Received: 20 October 2022 | Revised: 22 July 2023 | Accepted: 18 September 2023

DOI: 10.1002/gepi.22537

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

Data-adaptive and pathway-based tests for association studies between somatic mutations and germline variations in human cancers

Zhongyuan Chen[1] | **Han Liang**[2] | **Peng Wei**[3]

[1]Division of Biostatistics, Medical College of Wisconsin, Milwaukee, Wisconsin, USA

[2]Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, Houston, Texas, USA

[3]Department of Biostatistics, MD Anderson Cancer Center, Houston, Texas, USA

Correspondence
Zhongyuan Chen, Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI 53226, USA.
Email: zhchen@mcw.edu

Peng Wei, Department of Biostatistics, MD Anderson Cancer Center, Houston, TX 77030, USA.
Email: pwei2@mdanderson.org

Abstract

Cancer is a disease driven by a combination of inherited genetic variants and somatic mutations. Recently available large-scale sequencing data of cancer genomes have provided an unprecedented opportunity to study the interactions between them. However, previous studies on this topic have been limited by simple, low statistical power tests such as Fisher's exact test. In this paper, we design data-adaptive and pathway-based tests based on the score statistic for association studies between somatic mutations and germline variations. Previous research has shown that two single-nucleotide polymorphism (SNP)-set-based association tests, adaptive sum of powered score (aSPU) and data-adaptive pathway-based (aSPUpath) tests, increase the power in genome-wide association studies (GWASs) with a single disease trait in a case–control study. We extend aSPU and aSPUpath to multi-traits, that is, somatic mutations of multiple genes in a cohort study, allowing extensive information aggregation at both SNP and gene levels. *p*-values from different parameters assuming varying genetic architecture are combined to yield data-adaptive tests for somatic mutations and germline variations. Extensive simulations show that, in comparison with some commonly used methods, our data-adaptive somatic mutations/germline variations tests can be applied to multiple germline SNPs/genes/pathways, and generally have much higher statistical powers while maintaining the appropriate type I error. The proposed tests are applied to a large-scale real-world International Cancer Genome Consortium whole genome sequencing data set of 2583 subjects, detecting more significant and biologically relevant associations compared with the other existing methods on both gene and pathway levels. Our study has systematically identified the associations between various germline variations and somatic mutations across different cancer types, which potentially provides valuable utility for cancer risk prediction, prognosis, and therapeutics.

KEYWORDS

association studies, data-adaptive pathway-based test, germline variations, somatic mutations

1 | INTRODUCTION

Germline variations and somatic mutations play key roles in cancer. According to Knudson's "two-hit" theory (Knudson, 1971), cancer results from the combination of germline variations and somatic mutations in the same gene, as illustrated in Figure 1. Later, this theory has been extended and is shown to be also applicable to different genes. Studies on interactions between germline variations and somatic mutations can help us seek how commonly inherited variations can affect later somatic mutations and progression of tumors (Carter et al., 2017; Ramroop et al., 2019; Vali-Pour et al., 2022).

Many studies indicate that there exist intrinsic relationships between germline variations and somatic mutations in cancer. For example, breast cancer patients with inherited *BRCA1* variations are more likely to lose or gain nucleotides on certain chromosomes (Jonsson et al., 2005; Stefansson et al., 2009). Melanoma patients with *MC1R* germline variations are more likely to have *BRAF* somatic mutations than those without (Landi et al., 2006; Maldonado et al., 2003). Some *RBFOX1* germline variants were found to significantly increase *SF3B1* somatic mutations (Carter et al., 2017). Recently, the interactions between germline and somatic variation information were discovered for biomarkers in prostate cancer (Mamidi et al., 2019a). Researchers also discovered that in colorectal cancer a single-nucleotide polymorphism (SNP) variant (rs78963230) is associated with both *TLR3* and *FBXW7* somatic mutations (Barfield et al., 2022). The germline–somatic variant interactions were revealed in urothelial cancer (Vosoughi et al., 2020). It is thus very important to study germline–somatic interactions in cancer.

Note that it is more important to study the associations between germline variations and multiple somatic mutations rather than a single somatic mutation. That is because in tumors germline variations are often associated with somatic mutations of multiple genes, which together complete a certain biological function. For example, it was mentioned that multiple somatic mutations were identified to occur simultaneously (Chan & Gordenin, 2015). Also, in tumors the germline variations are often associated with mutation patterns. For example, cancer patients who have germline variants near *APOBEC3* have reduced levels of *APOBEC* mutational signatures (instead of a single-gene's mutation), which is featured by replacing *C* with either *T* or *G* within *TCA* or *TCT* motifs (Middlebrooks et al., 2016; Nik-Zainal et al., 2014).

Although researchers have made significant efforts to associate somatic mutations with germline genotypes (Dworkin et al., 2010) and to detect genetic linkage for somatic mutations (LaFramboise et al., 2010), it had not been possible to conduct genome-wide investigation until the recent availability of both germline and tumor DNA sequence data (genome-wide association studies [GWASs] and whole-exome/genome sequencing [WES/WGS]) as afforded by large-scale cancer genomics consortia, such as the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) projects. However, current research into germline–somatic mutation interactions based on the ICGC or TCGA resource suffers from several limitations. The first is the use of the mutagenesis signatures, rather than somatic mutations themselves, as the outcome variables, leading to association results that are non-specific regarding germline variant–somatic mutation pairs and thus difficult to validate in functional studies (Chen et al., 2019; Waszak et al., 2017). The second limitation is that the employed statistical methods have been largely limited to conventional SNP-by-SNP GWAS association analysis by Fisher's exact test or logistic regression, leading to a loss of statistical power in genomic discoveries (Carter et al., 2017).

On the other hand, new SNP-set-based association tests have been proposed and applied to GWAS and WES/WGS of conventional quantitative and disease traits, such as cholesterols and cancer risk (Gusev et al., 2016; Ma & Wei, 2019; Pan et al., 2014, 2015; Wei et al., 2016; Wu et al., 2011; Xu et al., 2017; Yang et al., 2019). Nevertheless, these powerful methods have never been applied in the context of germline–somatic mutation interaction analysis.

In this paper, our objective is to adapt and develop powerful, data-adaptive statistical tests for genome-wide

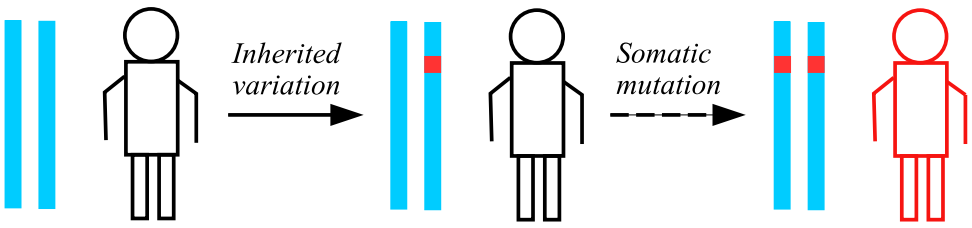

FIGURE 1 Illustration of Knudson's "two-hit" theory.

interrogation of germline–somatic mutation interactions that capitalize on the unique features of the germline and tumor genomes and address the challenges entailed by the interplay of these two genomes in human cancers. Our tests for the study of the associations between germline variations and somatic mutations are based on the adaptive sum of powered score (aSPU) (Pan et al., 2014) and data-adaptive pathway-based test (aSPUpath) (Pan et al., 2015) tests but extend to multitraits, that is, the somatic mutations of multiple genes especially driver genes in human cancers. Also, previous aSPU (Pan et al., 2014) and aSPUpath (Pan et al., 2015) tests were used for case–control study common in GWASs, but in our work we extend these two tests to cohort studies where all the samples are from cancer patients. We will group multiple SNPs into a gene, and further integrate functionally related genes into a pathway. This helps aggregate signals from multiple SNPs and genes and can potentially boost the statistical power and maintain type I error at the same time.

