutations within a gene (gSAME). We have implemented SAME in an R package, and it is computationally efficient enough for genome-wide analysis. We evaluated the performance of SAME through extensive simulations and a real data application using the data from 12 cancer types of The Cancer Genome Atlas (TCGA) project. Our results demonstrated that SAME controls type I error and has improved statistical power compared to the competing methods that ignore somatic mutation calling uncertainty.

Materials and methods

mSAME model

We first describe the mSAME test that works on a single somatic mutation. To simplify notations, we omit the index for somatic mutations in the following discussions. For a specific somatic mutation, we denote the mutation call and true mutation status in the i-th sample by O_i and S_i, respectively, where $1 \leq i \leq n$ and n is sample size. S_i equals to 1 if this mutation is present in the i-th sample, and 0 otherwise. The value of O_i depends on the read-depth information. Let the read-depth and the number of alternative reads of this mutation in the i-th sample be D_i and A_i, respectively. A somatic mutation can be called only if there is enough coverage, i.e, $O_i = 0$ or 1 as mutation call indicator if $D_i \geq d_0$, and O_i is unobserved if $D_i < d_0$, where d_0 is a threshold used in the mutation calling process. Denote the outcome variable of the i-th sample by Y_i and the set of additional covariates by x_i. Let $\rho_0 = P(S_i = 0)$ and $\rho_1 = 1 - \rho_0 = P(S_i = 1)$, then the likelihood for the observed data can be written as

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^{n}\sum_{j=0}^{1}\rho_j f_{Y,A,D,O}(Y_i, A_i, D_i, O_i|S_i = j) \\ &= \prod_{i=1}^{n}\sum_{j=0}^{1}\rho_j f_Y(Y_i|S_i = j) f_{A,D,O}(A_i, D_i, O_i|S_i = j), \end{aligned} \tag{1}$$

where f_T denotes the density function for random variable T.

We further assume that the conditional distribution of Y_i given S_i (i.e., $f_Y(Y_i|S_i = j)$ in Eq (1)) can be modeled by a generalized linear model with mean

$$E(Y_i) = g^{-1}(x_i^T\alpha + S_i\beta), \tag{2}$$

and a dispersion parameter ϕ, where $g(\cdot)$ is a link function, and α, β are the regression coefficients. We are interested in the association testing problem H_0: $\beta = 0$. For continuous outcomes, we can write $f_Y(Y_i|S_i)$ as a normal density with the identity link function $g(t) = t$. For binary outcomes, we write $f_Y(Y_i|S_i)$ as a Bernoulli density using the logit link function $g(t) = \log(t/(1 - t))$.

For the distribution of read counts and observed mutation calls (i.e., $f_{A,D,O}(A_i, D_i, O_i|S_i = j)$ in Eq (1)), we use beta-binomial distributions to model allele-specific read counts A_i given D_i, O_i and S_i, and use a Bernoulli distribution to model O_i given S_i. Beta-binomial distributions have been used to model allele-specific read counts from ChIP-seq [20], RNA-seq [21], DNA sequencing [22], and somatic mutations [23, 24]. The Bernoulli likelihood of observed somatic mutation calls given true somatic mutation status has been used to model somatic mutation calls from single cell DNA sequencing data [25, 26]. These previous work have shown that these distributions are appropriate for real data. We have also compared the distributions of observed read counts versus expected ones from beta-binomial model fit and they agree very well (Fig S1 in S1 Appendix).

We denote the unknown parameters in the model by θ and the likelihood-ratio test statistic for the mSAME model is

$$T = -2[\log \mathcal{L}(\hat{\theta}_0; Y, A, D, O) - \log \mathcal{L}(\hat{\theta}; Y, A, D, O)], \tag{3}$$

where $\hat{\theta}$ is the maximum likelihood estimator of θ in the whole parameter space, and $\hat{\theta}_0$ is the maximum likelihood estimator of θ under H_0: $\beta = 0$. All the technical details for the likelihood function and parameter estimation can be found in Section 1.1-1.4 of S1 Appendix. Under H_0, the test statistic T asymptotically follows a Chi-square distribution with degree of freedom 1, thus we can reject H_0 if $T > \chi_1^2(1-\xi)$ where $\chi_1^2(1-\xi)$ is the $(1 - \xi)$ quantile of this Chi-square distribution.

gSAME model

Next we discuss our gSAME model that aggregates the information of multiple somatic mutation loci within a gene (or any arbitrarily defined unit) for association testing. We start by defining some notations. Suppose that there are p mutation loci within a gene of interest, and we drop the index for gene for notational convenience. We use superscripts ^{m} and ^{g} to denote mutation-level and gene-level data, respectively. We denote the observed mutation calls for the i-th sample by $O_i^m = \{O_{i1}^m, \cdots, O_{ip}^m\}$, the read-depth and the number of the alternative reads by $D_i^m = \{D_{i1}^m, \cdots, D_{ip}^m\}$ and $A_i^m = \{A_{i1}^m, \cdots, A_{ip}^m\}$, respectively. Analogously, we denote the underlying true mutation status by $S_i^m = \{S_{i1}^m, \cdots, S_{ip}^m\}$. We define the gene-level mutation status to be 1 if there is one or more mutations within this gene:

$$S_i^g = \begin{cases} 1 & \text{if any } S_{ij}^m = 1, \\ 0 & \text{if all } S_{ij}^m = 0. \end{cases} \tag{4}$$

The outcome variable Y_i and the covariates x_i are defined as before. In gene-level analysis, we model Y_i as a function of S_i^g and x_i. Then the likelihood function is

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^{n} \sum_{j=0}^{1} \rho_j^g f_{Y,A,D,O}(Y_i, A_i^m, D_i^m, O_i^m | S_i^g = j) \\ &= \prod_{i=1}^{n} \sum_{j=0}^{1} \rho_j^g f_Y(Y_i | S_i^g = j) f_{A,D,O}(A_i^m, D_i^m, O_i^m | S_i^g = j), \end{aligned} \tag{5}$$

where $\rho_0^g = P(S_i^g = 0)$ and $\rho_1^g = 1 - \rho_0^g = P(S_i^g = 1)$.

Since read count data (i.e., D_i^m and A_i^m) and mutation calls (O_i^m) are collected for each mutated locus, their distributions are modeled given S_i^m. Then the remaining steps to complete this likelihood is to model S_i^m conditional on S_i^g. When $S_i^g = 0$, it is clear that $S_{ij}^m = 0$ for all the p mutations. When $S_i^g = 1$, S_i^m can have $2^p - 1$ possible values, which is computationally onerous to enumerate for large p. We notice that in practice, it is impossible to call a somatic mutation if the corresponding number of alternative reads equals to 0. Hence to reduce computational burden, we assume that the j-th mutation may occur only if $A_{ij}^m > 0$, otherwise we assign $S_{ij}^m = 0$ directly. Thus the number of the combinations is limited to $2^{m_i} - 1$, where m_i is the number of mutations with $A_{ij}^m > 0$.

Let θ^g be the unknown parameters in the gSAME model. The likelihood ratio test statistic of gSAME model for testing the effect of somatic mutation S_i^g is

$$T = -2[\log \mathcal{L}(\hat{\theta}_0^g; Y, A^m, D^m, O^m) - \log \mathcal{L}(\hat{\theta}^g; Y, A^m, D^m, O^m)], \tag{6}$$

where $\hat{\theta}^g$ is the estimator of θ^g in the whole parameter space, and $\hat{\theta}_0^g$ is the estimator of θ^g under H_0. All the technical details for the likelihood function and parameter estimation can be found in Section 1.5-1.6 of S1 Appendix.