As a case study, we systematically examine the associations between germline variations and somatic mutations using a large-scale pan-cancer genomics ICGC data set (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020) (2583 cancer patients from 38 cancer types). After a sequence of data quality control processing, we apply the modified aSPU and aSPUpath tests to uncover the interactions between germline variations (both gene level and pathway level) and somatic mutations across all cancer types in humans. We also show the results using the aSPU test outperform those using Fisher's exact test and some other commonly used germline-somatic association test methods in terms of detecting more association signals. On the pathway level, using the aSPUpath test, we systematically detect some hotspots such as *CTNNB1* and *KRAS* somatic mutation genes that are associated with almost all the pathway genes' germline variations, and pathway RTK/RAS's germline variations are associated with the somatic mutations of a large proportion of driver genes. We further study the dominant contributing genes in a pathway to the pathway–somatic associations. Finally, we generate a network of the pathway germline variations and somatic mutations to give a systematic view of the interactions.

Our research provides valuable statistical tools for cancer risk prediction. The work systematically identifies the associations between various germline variations and somatic mutations across different cancer types, which in turn indicates the chance of cancer risks following Knudson's "two-hit" theory. The work also helps further understanding of molecular mechanisms of specific cancer genes and provides new insights into the development of novel cancer therapy.

The remainder of the article is organized as follows. In Section 2, we develop statistical association tests for somatic mutations and germline variations. In Section 3, we focus on the data-adaptive pathway-based test for association studies between gene somatic mutations and germline variations. In Section 4, we conduct extensive simulation studies to show the high statistical power of our method compared with others. Section 5 shows the results of applying the proposed methods to the association studies between somatic mutations and germline variations on both gene and pathway levels using a large-scale pan-cancer data set ICGC. The conclusions and some discussions are given in Section 6.

2 | A POWERFUL AND ADAPTIVE TEST FOR ASSOCIATION STUDIES BETWEEN SOMATIC MUTATIONS AND GERMLINE VARIATIONS

In our work, we perform association studies between somatic mutations and germline variations in a cohort study. For a given gene, we use hypothesis testing to study whether its somatic mutation is associated with any germline variations across multiple cancer types. Previous studies are only concerned with associations between single germline SNP and somatic mutations. Here, however, we inspect the associations between whole genes (including all SNPs in one gene [Section 2] or even multiple genes in a pathway [Section 3]) and somatic mutations. This not only increases statistical power, but also is biologically more meaningful.

2.1 | Notation

Suppose for n subjects $(i = 1, 2, ..., n)$, $X_i = [X_{i1}, ..., X_{ij}, ..., X_{ip}]^T$ is a length-p vector for the genotype scores, where p is the number of SNPs in the gene under consideration and $X_{ij} = 0$ or 1 or 2 is the genotype score, representing the number of the minor allele, at SNP j for subject i. Let $Y_i = 0$ or 1 be a vector of binary traits, where $Y_i = 1$ indicates there is a somatic mutation in a given gene under consideration of subject i.

Let $L(\boldsymbol{\beta}; X)$ be the likelihood function. The score vector $U = [U_1, ..., U_p]^T$ associated with L is $U_i(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_i} \log L(\boldsymbol{\beta}; X)$. For binary Y_i, we consider a logistic regression model

$$\text{logit}[\Pr(Y_i = 1)] = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j \equiv \beta_0 + X_i^T\boldsymbol{\beta}. \quad (1)$$

The null hypothesis to be tested is $H_0: \boldsymbol{\beta} = 0$, which indicates that there is no association between any SNPs and somatic mutation of a gene under H_0.

In standard score tests, the score vector has a specific form as follows:

$$U = \sum_{i=1}^{n} X_i(Y_i - \bar{Y}) = X^T(Y - \bar{Y}), \tag{2}$$

where

$$X = [X_1, ..., X_n], \quad Y = [Y_1, ..., Y_n]^T, \tag{3}$$

and $\bar{Y}$ is the sample mean of all Y_i's. U has the covariance matrix $C = \mathrm{Cov}(U) = \bar{Y}(1 - \bar{Y}) \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$. U also has an asymptotic normal distribution $\mathcal{N}(0, C)$. The statistics in some commonly used score test schemes for GWASs are in Supplementary Information Table S1. See, for example, Pan et al. (2015) for a review.

2.2 | aSPU: the data-adaptive test

A class of sum of powered score (SPU) tests was proposed by Pan et al. (2014) for different types of traits and adjusted for covariates: $T_{SPU}(\gamma) = \sum_{j=1}^{k} U_j^{\gamma}$. The Sum ($\gamma = 1$) and sum of squared score (SSU) ($\gamma = 2$) tests are two special cases of the SPU tests. The score statistic-based tests perform differently for different situations. For example, if the association directions are different so that U_j's have different signs and T_{Sum} is small, then the sum test loses its power in rejecting H_0. On the other hand, this does not happen to the SSU test (Pan, 2009). The SPU test is more powerful if most association directions are the same. With $\gamma \to \infty$, the SPU test is more powerful if one or few variables have large association effect sizes. Interestingly, when $\gamma \to \infty$, the SPU test approaches to the minimum p-value (UminP) test under certain conditions (Pan et al., 2015).

To enhance the statistical power, some adaptive test methods were further proposed, especially for high-dimensional data. Examples include the adaptive Neyman test (Fan, 1996), the adaptive SSU test (Pan & Shen, 2011), the adaptive SPU (aSPU) test (Pan et al., 2014), and a pathway-based adaptive SPU (aSPUpath) test (Pan et al., 2015).

The aSPU test in Pan et al. (2014) was originally designed for association studies between a binary trait and a set of rare variants in a case–control study. It is shown to be a powerful test and also data adaptive. The data adaptivity can be understood as follows. Consider the SSU statistic in Supplementary Information Table S1, which can be rewritten as

$$T_{SPU}(\gamma) = \sum_{j=1}^{k} U_j^{\gamma} = \sum_{j=1}^{k} U_j^{\gamma-1} U_j, \tag{4}$$

where $\zeta_j = U_j^{\gamma-1}$. That is, the SSU test can be considered as an adaptive Sum test in Supplementary Information Table S1 with the weights ζ_j depending on the data itself. Because different γ values are more suitable for different data situations, it is desirable to choose γ adaptively depending on the data. The idea of the aSPU test is to choose various γ parameters from a candidate set Γ and select the best statistic. Multiple γ parameters yield multiple SPU p-values $P_{SPU}(\gamma)$. A minimum p-value is then selected as in the minimum p method (Tippett, 1931). The minimum is not a genuine p-value anymore, so it is treated as a new test statistic. Permutations can be used to estimate its p-value. We discuss the details as follows for our setting.

2.3 | aSPU test for gene-based association analysis of somatic mutations and germline variations

In the aSPU association test, we first use X and Y to compute the score vector (2) and the SPU statistic (4) for a given parameter γ. We can use resampling methods such as permutations (Churchill & Doerge, 1994) to estimate the p-value of SPU(γ). That is, permute Y to obtain a set of K trait vectors $Y^{(k)}, k = 1, 2, ..., K$ and compute the corresponding score vector $U^{(k)}$ and SPU statistics $T_{SPU}^{(k)}(\gamma)$. Then we can (approximately) calculate the p-value as

$$P_{SPU}(\gamma) = \frac{1}{K+1}\left(\sum_{k=1}^{K} \mathcal{I}\left(\left|T_{SPU}^{(k)}(\gamma)\right| \geq |T_{SPU}(\gamma)|\right) + 1\right), \tag{5}$$

where $\mathcal{I}(\cdot)$ is the indicator function.

Because the SPU with a specific γ works well for specific data, it is desirable to try multiple γ values, which leads to the aSPU test (Pan et al., 2014): a set of candidate γ values is used and leads to different $P_{SPU}(\gamma)$ values. Candidate sets such as $\Gamma = \{1, 2, ..., 8\}$ can be used. The aSPU test simply follows the minimum p method (Tippett, 1931) by taking the minimum p-value:

$$T_{aSPU} = \min_{\gamma \in \Gamma} P_{SPU}(\gamma). \tag{6}$$

T_{aSPU} is not a genuine p-value anymore. Thus, T_{aSPU} is treated as a new test statistic. Permutations or bootstraps can be used to estimate its p-value. Note

that we can reuse the above permutations in the computation of the $P_{SPU}(\gamma)$ values so as to avoid double permutations. That is, we can reuse the values $T^{(j)}_{SPU}(\gamma), j = 1, 2, ..., k-1, k+1, ..., K$ in the estimation of $P^{(k)}(\gamma)$. The p-value for each simulation k is then (approximately)

$$P^{(k)}(\gamma) = \frac{1}{K}\sum_{j\neq k}\left(\mathcal{I}\left(T^{(j)}_{SPU}(\gamma) \geq T^{(k)}_{SPU}(\gamma)\right) + 1\right).$$

Letting $T^{(k)}_{aSPU} = \min_{\gamma\in\Gamma} P^{(k)}(\gamma)$, the aSPU test p-value is then

$$P_{aSPU} = \frac{1}{K+1}\sum_{k=1}^{K}\left(\mathcal{I}\left(T^{(k)}_{aSPU} \leq T_{aSPU}\right) + 1\right). \quad (7)$$

3 | A DATA-ADAPTIVE PATHWAY-BASED TEST FOR ASSOCIATION STUDIES BETWEEN GENE SOMATIC MUTATIONS AND GERMLINE VARIATIONS

3.1 | Cancer signaling pathways

Here, we further consider using pathways in the association studies between germline variations and somatic mutations. Because a single SNP or gene may have a low mutation frequency, signals from multiple functionally related genes can be aggregated together and studied in a relevant pathway. In fact, pathways are important tools in advanced statistical genomics. Using pathways instead of individual genes in association testing can improve the statistical power.

Using pathways is also biologically more meaningful. To adapt to the environment, cells in human beings need to process signals from the outside. The function of a pathway is to communicate between outside signals and cell nucleus or between different cells. In cancer cells, certain genes in some pathways are mutated, which dysregulates signaling in cancer and leads to the change of some characteristics of tumor cells.

3.2 | aSPUpath test for pathway-based association analysis of somatic mutations and germline variations

Following the pathway-based aSPU method aSPUpath in Pan et al. (2015), we study association testing between the SNPs in a pathway and somatic mutations, and expect to gain higher statistical powers. Specifically, given a pathway S with $|S|$ genes, partition the score vector U in (2) according to the genes as

$$U = \left[U_1^T, U_2^T, ..., U_{|S|}^T\right]^T,$$

where $U_g = [U_{g1}, U_{g2}, ..., U_{gl_g}]^T$ is the score subvector with l_g SNPs for gene g.

For each gene g, the SPU test statistic is calculated as

$$T_{SPU}(\gamma_1; g) = \left(\frac{1}{l_g}\sum_{j=1}^{l_g}(U_{gj})^{\gamma_1}\right)^{1/\gamma_1}, \quad (8)$$

where $\gamma_1 > 0$ is a scalar parameter. Note that standardization is used in (8) so as to balance genes with different sizes. For the pathway, calculate the following test statistic:

$$T_{SPUpath}(\gamma_1, \gamma_2; S) = \sum_{g\in S}(T_{SPU}(\gamma_1; g))^{\gamma_2}, \quad (9)$$

where $\gamma_2 > 0$ is a scalar parameter. For given parameters γ_1 and γ_2, resampling methods such as permutations can be used to calculate the p-value $P_{SPUpath}(\gamma_1, \gamma_2; S)$. That is, permute Y to obtain a set of K trait vectors $Y^{(k)}$ and compute the corresponding score vectors $U^{(k)}$ and SPUpath statistic $T^{(k)}_{SPUpath}(\gamma_1, \gamma_2; S)$. We then have (estimates of) the p-value

$$P_{SPUpath}(\gamma_1, \gamma_2; S) = \frac{1}{K+1}\left(\sum_{k=1}^{K}\mathcal{I}\left(\left|T^{(k)}_{SPUpath}(\gamma_1, \gamma_2; S)\right| \geq |T_{SPUpath}(\gamma_1, \gamma_2; S)|\right) + 1\right).$$

To accommodate different data situations, we choose different parameters γ_1 and γ_2 and define a pathway-based aSPU test statistic as

$$T_{aSPUpath}(S) = \min_{\gamma_1\in\Gamma_1, \gamma_2\in\Gamma_2} P_{SPUpath}(\gamma_1, \gamma_2; S), \quad (10)$$

where Γ_1 and Γ_2 are the candidate parameter sets.

Finally, the corresponding p-value $P_{aSPUpath(S)}$ can be computed also with permutations. Just like in the aSPU case, there is no need to use double permutations. Instead, let

$$P^{(k)}_{SPUpath}(\gamma_1, \gamma_2; S) = \frac{1}{K}\left(\sum_{j\neq k}\mathcal{I}\left(\left|T^{(j)}_{SPUpath}(\gamma_1, \gamma_2; S)\right| \geq \left|T^{(k)}_{SPUpath}(\gamma_1, \gamma_2; S)\right|\right) + 1\right).$$

Then let

$$T^{(k)}_{aSPUpath}(S) = \min_{\gamma_1\in\Gamma_1,\gamma_2\in\Gamma_2} P^{(k)}_{SPUpath}(\gamma_1,\gamma_2;S),$$

so as to obtain the p-value

$$P_{aSPUpath(S)} = \frac{1}{K+1}\sum_{k=1}^{K}\left(\mathcal{I}\left(T^{(k)}_{aSPUpath}(S) \le T_{aSPUpath}(S)\right)+1\right).$$

Compared with the aSPU model, the aSPUpath model requires to provide SNP information and gene information. The SNP information includes SNP IDs, chromosome numbers, and SNP locations. The gene information includes gene IDs, chromosome number, and start and end positions of the gene. In real-world data sets such as ICGC, all such information is available. This is described in detail in Supplementary Information Section A.

In our later discussions, the test methods aSPU and aSPUpath will be thoroughly evaluated through both simulations and real-world tests. In our simulation studies, unlike previous aSPU and aSPUpath tests, we consider cohort studies instead of case–control studies. The details will be discussed in Section 4.1. For real-world data analysis, the SNPs will be prescreened via a minor allele frequency (MAF) criterion. For example, MAF > 5% is used for common variants and MAF < 5% is used for rare variants. Then we identify SNPs that need to be mapped to genes and genes that need to be mapped to pathways. The details are given in Supplementary Information Section A.