Results

Simulation studies

Simulations for mSAME model. We generated a dataset of $n = 400$ samples. For the i-th sample, we simulated the true somatic mutation value S_i by a Bernoulli distribution with success probability ρ_1, and we vary ρ_1 in different simulation setups. A continuous outcome Y_i was simulated by $Y_i = 1 + x_i + \beta S_i + \epsilon_i$, where x_i and ϵ_i were generated by the standard normal distribution independently. A dichotomous outcome Y_i was simulated from a Bernoulli distribution with success probability p_i, and $\log(p_i/(1 - p_i)) = -0.5 + x_i + \beta S_i$. Based on the true mutation value S_i, we simulated the observed mutation O_i by the Bernoulli distributions specified in equation (3) of S1 Appendix, with the sensitivity and specificity being $\gamma_1 = 0.9$ and $\gamma_0 = 0.98$, respectively. These choices of sensitivity and specificity are based on the evaluation of somatic mutation callers in a previous study [12]. It is desirable to generate O_i by actually performing somatic mutation calling. However, we did not pursue on this direction because it would require generation of bam files and simulating many factors, such as sequencing quality scores, mapping quality scores, clustering of reads due to amplification artifacts, and strand bias, and we are not aware of any existing tool to generate such bam files.

We simulated read-depth of somatic mutations to mimic the read-depth data observed in a TCGA exome-seq dataset (across 133,463 somatic mutations in 433 colon cancer patients). More details of this dataset are presented in the following sections on TCGA data analysis and S3 Appendix. In exome-seq data, read-depth varies across genomic loci and across samples, and thus we simulated read-depth in two steps. First, we simulated mean read-depth for each mutation by a negative binomial distribution with mean $\mu = 113$ and over-dispersion $\phi = 3.28$, so the standard deviation is $\sqrt{\mu + \mu^2/\phi} \simeq 63.3$. Next, for each mutation, we simulated read-depth across samples by a negative binomial distribution with mean value simulated in the first step, and over-dispersion 1.9. When $D_i \geq d_0 = 20$, we simulated A_i by a beta-binomial distribution specified in equation (2) of S1 Appendix, with parameters $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}) = (0.001, 0.002, 0.1179, 0.3207)$ and $(\varphi_{00}, \varphi_{01}, \varphi_{10}, \varphi_{11}) = (0.0006, 0.3457, 0.0001, 0.1018)$. Later in simulation studies, we estimate these parameters based on 50 simulated somatic mutations across 400 samples, and the estimates are fairly accurate. When $D_i < 20$, the number of alternative reads A_i was generated by a beta-binomial distribution (see equation (4) of S1 Appendix) with parameters $\pi_0 = 0.001$, $\varphi_0 = 0.001$, $\pi_1 = 0.146$, and $\varphi_1 = 0.10$. The parameters of these negative binomial and beta binomial distributions are all estimated from the TCGA colon cancer dataset.

Using this simulated dataset, we compare the performance of mSAME and a naive generalized linear model, in terms of type I error and power for testing the hypothesis $\beta = 0$. The generalized linear model does not account for somatic mutation calling uncertainty, but simply treats the observed somatic mutation call O_i as the true somatic mutation status and performs a Wald test on the regression coefficient β.

Across different simulation settings, we considered various mutation frequencies $\rho_1 = 0.02$, 0.05, 0.10 and effect size β, and evaluate the performance over 1,000 replicates. We set $\beta = 0$ to evaluate the type I error at the significance level 0.05. For the power performance, we set $\beta = 0.2, 0.4, 0.6, 0.8, 1.0$ for the continuous trait and $\beta = 0.4, 0.8, 1.2, 1.6, 2.0$ for the binary trait. In all the scenarios, the type I errors of both methods are well controlled, and the mSAME has higher power than GLM for all simulation settings and for both continuous and binary traits (Fig 1, Table S1 in S2 Appendix). In addition, mSAME has more accurate estimates of β, evaluated by the mean square error (MSE) (Fig S2 in S2 Appendix).

Simulations for gSAME model. For the gene-level mutation analysis, we are interested in the association between an outcome Y and a gene-level mutation S_i^g. We assume that there are $p = 10$ mutations within the gene, and denote the frequency of the gene-level mutation as $P(S_i^g = 1) = \rho_1^g$. We set the sample size $n = 400$. For the i-th sample, we generated the true mutation values $S_{ij}^m, j = 1, \cdots, p$ independently by a Bernoulli distribution with $P(S_{ij}^m = 1) = 1 - (1 - \rho_1^g)^{1/p}$. Then the gene-level mutation value S_i^g can be obtained by collapsing mutation level data (Eq (4)). The continuous outcome variable Y_i was simulated by $Y_i = 1 + x_i + \beta S_i^g + \epsilon_i$, where x_i and ϵ_i were simulated by the standard normal distribution independently. The binary outcome Y_i was simulated from a Bernoulli distribution so that $\text{logit}[p(Y_i = 1)] = -0.5 + x_i + \beta S_i^g$. In addition, for the j-th mutation, the observed mutation call O_{ij}^m, the read-depth D_{ij}^m and the number of alternative reads A_{ij}^m were simulated based on the true mutation value S_{ij}^m, following the same procedure as the mutation-level simulations.

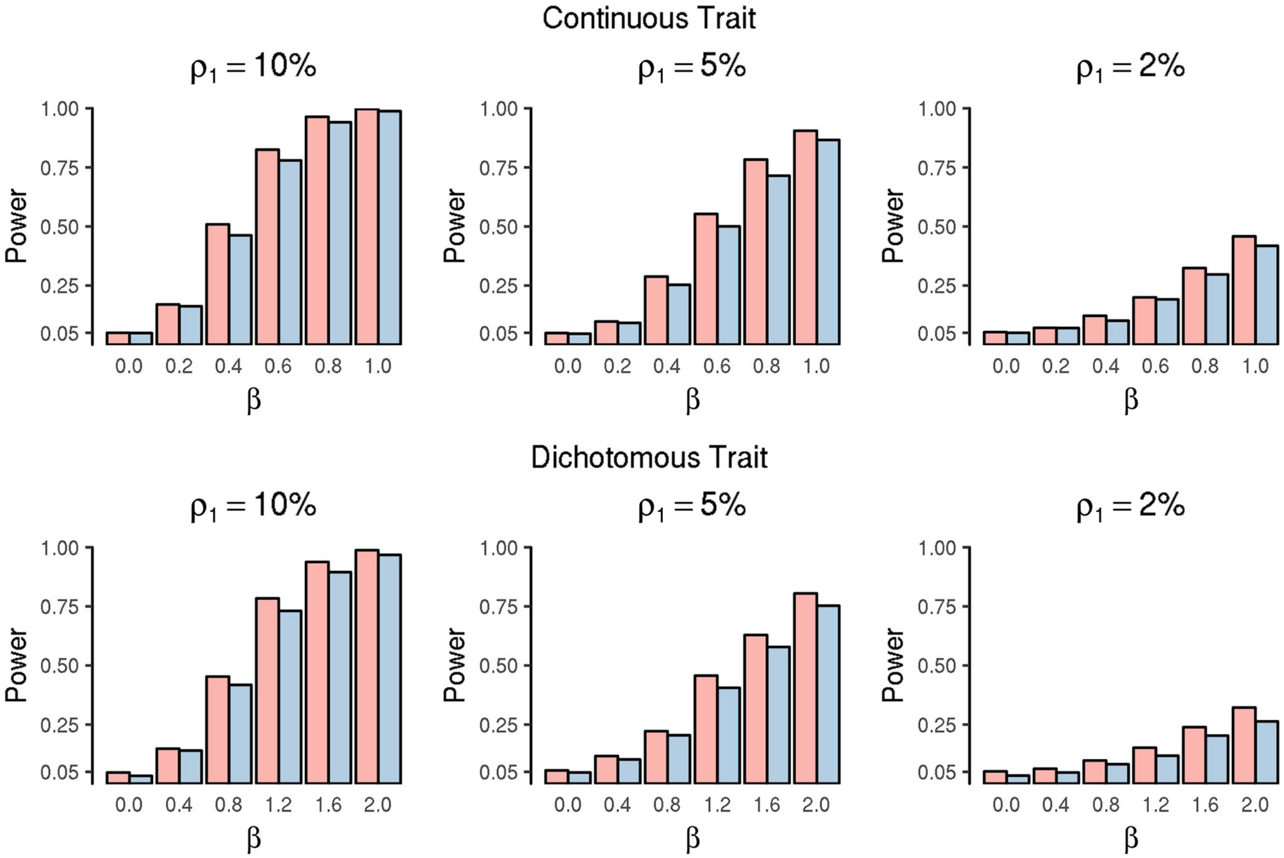

Fig 1. Power comparison of mSAME (red bars) and GLM (blue bars) for mutation-level simulations.