4 | SIMULATION STUDIES

4.1 | Simulation setup

We used extensive simulation studies to evaluate the performance of the proposed methods in terms of type I error rate and power. To evaluate the adaptive pathway-based testing for association studies between somatic mutations and germline variations, we modified function "simPathAR1Snp" in the R package "aSPU" (Kwak et al., 2021) and generated the simulation data (SNP matrix X and trait vector Y) as follows.

The germline SNP data matrix X was simulated following (Pan et al., 2015; Wang & Elston, 2007). First, we generated a latent vector $x = [x_1, ..., x_p]^T$ from a multivariate normal distribution based on the following autoregressive covariance structure for latent variables x_i and x_j:

$$Corr(x_i, x_j) = \theta^{|i-j|}. \quad (11)$$

If $\theta = 0$, the SNPs are independent. If $\theta > 0$, the SNPs are correlated. A haplotype was then obtained based on certain MAFs. Two such independent haplotypes were combined to yield the genotype data X_i for subject i. For the null case, all $\beta_j = 0, j = 1, ..., p$. For the non-null case, certain SNPs within some genes are set to be causal with $\beta_j = \log \mathrm{OR} \neq 0$.

To simulate single-gene somatic mutation data, the mutation status Y_i of patient i was generated from the logistic regression model. In case–control simulation studies in Pan et al. (2015) and Wang and Elston (2007), there are two sample groups: normal samples (without disease) and tumor samples (with diseases). The case–control study is to investigate whether the disease is significantly associated with the predefined risk factor, say, gene mutations, where the disease group is typically oversampled from a population.

Here, unlike the case–control simulations in Pan et al. (2015) and Wang and Elston (2007), we consider cohort studies, where all the subjects are patients because we only have tumor samples. There are two sample groups: samples with somatic mutations in a certain gene and those without in the same gene. That is, we investigate whether there is a statistically significant association between germline variations of genes and somatic mutations of the same or different genes. In our study, we have multi-traits corresponding to the response Y vector because we have multiple somatic mutation genes. On the other hand, in Pan et al. (2014, 2015) a single trait corresponding to Y vector is studied because one disease or other single trait, for example, cholesterol level, is focused.

Following the logistic regression model in (1), we define p_i as the probability of obtaining samples with somatic mutations ($Y = 1$) and $1 - p_i$ as the probability of obtaining those without somatic mutations ($Y = 0$), where

$$p_i = \frac{e^{\beta_0 + X_i^T\boldsymbol{\beta}}}{1 + e^{\beta_0 + X_i^T\boldsymbol{\beta}}}.$$

That is, the binary trait $Y_i = 1$ or 0 is sampled following the probability p_i. To decide a certain background mutation probability p_0, we use

$$p_0 = \frac{e^{\beta_0}}{1 + e^{\beta_0}}.$$

4.2 | Simulation results

We conducted 12 sets of simulations to extensively study the performances of the association test models. See Supplementary Information Table S2 for a list of all the

setup situations. Some of the test parameters are similar to those in Pan et al. (2015, 2022).

- The simulations involve 20 genes each with a random number of SNPs.
- Two sets of simulated genes are involved, where the total number of SNPs is either 302 (with the number of SNPs in each gene between 1 and 20) or 1042 (with the number of SNPs in each gene between 3 and 100).
- The first SNP of 1, 5, or 10 genes is set to be a causal SNP.
- The MAFs used in the haplotype generation are randomly selected between 0.05 and 0.4.
- We set the background disease prevalence $p_0 = 0.1$, that is, $\beta_0 = \log(0.1/0.9)$ for a 10% background mutation probability.
- $n = 500$ patient data is simulated in the cohort study.
- $K = 500$ permutations are used in the p-value computation.
- 1000 simulation replications are used to evaluate the power.
- We also simulate both independent SNPs (linkage equilibrium) with $\theta = 0$ in (11) and correlated SNPs (linkage disequilibrium) with θ following a uniform distribution $\mathcal{U}(0, 0.8)$.

We demonstrated the powers of the adaptive methods aSPU and aSPUpath as described in Sections 2 and 3. aSPU uses the candidate parameters as suggested in Pan et al. (2015):

$$\gamma \in \Gamma = \{1, 2, ..., 8\}, \tag{12}$$

aSPUpath uses candidate parameters

$$\gamma_1 \in \Gamma_1\{1, 2, 3, 4, 5, 6, 7, 8\}, \quad \gamma_2 \in \Gamma_2\{1, 2, 4, 8\}. \tag{13}$$

The test functions were written in R and were based on the package "aSPU" (Kwak et al., 2021).

Some other commonly used methods (hybrid set-based test [HYST], Li et al., 2012; Gates-Simes, Gui et al., 2011; SSU and UminP) were also compared. Here, both HYST and Gates-Simes combine p-values via a gene-level test (Li et al., 2011). The former is based on Fisher's method, and the latter is based on an extended Simes method to obtain p-values. Both methods become low powered if there are weak or no associations between many SNPs/genes and the trait. As shown in Supplementary Information Table S1, SSU corresponds to SPU with $\gamma = 2$. Both SSU and UminP lose power if there are many SNPs with weak associations.

4.2.1 | Statistical power

Figure 2 shows the powers of the methods with different parameter setups in Supplementary Information Table S2. In general, the powers from all the methods increase with the increase of number of causal SNPs regardless of whether the SNPs are independent or correlated. Both aSPU and aSPUpath perform reasonably well in general. In particular, with the number of causal SNPs increases (5 and 10 causal SNPs), aSPU and aSPUpath are more powerful than the other methods. When the number of causal SNPs stays fixed and the number of total SNPs increases, the powers of the methods are a little lower. See the comparison between Figures 2a–c and 2g–i (and also Figures 2d–f and 2j–l).

4.2.2 | Type I error

We also evaluated the type I errors for each setup with the test methods. See Table 1. The type I errors of aSPU and aSPUpath maintain around the nominal significance level of 0.05, which is comparable to the other existing association tests, such as HYST, Gates-Simes, SSU, and UminP. The previous literature (Pan et al., 2014) has demonstrated that the type I error of aSPU tests could be controlled at 10^{-3} or 10^{-4} level with 10^5 simulation replicates, but it is computationally very expensive. From the simulations it can be seen that the data-adaptive schemes aSPU and aSPUpath are generally more powerful while maintaining similar type I errors as some other methods.