https://doi.org/10.1371/journal.pgen.1007746.g001

When simulating O_{ij} within a gene, we randomly chose the specificity to be 0.98 or 1 with equal probabilities and randomly chose the sensitivity to be 0.9 or 1 with equal probabilities.

We compared the performance of gSAME with a naive GLM method. For the naive GLM method, we regress the response Y_i on x_i and the observed gene-level somatic mutation O_i^g, where O_i^g is defined as

$$O_i^g = \begin{cases} 1 & \text{if any } O_{ij}^m = 1, \\ 0 & \text{if all } O_{ij}^m = 0. \end{cases} \tag{7}$$

For mutation-level associations, we have considered the mutation frequencies of $\rho_1 = 0.02$, 0.05, or 0.10. Since the gene-level mutation frequencies are usually higher than mutation level mutation frequencies, for gene-level mutation, we set $\rho_1^g = 0.05, 0.10$, or 0.15. The regression coefficient β was set to be the same as in the mutation-level analysis. All the results were evaluated over 1,000 replicates. Overall, both gSAME and GLM control the Type I error, and gSAME always has higher power than GLM (Fig 2, Table S2 in S2 Appendix), and more accurate estimates of β (Fig S2 in S2 Appendix). Given the same mutation frequency, the power of gene-level analysis is lower than mutation-level analysis because the mutation-level measurement errors aggregate and become larger at gene-level.

In conclusion, SAME has higher power than the naive GLM approach that ignores mutation calling uncertainty, even if the mutation calling is relatively accurate (sensitivity 0.9 and specificity 0.98) and read-depth is relatively high (average read-depth of 113). This is because

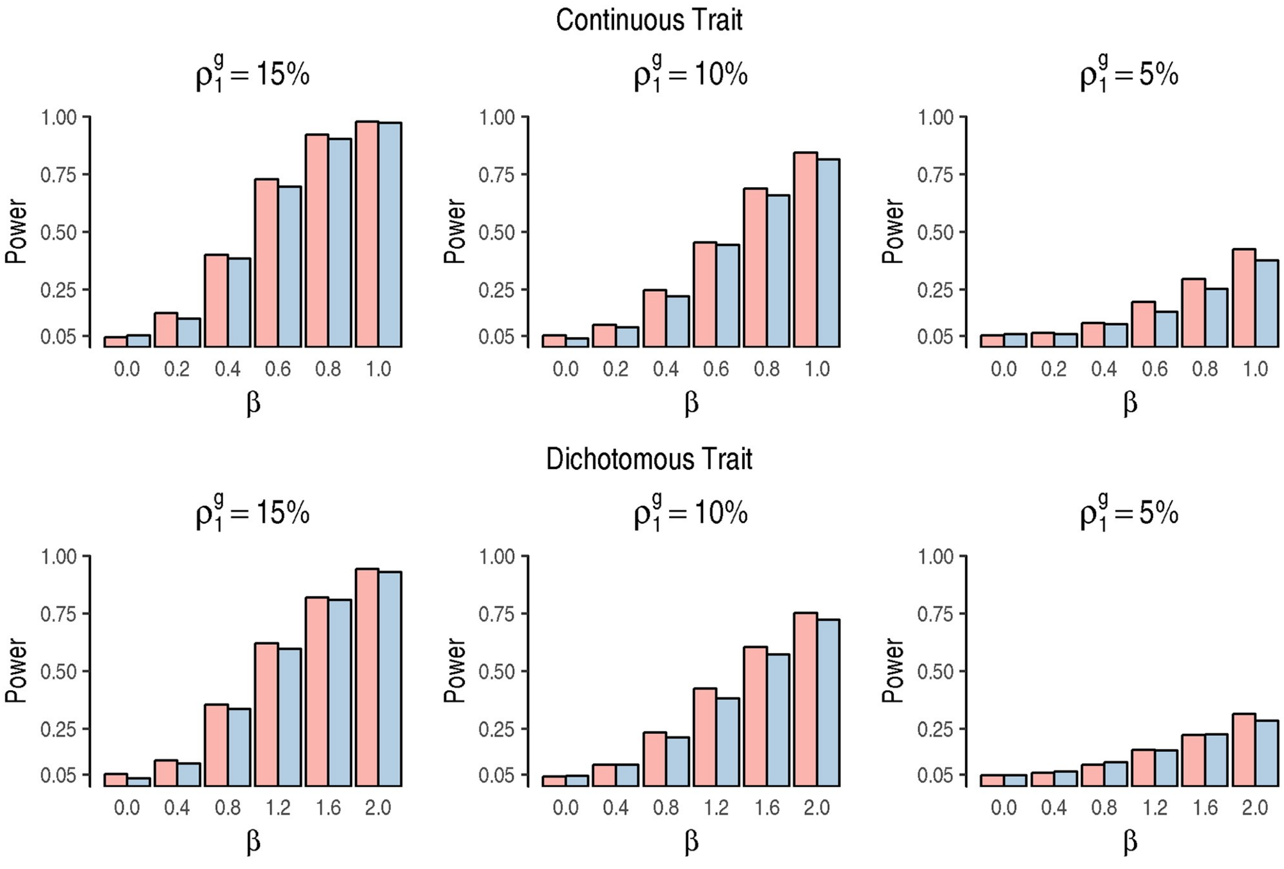

Fig 2. Power comparison of gSAME (red bars) and GLM (blue bars) for gene-level simulations.

https://doi.org/10.1371/journal.pgen.1007746.g002

our model accounts for the imperfect sensitivity and specificity, and read-depth can be low for some genes in some samples, due to uneven coverage of exome-seq data (Fig S3 in S2 Appendix). When the mutation calling becomes less accurate (e.g., sensitivity 0.9 and specificity 0.95) or the read-depth becomes lower (e.g., for whole genome sequencing data, where the typical read-depth is 20x to 40x), SAME has even larger power gain than GLM. For example, in an additional simulation setup, we simulated read-depth by a negative binomial distribution with mean 40 and over-dispersion 1.9, resembling a whole-genome sequencing situation. For a continuous trait, with mutation frequency 5% and effect size $\beta = 0.2$ or 0.4, the power gain of mSAME vs. GLM is around 40% for this whole-genome sequencing simulation setup (Fig S4 in S2 Appendix). In contrast, our main simulation setup resembles an exome-seq data where the power gain is around 10%. See S2 Appendix for more details of additional simulation setup and results.

TCGA colon cancer eQTL analysis

We applied the proposed mSAME and gSAME methods as well as GLM to study the associations between somatic mutations and genome-wide gene expression in TCGA colon cancer patients. Briefly, we downloaded the bam files of exome-seq data for paired tumor-normal samples from NCI Genomic Data Commons (GDC) Data Portal. We called somatic mutations using the intersection of MuTect and Strelka, followed by the read-depth filter to keep those mutations with read-depth $\geq$20 in both tumor and paired-normal samples (Section 3.1 in S3 Appendix). Colon cancer patients can be separated into two subtypes based on mutation load [27]. We classified a sample as hyper-mutated if it has more than 375 non-silent mutations and this cutoff is chosen to separate the two modes of the distribution of mutation load (Fig S9 in S3 Appendix). Our analysis requires allele-specific read counts for each mutation across all samples. While collecting such information, we noticed that 24 samples have much smaller number of allele-specific read counts than the remaining samples and we removed them from our data analysis (Section 3.3 in S3 Appendix).

For gene expression data, we downloaded the `.htseq.counts` files from NCI GDC, which include the number of RNA-seq reads mapped to 60,483 genomic features. Most of these features are non-coding RNAs or pseudo genes that have zero or very small number of RNA-seq read counts across most tumor samples. We selected 17,986 genes that have at least 20 reads in more than 25% of the samples for the down-stream analysis. Let T_{ij} be the read count for the j-th gene in the i-th sample. We correct for read-depth variation across samples using T_{ij}/d_i, where d_i is the 75 percentile of gene expression within the i-th sample, a robust measurement of read-depth [28]. Then we quantified gene expression by $\log(T_{ij}/d_i)$, to make variation of gene expression similar across orders of expression levels [29]. We further regressed out copy number effect from gene expression data (Section 3.4 in S3 Appendix). Since copy number measurement may be missing for some genes (usually the genes around the beginning/end of a chromosome or around a centromere), we removed genes with missing copy number information, and ended up with 16,339 genes for the following analysis.

Taking the intersection of the samples with somatic mutations and gene expression, we obtained 386 samples. We further included age, gender, and hyper-mutation status as covariates. We also removed those potential germline mutations by checking the read-depth data in the paired normal samples (Section 3.5 in S3 Appendix). In the following analysis, we only studied non-silent mutations because silent mutations in exonic regions are most likely to be passenger mutations that do not have functional impact.

mSAME results. For the mutation-level association analysis, we selected 37 mutations which have occurred in at least 5 of the 386 samples, corresponding to a mutation frequency of

1.3%. For the association analysis between these 37 mutations and all the 16,339 genes, the Bonferroni correction was adopted for multiple testing correction, i.e., we rejected the null hypothesis if the p-value was less than $0.05/(37 \times 16339) \approx 8.27 \times 10^{-8}$.

We applied both mSAME and GLM to assess the associations between the somatic mutations and gene expression. Recall that we model alternative read count by a beta-binomial distribution, and the parameters of this distribution need to be estimated *a priori*. We estimated these parameters using the 3, 359 mutations used in gene-level analysis since more mutations can be included in gene-level analysis. In addition, the sensitivity and specificity for each mutation were estimated as described in Section 1.3 of S1 Appendix.

At the significance level with Bonferroni correction, mSAME identified 109 significant associations while GLM detected 100 significant associations that is a subset of the 109 associations identified by mSAME (Table S7 in S3 Appendix). Most of these significant associations (100 out of 109) are with respect to the BRAF V600E mutation (chr7:140753336), which is a single nucleotide variant that results in an amino acid change from a valine (V) to a glutamic acid (E). The high frequency of BRAF V600E mutation associations is partly due to its high mutation frequency (11.66%). In contrast, the secondly most frequently mutated locus, which is located in gene PIK3CA, is observed in only 3.63% of the samples (Fig S11 in S3 Appendix). Since this mutation has no detectable calling errors, the p-values of mSAME are in line with those of GLM in general.

In total, mSAME identified 9 additional findings that were missed by GLM. Here we briefly discuss two interesting examples and list all of them in Table 1. The first example is that the TP53 mutation "chr17:7673803" is associated with the gene expression CDX1. Previous work has indicated that the gene expression of CDX1 is abnormally down-regulated in colon cancer-derived cell lines [30, 31], and our finding suggests that this TP53 somatic mutation is partly responsible for dysregulation of CDX1's expression in colon cancer. The second example is that TP53 mutation "chr17:7674894" associated with its own gene expression.

gSAME results. We collapsed mutations within the same gene and obtained 17,386 gene-level mutations. Among these mutations, we conducted the association analysis for 180 genes that are mutated in at least 5 samples and are known to be associated with colon cancer. In total, these 180 gene-level mutations correspond to 3,359 individual mutations. We applied gSAME and GLM for all the 180 × 16, 339 tests, and uses Bonferroni corrected significance level of 1.70×10^{-8}.

At this significance level, gSAME and GLM both identified 63 significant associations where 59 of them are in common, and hence 67 associations in total (Table S8 in S3 Appendix). Gene-level mutation status of BRAF is associated with the expression of 36 genes and all

Table 1. Nine significant results detected uniquely by mSAME. An eQTL is a local eQTL if the distance between the somatic mutation and the gene is smaller than 1Mb. Otherwise it is a distant eQTL.

Mutation(Gene)	Associated Gene	eqtl type	mSAME	GLM
chr3:25627236(TOP2B)	C4orf19	distant	6.98e-8	8.64e-8
chr5:112838007(APC)	ZWILCH	distant	5.71e-9	8.58e-6
chr7:74824936(GTF2IRD2)	SDR42E1	distant	3.02e-10	1.30e-5
	LINC00675	distant	1.47e-8	1.02e-3
chr7:140753336(BRAF)	EPM2AIP1	distant	6.49e-8	8.44e-8
	LRRC19	distant	8.27e-8	9.72e-8
	ETV5	distant	6.91e-8	8.50e-8
chr17:7673803 (TP53)	CDX1	distant	1.77e-9	3.46e-2
chr17:7674894 (TP53)	TP53	local	3.30e-8	2.12e-7

https://doi.org/10.1371/journal.pgen.1007746.t001

of these associations have been identified in mutation-level analysis with respect to the V600E mutation. This may not be surprising because V600E mutation is present in more than 80% of the samples with at least one BRAF mutation.

Another gene-level mutation that is associated with the expression of several genes is gene-level mutation of TP53. Each of the 68 individual mutations within TP53 has relatively low mutation frequency (the highest frequency is 5.70%), however, after aggregating all the mutations, the gene-level mutation TP53 is present in 39.38% of the samples. The expression of 11 genes are associated with TP53 gene-level mutation (including two gSAME-specific findings and one GLM-specific finding). Using the DAVID Tools for the gene enrichment analysis on these 11 genes (https://david.ncifcrf.gov/), we found that the following 4 genes are in the KEGG p53 signaling pathway: FAS, MDM2, DDB2, ZMAT3 (with enrichment p-value 3.1e-5 after Benjamini correction). Among them, FAS and ZMAT3 were only detected by gSAME (Table S8 in S3 Appendix). Intrigued by this functional enrichment, we further explored the gene-level associations for TP53 at a more liberal p-value cutoff of $0.05/16339 \approx 3.06 \times 10^{-6}$. gSAME and GLM both detected 27 associated genes, while 22 of them are in common. Among these 32 genes, the following seven genes belong to the KEGG p53 signaling pathway: BAX, FAS, MDM2, CDKN1A, DDB2, TP53I3, ZMAT3 (with enrichment p-value 3.3e-8 after Benjamini corretion), where BAX and TP53I3 are detected only by gSAME but missed by GLM.

The complete list of mutation level and gene level eQTL results can be found in two text files as in S1 and S2 Files.