4.2.3 | Computational burden

In addition, we compared the computational burden of aSPU and aSPUpath with the other methods, such as SSU, UminP, HYST, and Gates-Simes using the simulated data (scenario see setup D in Supplementary Information Table S2). In this scenario, there are 20 genes (including 302 SNPs) and 500 patients. The results are shown in Table 2. In general, aSPU and aSPUpath do not show the advantage in the aspect of computational burden. This is consistent with the results of Ma and Wei (2019). It is not surprising because aSPU and aSPUpath involved quite a few parameters and many

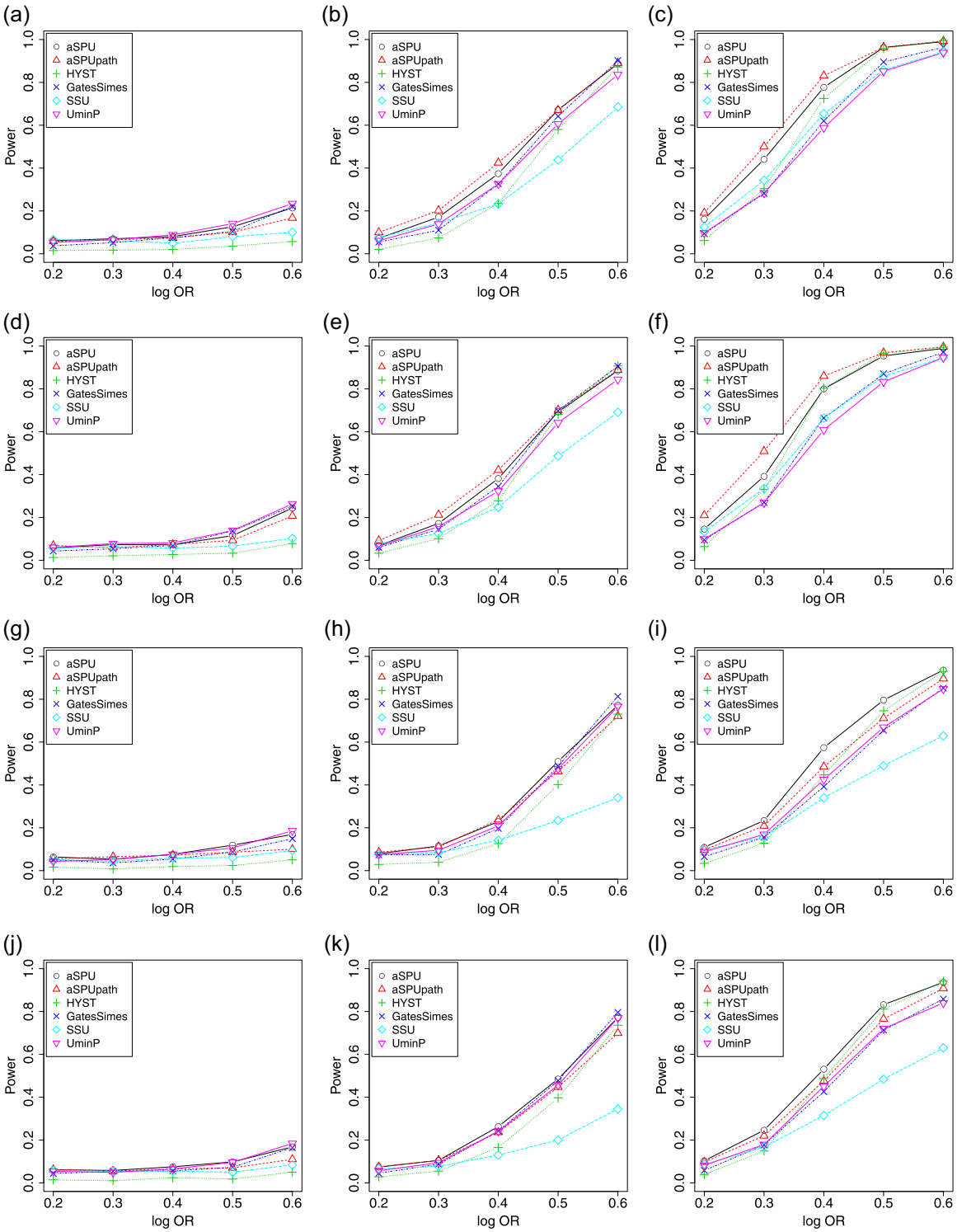

FIGURE 2 Powers of different test methods with 12 test setups. The number of total SNPs is 302 (a–f) or 1042 (g–l). The SNPs are independent (a–c, g–i) or correlated (d–f, j–l). (a) 1 causal SNP, (b) 5 causal SNPs, (c) 10 causal SNPs, (d) 1 causal SNP, (e) 5 causal SNPs, (f) 10 causal SNPs, (g) 1 causal SNP, (h) 5 causal SNPs, (i) 10 causal SNPs, (j) 1 causal SNP, (k) 5 causal SNPs, and (l) 10 causal SNPs. aSPU, adaptive sum of powered score; aSPUpath, data-adaptive pathway-based test; Gates-Simes, gene-based association test using extended Simes procedure; HYST, hybrid set-based test; SNP, single-nucleotide polymorphism; SSU, sum of squared score; UminP, univariate minimum *p*-value.

TABLE 1 Type I errors for the tests of the setups in Supplementary Information Table S2.

Method	aSPU	aSPUpath	HYST	Gates-Simes	SSU	UminP
Setup A	0.049	0.057	0.009	0.037	0.046	0.047
Setup B	0.068	0.070	0.018	0.043	0.057	0.064
Setup C	0.053	0.051	0.012	0.043	0.049	0.054
Setup D	0.053	0.064	0.011	0.040	0.051	0.045
Setup E	0.059	0.053	0.010	0.041	0.054	0.048
Setup F	0.058	0.052	0.020	0.049	0.052	0.051
Setup G	0.036	0.042	0.006	0.040	0.039	0.050
Setup H	0.055	0.069	0.014	0.049	0.057	0.054
Setup I	0.055	0.056	0.006	0.042	0.047	0.048
Setup J	0.050	0.057	0.010	0.056	0.042	0.055
Setup K	0.070	0.062	0.008	0.027	0.058	0.057
Setup L	0.052	0.057	0.020	0.037	0.047	0.054

Abbreviations: aSPU, adaptive sum of powered score; aSPUpath, data-adaptive pathway-based test; Gates-Simes, gene-based association test using extended Simes procedure; HYST, hybrid set-based test; SSU, sum of squared score; UminP, univariate minimum *p*-value.

TABLE 2 Comparison of the computational burden (unit: second) in different methods (Scenario D in Supplementary Information Table S2).

aSPU Mean (SD)	aSPUpath Mean (SD)	SSU Mean (SD)	UminP Mean (SD)	HYST Mean (SD)	Gates-Simes Mean (SD)
0.8902 (0.0265)	1.1357 (0.0377)	0.8087 (0.0297)	0.7940 (0.0303)	1.0718 (0.0166)	1.0700 (0.0122)

Note: Mean and standard deviation (SD) values of the computational time from 1000 simulations are reported.

Abbreviations: aSPU, adaptive sum of powered score; aSPUpath, data-adaptive pathway-based test; Gates-Simes, gene-based association test using extended Simes procedure; HYST, hybrid set-based test; SSU, sum of squared score; UminP, univariate minimum *p*-value.

permutations which is expensive for large data. Actually, we are working on a method for saving the computational time of aSPU and aSPUpath tests using a low-rank parameter preselection, which is expected to appear in Chen et al. (2023).

4.2.4 | Sample size

The last part but not the lest, we investigated the proper sample size needed for the different tests. We performed the simulations (10 causal SNPs, Scenarios C, F, I, L, fixed Log OR = 0.4) varying the number of patients from 100, 200, 300, 400, 500, to 1000. The results are shown in Figure 3. In general, when the sample size goes up to 1000, the statistical power of all the tests equals to 1. aSPU and aSPU outperform the other methods in terms of the power across different sample sizes, especially when the sample size is relatively small (less than 300). Interestingly, when the sample size goes down to 100, the powers of aSPU and aSPU tests show better than those of the other methods when the number of SNPs is relatively small. (see Scenarios C and F with 302 number of SNPs).

5 | DATA EXAMPLE: INTEGRATIVE ANALYSIS OF ASSOCIATIONS BETWEEN SOMATIC MUTATIONS AND GERMLINE VARIATIONS IN THE ICGC

In this section, we study the associations between somatic mutations and germline variations for a real-world pan-cancer data set called the ICGC. Our aim is to systematically characterize such associations across multiple cancer types. The ICGC data first undergoes a sequence of preprocessing so that we can extract data that can be analyzed with statistical tests. We then apply the tests in the previous sections to the resulting data and perform a comprehensive analysis of the underlying interactions between germline variations and somatic mutations.

5.1 | ICGC data and processing

We briefly summarize the ICGC data preprocessing procedure here (see details in Supplementary Information

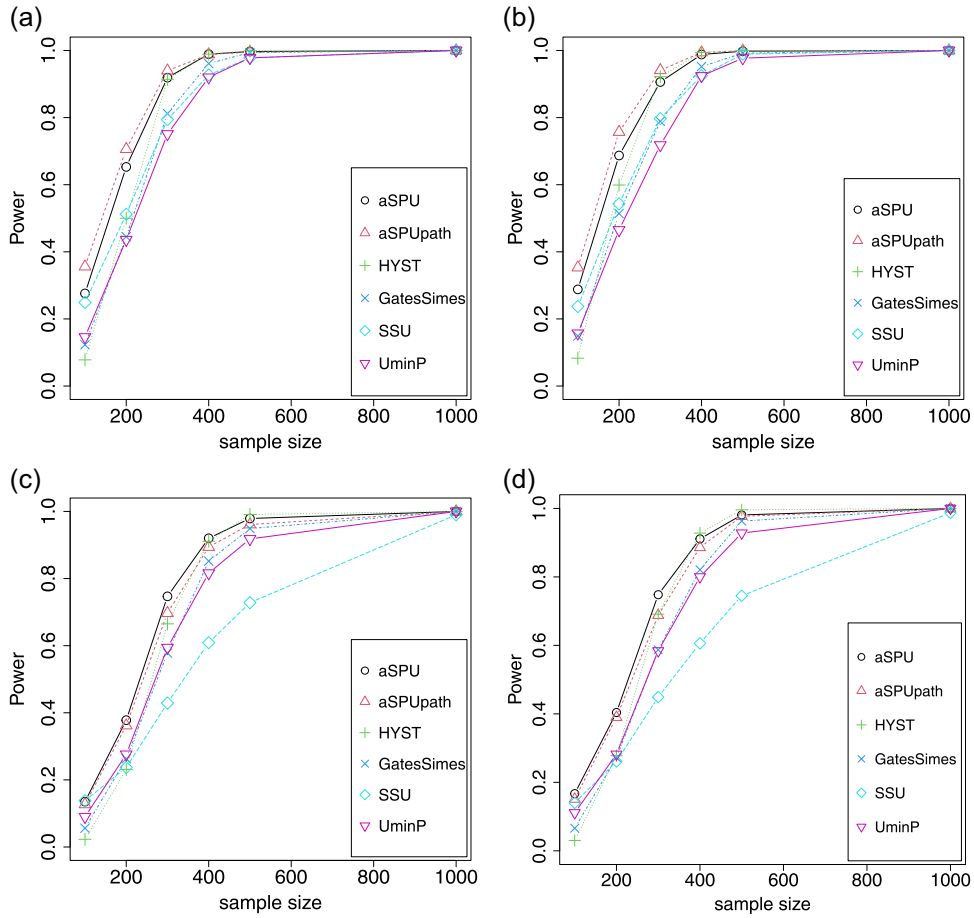

FIGURE 3 The effect of sample size on the powers of different test methods with four test setups as in Supplementary Information Table S2. (a) Scenario C, (b) Scenario F, (c) Scenario I, and (d) Scenario L. aSPU, adaptive sum of powered score; aSPUpath, data-adaptive pathway-based test; Gates-Simes, gene-based association test using extended Simes procedure; HYST, hybrid set-based test; SSU, sum of squared score; UminP, univariate minimum *p*-value.

Section A). The ICGC Pan-Cancer Analysis of Whole Genomes (PCAWG) data set includes 2583 cancer patients (white list) with 38 cancer types (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). We preprocessed the ICGC germline variation and somatic mutation data regarding the samples and the SNPs/genes. A series of steps of preprocessing the SNPs included removing indels, removing SNPs in sex chromosomes, and keeping the common variants (MAF > 5%). After removing hypermutated samples, 2575 samples were left. Then we found the number of overlapping samples from germline samples and somatic samples was 2561. Finally, we had the SNP data matrix which included genotype scores (0, 1, 2) from 6,495,525 SNPs for 2561 samples. That is, $\mathbf{X}$ is a $2561 \times 6{,}495{,}525$ matrix with 0, 1, 2 entry in each cell. SNPs were then annotated to genes via a gene annotation software package Oncotator (Ramos et al., 2015) (http://portals.broadinstitute.org/oncotator/). For somatic mutation data, we removed the genes that are annotated as silent and Intergenic region (IGR). Finally, the cleaned somatic matrix $\mathbf{Y}$ had 24,837 genes for 2561 samples. That is, $\mathbf{Y}$ is a $2561 \times 24{,}837$ matrix, and its columns can be used to extract the trait vectors Y in (3) for the genes with somatic mutation. From ICGC data, there are 159 driver genes with protein-coding point mutations, among which 150 genes overlap with our annotated germline variation genes and 156 genes overlap with our annotated somatic mutation genes.

5.2 | aSPU association test results

5.2.1 | aSPU association test results

The ICGC driver gene list (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020) was used for testing the associations between somatic mutations and germline variations. We first performed aSPU tests on ICGC driver genes' somatic and germline data.

For each association test, the X data matrix in (3) includes all SNPs in one gene and the binary trait vector Y in (3) is for one gene's somatic mutation status. We choose bootstrapping as the resampling method and choose γ from the candidate set (12). The permutation number is 1000. In all, 150 overlapping genes between our annotated common germline SNPs and the driver gene list were chosen in our tests. Also, 156 overlapping genes between our annotated somatic mutation genes and the driver gene list were used. On the basis of the mapping table, germline variation samples and somatic mutation samples were matched to the same patients. The total number of association tests between the 150 germline variation genes with SNPs and the 156 somatic mutation genes was $150 \times 156 = 23{,}400$.

We discovered that the p-values from 1728 out of 23,400 associations tests (a rate of 7.38%) were less than the nominal significance level 0.05. Supplementary Information Tables S5 and S6 show selected p-values. Totally, seven out of 150 germline genes had significant associations with 20 or more somatic mutation genes. False discovery rate (FDR) multiple testing correction was then performed. The number of FDR values less than 0.2 was 210.

We then drew heatmaps for both p-values and FDR values from the tests (Supplementary Information Figure S1A and Figure 4a). For the purpose of easy visualization, in all the heatmaps, we show the values $-\log_{10}(p\text{-value or FDR})$. For the purpose of postquality control, we only kept the genes whose somatic mutation frequency is more than 1% across all cancer types in the heatmap result, although we do the association tests for all the driver genes.

Interestingly, we found that somatic mutations in the gene *CTNNB1* were associated with almost all of the germline gene variations (Figure 4a). The gene KRAS was the second hotspot besides *CTNNB1*. In addition, germline variations in the gene *KMT2C* are associated with the somatic mutations in a large proportion of genes. Literatures for supporting these findings will be discussed in Section 5.7.