TCGA colon cancer subtype analysis

To illustrate somatic mutation association analysis using dichotomous outcomes, we applied both mSAME and gSAME to identify somatic mutations associated with colon cancer subtypes defined by DNA methylation data. One of the most well known subtype of colon cancer is the hypermutation subtype [4, 27]. By definition, it is associated with many somatic mutations and thus we used it as a covariate in all the analysis of this paper. Here we consider another subtype, defined by clustering analysis of genome-wide DNA methylation data [32] (Fig S13 in S3 Appendix). See Section 3.8 in S3 Appendix for details of methylation data processing. We used this clustering results to classify the cancer patients into two groups and treated it as a binary outcome. Then we associated this subtype indicator with somatic mutations.

Similar to the eQTL analysis, we performed mutation-level association analysis using mSAME and GLM (logistic regression) on 37 mutations that are present in at least 5 samples. At the significance level $0.05/37 \approx 0.00135$, mSAME and GLM both detected one significant mutation of BRAF V600E, where mSAME yields a smaller p-value than GLM (Table 2). We also performed gene-level analysis by gSAME and GLM for the 180 gene-level mutations used in eQTL analysis. Both methods discovered two significant gene-level mutations: BRAF and KMT2C, using the p-value threshold $0.05/180 = 0.00028$. KMT2C is known as a tumor suppressor gene [33]. Our results suggest that the mutations of KMT2C are associated with DNA

Table 2. Summary for the association study of the subtypes on mutation level (top table) and gene level (bottom table).

Mutation	Gene	mSAME	GLM
chr7:140753336	BRAF	5.91e-8	5.74e-6
Mutation	chr	gSAME	GLM
BRAF	7	4.41e-5	8.50e-5
KMT2C	7	2.44e-4	7.05e-4

https://doi.org/10.1371/journal.pgen.1007746.t002

methylation, which is consistent with its role as histone methyltransferases because DNA methylation and histone methylation often work together to establish epigenetic landscape for gene expression regulation.

eQTL analysis for pan-cancer studies

Following the workflow of the eQTL analysis for TCGA Colon Adenocarcinoma (COAD) samples, we conducted eQTL analysis using somatic mutations for 11 other TCGA cancer types, including Bladder Urothelial Carcinoma (BLCA), Brain Lower Grade Glioma (LGG), Glioblastoma multiforme (GBM), Head and Neck squamous cell carcinoma (HNSC), Kidney renal clear cell carcinoma (KIRC), Liver Hepatocellular Carcinoma (LIHC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Ovarian serous cystadenocarcinoma (OV), Skin Cutaneous Melanoma (SKCM), and Stomach adenocarcinoma (STAD). These cancer types are chosen due to their relatively large sample sizes and relatively higher rates of somatic mutations.

We dowloaded the gene expression and somatic mutation data for association analysis from NCI GDC Data Portal, using the workflow of "HTSeq—Counts" for gene expression data, and the workflow of "MuTect2 Variant Aggregation and Masking" for somatic mutation data. For mutation-level association analysis, we selected the mutations that occur in at least 5 samples. For gene-level analysis, we selected the gene-level mutations that occur in at least 5% of the samples. For each mutated locus, we need read count data (read depth and the number of alternative reads) for all samples, regardless of mutation call status. For COAD analysis, we downloaded all the bam files to local server and then collected these counts. However, this approach is not feasible for pan-cancer study across 11 cancer types because downloading and storing all the bam files requires too many resources. Instead, we obtained the read-count data using the cloud service provided by The Seven Bridges Cancer Genomics Cloud [34] (Section 3.9 of S3 Appendix).

In all association studies, we included age and gender (except for gender-specific cancer PRAD and OV) as covariates. For LGG, we further adjusted for cancer subtype defined based on the IDH1 or IDH2 mutation and chromosome 1p and 19q co-deletion [35]. We recorded significant findings using genome-wide Bonferroni correction, and summarized the number of the significant findings by GLM or SAME in Table S9 in S3 Appendix (Section 3.10 of S3 Appendix). The complete lists of the results are provided as supplementary text files. Examining the number of significant eQTLs for each mutation or each gene across cancer types shows no apparent pattern: most mutation-level or gene-level eQTLs are not shared across cancer types. However, one exception is gene-level TP53 mutation (Fig 3), which is among the significant eQTLs in 7 out of the 12 cancer types. This is partly due to the fact that TP53 is mutated with relatively high frequency across cancer types and it is a transcription factor that can directly regulate gene expression. When we relax the p-value cutoff to use transcriptome-wide significance (i.e., p-value cutoff = 0.05/# of genes), gene-level TP53 eQTLs were identified in 9 cancer types. In addition, several other gene-level eQTLs are shared among multiple cancer types (Fig S14 in S3 Appendix). Overall the pattern of mutation/gene eQTLs shared across cancer types are similar between SAME and GLM (Fig S15-S16 in S3 Appendix), though in general mSAME/gSAME identify more eQTLs than GLM.

Next we examine the eGenes (genes whose expression are associated with an eQTL) associated with TP53 gene level mutation across cancer types. Since we focus on one mutation, we select the eGenes identified by transcriptome-wide significance. At this significance level, TP53 has no eGene by either gSAME or GLM in three cancer types: KIRC, LUSC and HNSC, and thus the following results only involve the remaining nine cancer types. We are interested