5.2.2 | *cis*/*trans* association results

Based on Knudson's "two-hit" theory, patients tend to develop cancer if they have both germline variations and somatic mutations of the same gene. Thus, we are interested in whether associations between the same genes' germline variations and somatic mutations tend to happen more often than different genes'. We refer to the same genes' associations in both germline variations and somatic mutations as "*cis*," and refer to different genes' association studies as "*trans*." We created the 2×2 contingency table (Table 3) for the *cis* significant association count, *trans* significant association count, *cis* nonsignificant association count, and *trans* nonsignificant association count. It demonstrated that the significant count from *cis* associations is not significantly higher/lower than *trans* associations (Fisher's exact test p-value $= 0.339$), although *cis* significant count percentage (9.5%) was slightly larger than *trans* significant count percentage (7.4%). That means, the same genes' germline variations and somatic mutations associations did not tend to happen more often than different genes' associations.

5.3 | aSPU test versus Fisher's exact test

To demonstrate that the data-adaptive aSPU test is more powerful than commonly used association test, we performed Fisher's exact test on the same ICGC data set and compared the results from aSPU. Note that Fisher's exact test is limited to single SNPs in germline variation genes. Given each germline gene, we performed Fisher's exact test between the variation status of each SNP in the gene and one somatic mutation status. A 2×3 contingency table was created with each row indicating one gene's somatic mutation status (which is binary with 0 for no mutation and 1 for mutation) and each column indicating germline SNP genotype score 0, 1, or 2 as defined before for the SNP matrix X in Section 2.1. Because there are multiple SNPs in each germline gene (Supplementary Information Table S4), there are multiple Fisher's exact tests for each germline variation and somatic mutation association. We chose the minimum p-value among multiple association tests corresponding to the SNPs within one germline gene. Then we performed Bonferroni multiple testing correction on the minimum p-value for the germline gene association by the number of SNPs in the gene. The rest settings are exactly the same as those in the aSPU test.

We found that the corrected p-values from 675 out of 23,400 associations tests are smaller than the nominal significance level 0.05 (a rate of 2.88%). Supplementary Information Tables S7 and S8 show significant p-values (Bonferroni corrected within one gene for multiple SNPs testing) for 7 selected germline genes as those from the aSPU test results in Supplementary Information Tables S5 and S6. Clearly, the numbers of significant associations are much smaller than those from the aSPU test.

We also plotted the heatmaps for both p-values and FDR values from Fisher's exact test. We kept the exact same gene list with a somatic mutation frequency of more than 1%. See Figure 4. By comparing

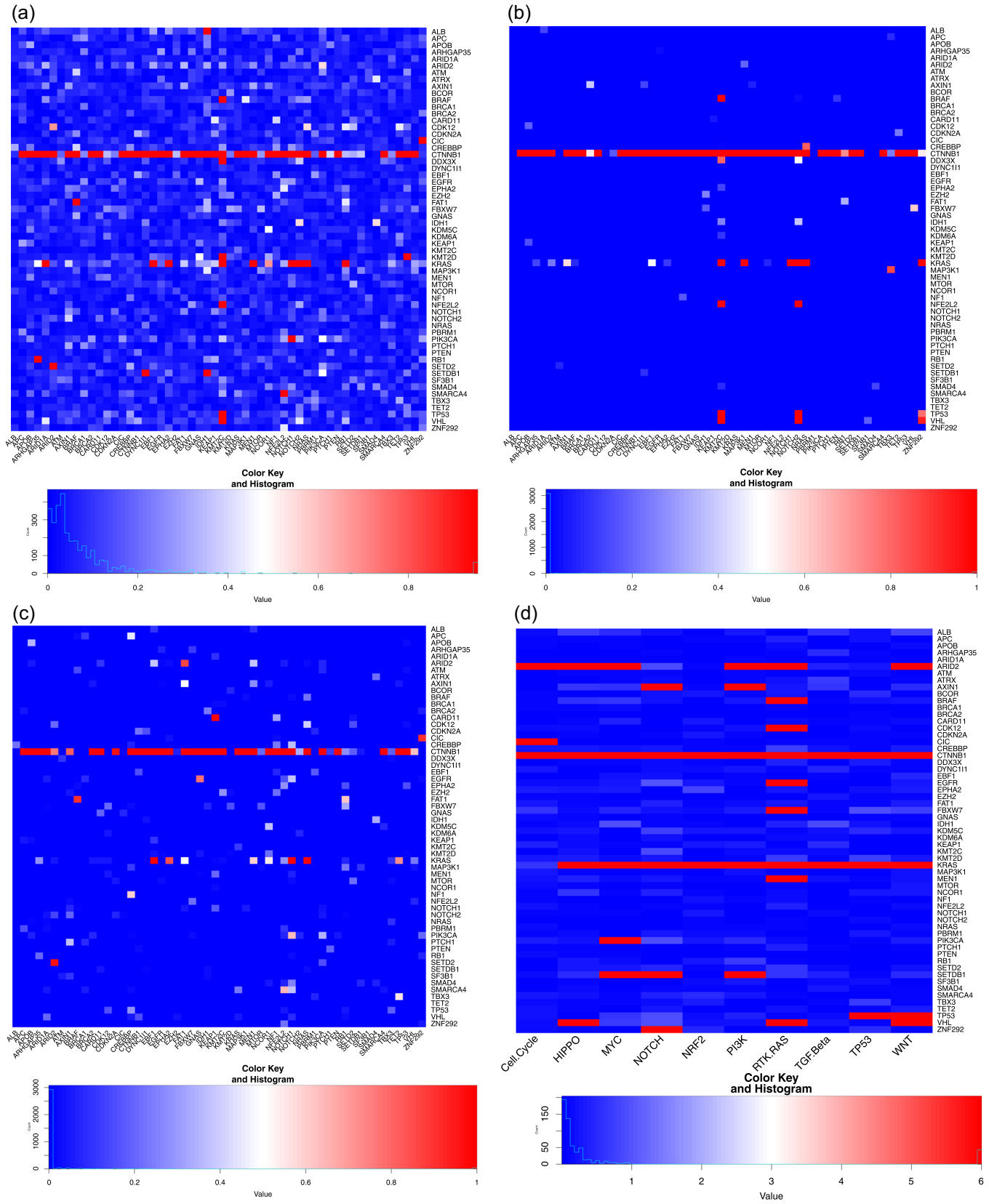

FIGURE 4 Heatmap of $-\log_{10}(\text{FDR})$ from (a) aSPU test, (b) Fisher's exact test, (c) MAGMA gene analysis, and (d) $-\log_{10}(\text{FDR}+10^{-6})$ from aSPUpath test between germline variations and somatic mutations of driver genes. The color range of the heatmaps is from blue to red: darker red of the cells in the heatmap indicates a more significant association. The threshold corresponding to FDR = 0.2 is $-\log_{10}(0.2) \approx 0.70$. FDR, false discovery rate; aSPU, adaptive sum of powered score; aSPUpath, data-adaptive pathway-based test.

Figure 4a against Figure 4b and Supplementary Information Figure S1A against Figure S1B, it is clear that aSPU detects many more significant associations or hotspots than Fisher's exact test for the same data set. Interestingly, although the significance signal from Fisher's exact tests in the heatmap is less than that from the aSPU test, we can still tell that somatic mutation genes *CTNNB1* and *KRAS* are two hotspots in the associations.