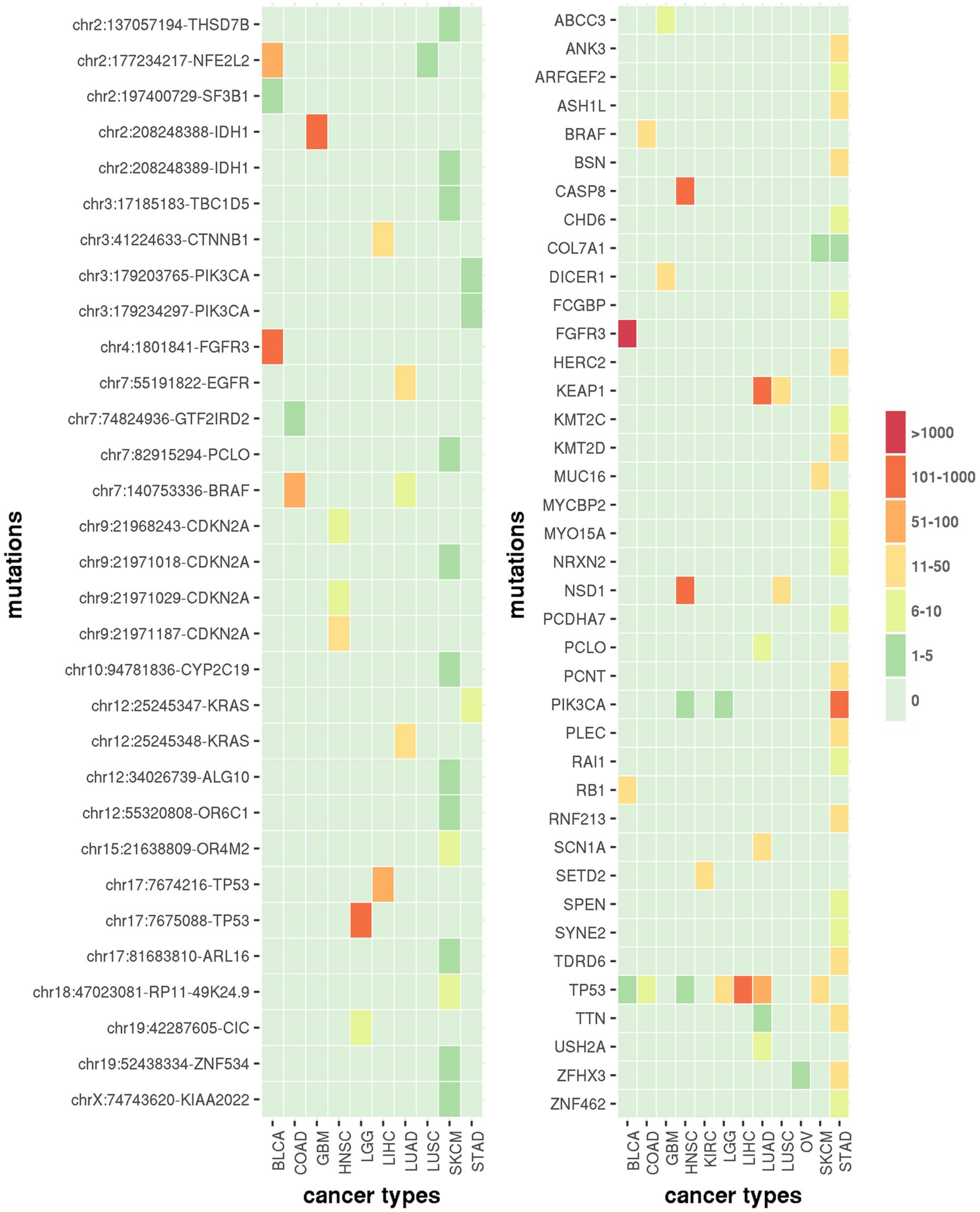

Fig 3. Summary of pan-cancer eQTL mapping results by SAME with Bonferroni multiple testing correction. Left panel: a heatmap of mutation-level eQTL mapping results by mSAME across cancer types. Each cell in the heatmap is colored according to the number of significant associations for one mutation (row) in one cancer type (column). Only those mutations that are associated with 2 or more genes across the 12 cancer types are shown. Note that OV cancer is not included since there is no significant eQTL in OV. Right panel: a heatmap of gene-level eQTL mapping results by gSAME across cancer types. Only those gene-level mutations that are associated with 6 or more genes across the 12 cancer types are shown.

https://doi.org/10.1371/journal.pgen.1007746.g003

in similarities of TP53 eGenes across cancer types. Towards this end, we examine the 50 eGenes identified in at least 3 cancer types by either gSAME (35 eGenes) or GLM (46 eGenes), with an intersection of 31 genes identified by both methods (Fig 4). The difference of gSAME and GLM results are most due to potential mutation calling errors in TP53 (Fig S17 in S3 Appendix).

The protein product of TP53, p53, is a very well studied tumor suppressor and is involved in different biological processes such as cell cycle arrest, DNA repair, and apoptosis [36]. Many target genes of p53 have been reported [37], and these target genes can be used to evaluate the relevance of the eGenes identified from our study. About 29% (10 out of 35) of the eGenes identified by gSAME and 20% (9 out of 46) identified by GLM are among 343 high confidence p53 target genes [37] (Fig 4). The only difference is gene CDKN1A (encoding protein p21) where gSAME and GLM identified it as an eGene for three and two cancer types, respectively. CDKN1A is one of the most important targets of p53 and is requested for p53-mediated cell cycle arrest [37].

Visualization of the mutation status of these 50 genes show an interesting pattern: three cancer types, LUAD, LIHC and SKCM are clustered together since many genes are eGenes only in these three cancer types (Fig 4). The relatively larger number of eGenes in these cancer types can not be explained by the mutation frequency of TP53 (Fig S18 in S3 Appendix) or genome-wide somatic mutation load (Fig S19 in S3 Appendix). None of these eGenes are among the 343 high confidence p53 target genes, suggesting that they may be indirectly regulated by p53. Gene ontology analysis shows that these eGenes are enriched with genes involved in cell cycle related biological processes such as chromosome segregation. Therefore our results suggest that somatic mutation of TP53 may have similar functional roles in cell cycle control in LUAD, LIHC and SKCM.

Discussion

Understanding the associations between somatic mutations and cancer-related traits is of fundamental importance for precision cancer therapy. In this paper, we present a statistically powerful and computationally efficient approach for association analysis of somatic mutations while accounting for measurement errors of somatic mutations. By modeling the calling uncertainty of the somatic mutations and including the read-depth data into our statistical model, the proposed SAME method can significantly improve the statistical power for the association analysis. The SAME method can accommodate both continuous and dichotomous outcomes, and it is applicable to both mutation-level and gene-level association testing. While we have demonstrated SAME using the publicly available exome-seq data, it will provide larger degree of power gain for whole genome sequencing studies where read depth are typically lower.

One practical question of using our method is that how to choose between mutation-level (mSAME) versus gene-level (gSAME) analysis. Our eQTL analysis results suggest that mSAME may be more suitable for recurrent mutations in oncogene (e.g., the BRAF V600E mutation). This is because an oncogene is often activated by some specific “gain of function” mutations that drive tumor growth, and such driver mutations are often recurrent across patients. Other rare mutations in the same gene may be passenger mutations, even if they are non-silent ones. For example, BRAF harbors 10 non-silent mutations in TCGA colon cancer dataset. Except for the V600E mutation, the remaining 9 mutations only occur in one or two samples, and thus are likely passenger mutations. When collapsing both driver and passenger mutations to create a gene-level mutation, the mutation pattern may be “diluted” by those passenger mutations, and thus gene-level associations may yield less significant results than

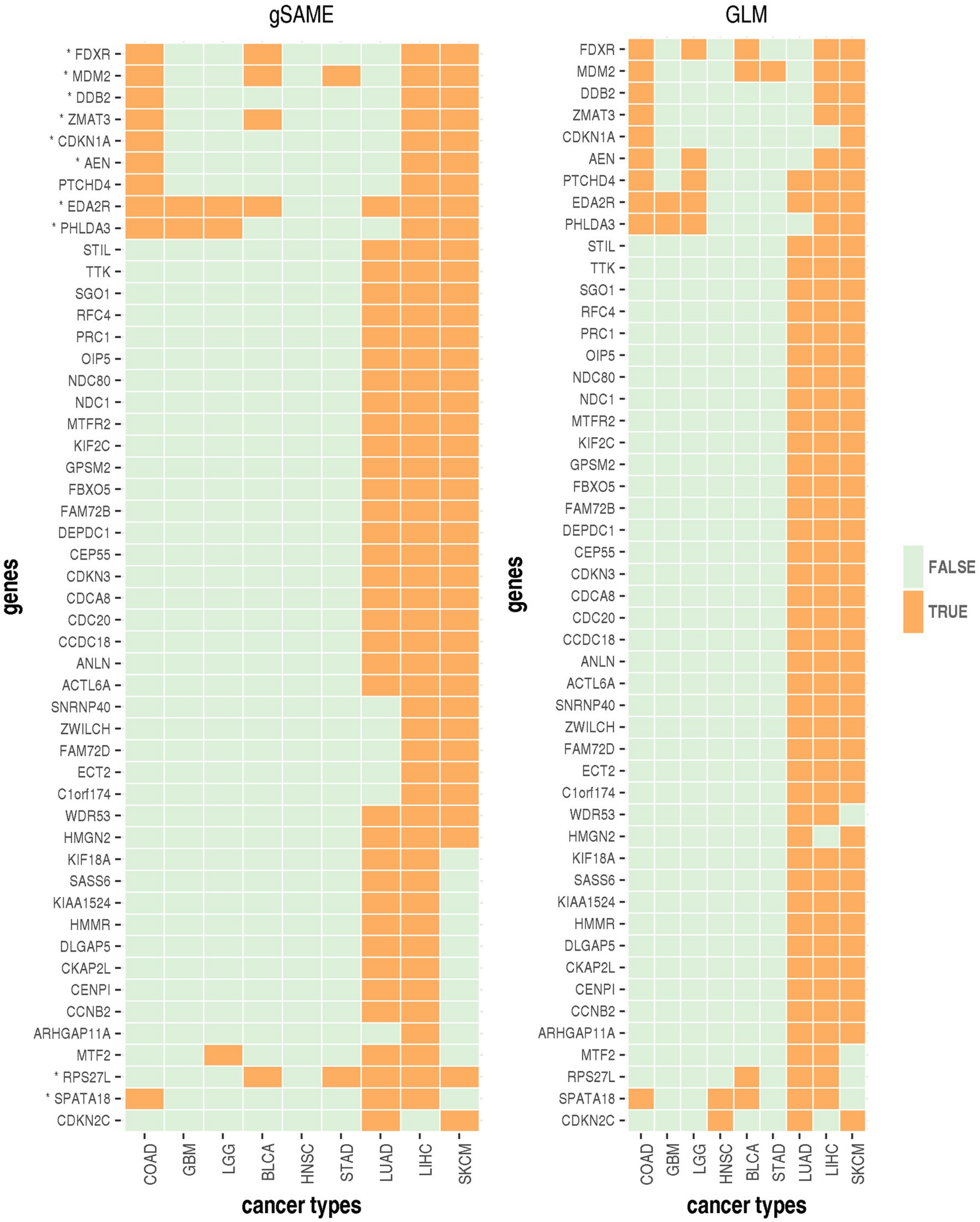

Fig 4. Summary of eGenes of TP53 with transcriptome-wide multiple testing correction. The results of gSAME (left panel) and GLM (right panel) were summarized by heatmaps showing whether a gene is eGene across cancer types. Only the genes that are eGenes for three or more cancer types by either method are shown. The asterisk (*) next to gene symbol indicates a gene is a high confidence TP53 target gene [37].

https://doi.org/10.1371/journal.pgen.1007746.g004

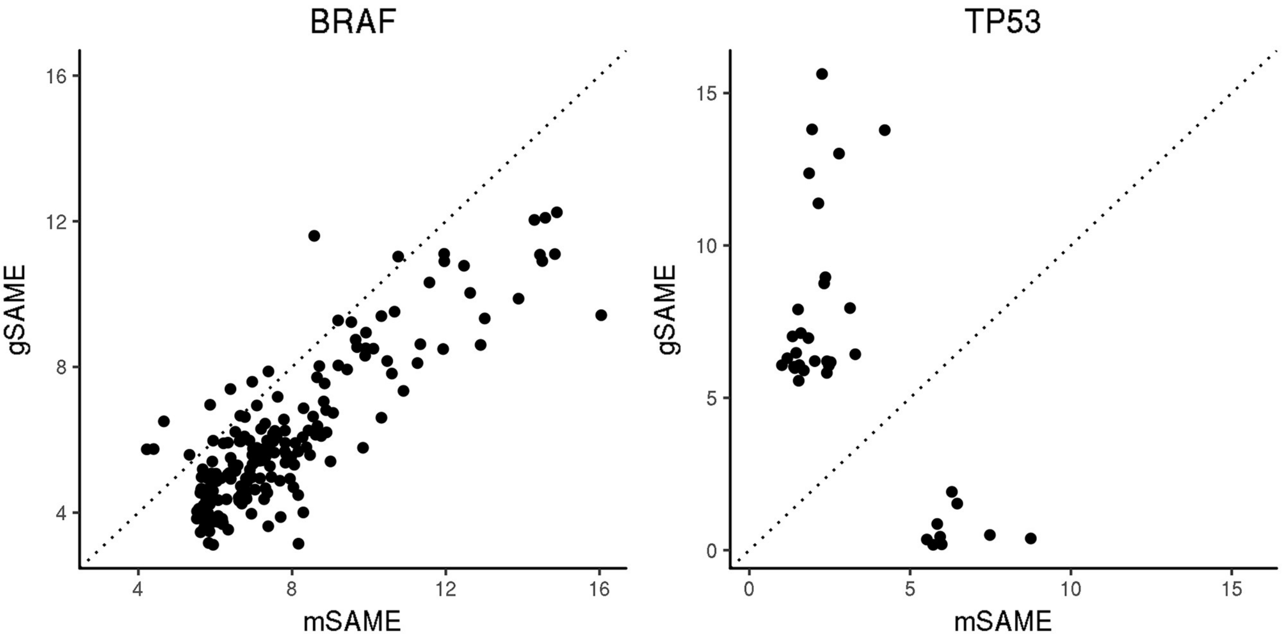

Fig 5. Comparison of eQTL mapping using mSAME p-values versus gSAME p-values on -log10 scale in colon cancer patients. Left panel: mutation-level versus gene-level analysis for the oncogene BRAF for all gene expression traits with eQTL p-value smaller than 0.05/16, 339 by either mSAME or gSAME. Right panel: similar results for the tumor suppressor gene TP53. Dashed line indicates $y = x$.

https://doi.org/10.1371/journal.pgen.1007746.g005

mutation-level associations. This is indeed the pattern observed when we compare the eQTL results for BRAF V600E mutation versus BRAF gene level mutations (Fig 5).

In contrast, gSAME may be more suitable for tumor suppressor gene (e.g., TP53). The function of a tumor suppressor gene may be perturbed by multiple "loss of function" mutations and thus there is no evolutionary pressure to select a specific one. Since all the loss of function mutations have similar functional consequence, gene-level association can have much larger power than mutation-level analysis. For example, TP53 has 68 individual mutations in TCGA colon cancer dataset, among which only 6 mutations occur at more than 2% of the samples and are significant eQTLs with transcriptome-wide multiple testing correction. For each gene expression trait, we take the minimum mutation-level p-value across these 6 mutations and compare it with gene-level p-value. In most cases, the gene level analysis yields stronger associations than mutation-level analysis (Fig 5).

We have carefully implemented mSAME/gSAME to maximize computational efficiency, so that it is computationally feasible for genome-wide eQTL mapping. However, it still takes about 1-5 seconds per association testing. In contract, GLM is computationally much more efficient, taking about 0.01-0.02 seconds per association testing. Therefore, when there is limited mutation calling error (e.g., with high quality samples and high sequencing coverage) one strategy to balance computational time and accuracy is to use GLM for a quick initial scan, and then apply mSAME/gSAME for a subset of associations at a relatively liberal p-value cutoff. In addition, gSAME will become computationally more inefficient for larger analysis units, such as several genes within a pathway. Further development is needed in such situations. In fact, simply collapsing individual mutations may not be a good strategy for pathway level association analysis and better strategies to summarize pathway level somatic mutations warrant further studies.

Somatic mutation association is a new field with great potential to deliver key findings for precision cancer therapy. Accounting for somatic mutation calling uncertainty and low read-depth is an initial step to develop more rigorous and powerful association methods. We expect that more methods will be developed to exploit other types of information, such as intra-tumor heterogeneity or pathway level analysis where mutation information across genes is aggregated.

Supporting information

S1 Appendix. Supplementary methods. It includes details for (1) Estimation of beta-binomial distributions when read-depth is high; (2) Estimating the specificity and sensitivity of an individual somatic mutation; (3) Details of the EM algorithm for mSAME model; (4) Estimation of beta-binomial distributions when read-depth is low for gene-level associations; (5) Details of the EM algorithm for gSAME model.
(PDF)

S2 Appendix. Additional simulation results. It includes the estimation error comparison of β by the EM algorithm and GLM, and additional simulations when the mutations calling becomes less accurate or the read-depth becomes lower compared with the simulations in the main text.
(PDF)

S3 Appendix. Additional methods/results for real data analysis. It includes details for the data processing for the real data analysis: (1) Pre-processing and somatic mutation calling; (2) Mutation load and hypermutation status; (3) Allele-specific read counts; (4) Removing copy number effect from gene expression data; (5) Removing potential germline mutations; (6) Mutation frequencies for individual mutations or gene-level mutations; (7) Processing DNA methylation data.
(PDF)

S1 File. The complete list of mutation level eQTL results.
(TXT)

S2 File. The complete list of gene level eQTL results.
(TXT)

Acknowledgments

The results presented in the paper are in part based on the read-count data generated by The Seven Bridges Cancer Genomics Cloud platform.

Author Contributions

Conceptualization: Wei Sun.

Data curation: Wei Sun.

Formal analysis: Yang Liu.

Methodology: Yang Liu, Qianchan He, Wei Sun.

Software: Yang Liu.

Supervision: Qianchan He, Wei Sun.

Writing – original draft: Yang Liu, Wei Sun.

Writing – review & editing: Qianchan He, Wei Sun.

References

1. 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(7457):214–218. https://doi.org/10.1038/nature12213 PMID: 23770567
2. Raphael BJ, Vandin F, Dobson JR, Oesper L. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Medicine. 2014; 6(1):5. https://doi.org/10.1186/gm524 PMID: 24479672
3. Hua X, Hyland PL, Huang J, Song L, Zhu B, Caporaso NE, et al. MEGSA: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations. The American Journal of Human Genetics. 2016; 98(3):442–455. https://doi.org/10.1016/j.ajhg.2015.12.021 PMID: 26899600
4. Guinney J, Dienstmann R, Wang X, De Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nature Medicine. 2015; 21(11):1350–1356. https://doi.org/10.1038/nm.3967 PMID: 26457759
5. Pereira B, Chin SF, Rueda OM, Vollan HKM, Provenzano E, Bardwell HA, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nature Communications. 2016; 7:11479. https://doi.org/10.1038/ncomms11479 PMID: 27161491
6. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. Cell. 2017; 168(4):613–628. https://doi.org/10.1016/j.cell.2017.01.018 PMID: 28187284
7. Do H, Dobrovic A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. Clinical Chemistry. 2015; 61(1):64–71. https://doi.org/10.1373/clinchem.2014.223040 PMID: 25421801
8. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research. 2012; 22(3): 568–576. https://doi.org/10.1101/gr.129684.111 PMID: 22300766
9. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics. 2012; 28(14): 1811–1817. https://doi.org/10.1093/bioinformatics/bts271 PMID: 22581179
10. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology. 2013; 31(3):213–219. https://doi.org/10.1038/nbt.2514 PMID: 23396013
11. Fan Y, Xi L, Hughes DS, Zhang J, Zhang J, Futreal PA, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biology. 2016; 17(1):178. https://doi.org/10.1186/s13059-016-1029-6 PMID: 27557938
12. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics. 2014; 15(1):244. https://doi.org/10.1186/1471-2164-15-244 PMID: 24678773
13. Krøigård AB, Thomassen M, Lænkholm AV, Kruse TA, Larsen MJ. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. PLoS One. 2016; 11(3):e0151664. https://doi.org/10.1371/journal.pone.0151664 PMID: 27002637
14. Lin D, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies. Journal of the American Statistical Association. 2006; 101(473):89–104. https://doi.org/10.1198/016214505000000808
15. Tzeng JY, Zhang D. Haplotype-based association analysis via variance-components score test. The American Journal of Human Genetics. 2007; 81(5):927–938. https://doi.org/10.1086/521558 PMID: 17924336
16. Hu YJ, Liao P, Johnston HR, Allen AS, Satten GA. Testing rare-variant association without calling genotypes allows for systematic differences in sequencing between cases and controls. PLoS Genetics. 2016; 12(5):e1006040. https://doi.org/10.1371/journal.pgen.1006040 PMID: 27152526
17. Masica DL, Karchin R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. Cancer Research. 2011; 71(13):4550–4561. https://doi.org/10.1158/0008-5472.CAN-11-0180 PMID: 21555372
18. Ding J, McConechy MK, Horlings HM, Ha G, Chan FC, Funnell T, et al. Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nature Communications. 2015; 6. https://doi.org/10.1038/ncomms9554
19. Fredriksson NJ, Ny L, Nilsson JA, Larsson E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nature Genetics. 2014; 46(12):1258–1263. https://doi.org/10.1038/ng.3141 PMID: 25383969

20. Wei Y, Li X, Wang Qf, Ji H. iASeq: integrative analysis of allele-specificity of protein-DNA interactions in multiple ChIP-seq datasets. BMC Genomics. 2012; 13(1):681. https://doi.org/10.1186/1471-2164-13-681 PMID: 23194258

21. Sun W. A statistical framework for eQTL mapping using RNA-seq data. Biometrics. 2012; 68(1):1–11. https://doi.org/10.1111/j.1541-0420.2011.01654.x PMID: 21838806

22. Wang W, Wang W, Sun W, Crowley JJ, Szatkiewicz JP. Allele-specific copy-number discovery from whole-genome and whole-exome sequencing. Nucleic Acids Research. 2015; 43(14):e90–e90. https://doi.org/10.1093/nar/gkv319 PMID: 25883151

23. Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. Nucleic Acids Research. 2013; 41(7):e89–e89. https://doi.org/10.1093/nar/gkt126 PMID: 23471004

24. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Computational and structural biotechnology journal. 2018; 16:15–24. https://doi.org/10.1016/j.csbj.2018.01.003 PMID: 29552334

25. Ross EM, Markowetz F. OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome biology. 2016; 17(1):69. https://doi.org/10.1186/s13059-016-0929-9 PMID: 27083415

26. Jahn K, Kuipers J, Beerenwinkel N. Tree inference for single-cell data. Genome biology. 2016; 17(1):86. https://doi.org/10.1186/s13059-016-0936-x PMID: 27149953

27. Cancer Genome Atlas Network and others. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487(7407):330–337. https://doi.org/10.1038/nature11252 PMID: 22810696

28. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010; 11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106 PMID: 20979621

29. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015; 43(7):e47–e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792

30. Lynch J, Keller M, Guo RJ, Yang D, Traber P. Cdx1 inhibits the proliferation of human colon cancer cells by reducing cyclin D1 gene expression. Oncogene. 2003; 22(41):6395–6407. https://doi.org/10.1038/sj.onc.1206770 PMID: 14508520

31. Suh ER, Ha CS, Rankin EB, Toyota M, Traber PG. DNA methylation down-regulates CDX1 gene expression in colorectal cancer cell lines. Journal of Biological Chemistry. 2002; 277(39):35795–35800. https://doi.org/10.1074/jbc.M205567200 PMID: 12124393

32. Hinoue T, Weisenberger DJ, Lange CP, Shen H, Byun HM, Van Den Berg D, et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Research. 2012; 22(2):271–282. https://doi.org/10.1101/gr.117523.110 PMID: 21659424

33. Ruault M, Brun ME, Ventura M, Roizès G, De Sario A. MLL3, a new human member of the TRX/MLL gene family, maps to 7q36, a chromosome region frequently deleted in myeloid leukaemia. Gene. 2002; 284(1):73–81. https://doi.org/10.1016/S0378-1119(02)00392-X PMID: 11891048

34. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, et al. The Cancer Genomics Cloud: collaborative, reproducible, and democratized–a new paradigm in large-scale computational research. Cancer Research. 2017; 77(21):e3–e6. https://doi.org/10.1158/0008-5472.CAN-17-0387 PMID: 29092927

35. Cancer Genome Atlas Research Network.Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. New England Journal of Medicine. 2015; 372(26):2481–2498. https://doi.org/10.1056/NEJMoa1402121 PMID: 26061751

36. Kastenhuber ER, Lowe SW. Putting p53 in context. Cell. 2017; 170(6):1062–1078. https://doi.org/10.1016/j.cell.2017.08.028 PMID: 28886379

37. Fischer M. Census and evaluation of p53 target genes. Oncogene. 2017; 36(28):3943. https://doi.org/10.1038/onc.2016.502 PMID: 28288132