5.4 | aSPUpath test results

5.4.1 | Pathway-based association results

Then, we performed aSPUpath tests to study the associations between the somatic mutations in the same driver genes as used in the aSPU test and the germline variations in the 10 key cancer pathways (Sanchez-Vega, 2018). See Supplementary Information Table S3 in Section A.5. The number of pathway genes overlapping with our annotated germline variation genes is 259. For each association test, the data include all the SNPs in all the genes in a pathway (for the X matrix in Equation 3) and one gene's somatic mutation status (for the Y vector in Equation 3). We chose the candidate parameters γ_1 and γ_2 as in (13). The number of permutations is 1000.

The aSPUpath test results showed that the p-values from 188 out of 1560 association tests were less or equal to the nominal significance level 0.05 (a rate of 12.05%).

TABLE 3 aSPU *cis*/*trans* result.

	Significant ($p \leq 0.05$)	Not significant ($p > 0.05$)
cis	14 (9.5%)	133 (90.5%)
trans	1714 (7.4%)	21,539 (92.6%)

Abbreviation: aSPU, adaptive sum of powered score.

Supplementary Information Tables S9 and S9 showed the significant p-values. FDR multiple testing correction was then performed and the FDR values from 89 association tests were less or equal to 0.2.

We also created heatmaps for aSPUpath association tests between pathway germline variations and somatic mutations with raw p-values and FDR values. See Supplementary Information Figure S1C (p-values) and -Figure 4c (FDR values). The heatmaps for the aSPUpath results for both raw p-values and FDR values demonstrated that the somatic mutation genes *CTNNB1* and *KRAS* were associated with germline variations of almost all the pathways, which was consistent with the results from aSPU tests. We also found that the pathway RTK.RAS's germline variations are associated with the somatic mutations of a large proportion of genes.

5.4.2 | Dominant contributions to the associations

To investigate whether there exists a single gene in a pathway that dominates the contributions to the pathway–somatic association or a large proportion of genes in a pathway contributes to the associations, we performed the following tests. For each pathway, we performed aSPU tests on germline gene variations and somatic mutations and then ranked germline genes in the pathway based on the corresponding p-values.

The results showed that both situations existed. For example, in aSPUpath tests, we detected that germline variations in the TP53 pathway were significantly associated with *KRAS* somatic mutation (aSPUpath $p < 0.001$). In aSPU tests, the p-values from the associations between each germline gene in TP53 pathway and *KRAS* somatic mutation were given in Table 4 (part A). We found that the

TABLE 4 p-Values from the aSPU association tests between each germline gene in the TP53 pathway and the *KRAS* somatic mutation, and association tests between each germline gene in the MYC pathway and the *KRAS* somatic mutation.

(A) TP53 pathway					
Gene	**MDM4**	**CHEK2**	**MDM2**	**ATM**	**TP53**
p-value	0.000999	0.016983	0.0539461	0.2837163	0.4475524
(B) MYC pathway					
Gene	**MAX**	**MGA**	**MLXIP**	**MLXIPL**	
p-value	0.232767233	0.018981019	0.297702298	0.365634366	
Gene	**MNT**	**MXD1**	**MXD3**	**MXD4**	
p-value	0.733266733	0.85014985	0.02997003	0.045954046	
Gene	**MXI1**	**MYC**	**MYCL**	**MYCN**	
p-value	0.017982018	0.078921079	0.022977023	0.012987013	

Abbreviation: aSPU, adaptive sum of powered score.

gene *MDM4* made the dominant contribution to the association between the germline variations in the TP53 pathway and *KRAS* somatic mutation.

On the other hand, in the aSPUpath tests, we detected a significant association between MYC pathway germline variations and *KRAS* somatic mutation (aSPUpath, $p<0.001$). However, multiple germline genes in the MYC pathway contributed to the MYC pathway/KRAS somatic mutation association. See Table 4 (part B).

5.5 | Comparison of aSPU and aSPUpath tests with the other existing methods

Besides Fisher's exact test, with the same ICGC data set, we also tested other related existing methods, including MAGMA (Leeuw et al., 2015), SSU, UminP, HYST, and Gates-Simes for comparison. In general, these methods detected usually much smaller numbers of significant associations than those from aSPU and aSPUpath methods. In addition, some of the aforementioned five methods demonstrated the limitations of the association detections on both the gene level and the pathway level.

5.5.1 | MAGMA gene-based and gene-set-based analysis

MAGMA is a novel tool of gene and gene-set analysis for GWASs. It was shown that both the MAGMA gene analysis and the MAGMA gene-set analysis have significantly higher power than some other tools and they could detect more genes and gene sets in a real example (Leeuw et al., 2015). However, with the same ICGC data set, we discovered that the *p*-values of MAGMA gene analysis from only 1183 out of 23,400 association tests (a rate of 5.09%) were smaller than or equal to the nominal significance level 0.05 (aSPU: 1728 out of 23,400, a rate of 7.38%). After the FDR multiple testing correction, the number of the FDR values from MAGMA gene analysis less than 0.2 was 140 (aSPU: 210). Similarly, the *p*-values of MAGMA gene-set analysis from only 80 out of 1560 association tests (a rate of 5.13%) were smaller than or equal to the nominal significance level 0.05 (aSPUpath: 188 out of 1560, a rate of 12.05%). The number of the FDR values from MAGMA gene-set analysis less than 0.2 was only 2 (aSPUpath: 89). See the heatmaps for MAGMA gene analysis (FDR: Figure 4c and *p*-values: Supplementary Information Figure S1) and for MAGMA gene-set analysis (*p*-values and FDR: Figure 5).

In addition, one of the major limitations of the MAGMA is that we have to first calculate *p*-value for the association between each single SNP and each somatic mutation. However, due to the rank-deficient models for some associations between SNPs and somatic mutations, the *p*-values could not be obtained. This results in NAs of *p*-values. The percentage of NAs for the MAGMA gene analysis was 0.3328% (9516 NA *p*-values among 2,859,636 associations). The percentage of NAs for the MAGMA gene-set analysis was 0.4749% (30,888 NA *p* values among 6,504,420 associations). This type of information loss might contribute to the weak detection of the significant association tests.

5.5.2 | SSU and UminP

These two methods were designed for the gene-based association tests. We found that the *p*-values of SSU tests from 1693 out of 23,400 association tests were smaller than the nominal significance level 0.05 (a rate of 7.24%), which was slightly less than that with aSPU tests (a rate of 7.38%). The UminP tests showed 1729 out 23,400 association tests (7.39%) resulted in *p*-values less than 0.05, comparable to aSPU tests. After the FDR multiple testing correction, the number of FDR values less than 0.2 from HYST tests (146) was smaller than that from aSPU tests (210). So is the case with Gates-Simes tests (159). See the heatmaps for the SSU tests (FDR: Figure 6a) and for the UminP tests (FDR: Figure 6b).

5.5.3 | HYST and Gates-Simes

These two methods were designed for pathway-based association tests. Both HYST and Gates-Simes methods have the same limitation as MAGMA because the *p*-values for the association tests between a single SNP and the somatic mutation of a gene also need to be calculated first. We found that HYST and Gates-Simes detected the less significant numbers of associations than that from aSPUpath tests, especially after the FDR multiple testing correction. See the heatmaps for HYST tests (FDR: Figure 6c) and Gates-Simes tests (FDR: Figure 6d).

5.6 | Network of pathway germline variations and somatic mutations

To further investigate the interactions between germline variations and somatic mutations, we used the aSPUpath results to create a network via a software tool called Cytoscape (https://cytoscape.org/). Cytoscape is an open-source software tool to visualize the interactions from network views. We used Cytoscape to create a network of interactions between pathway germline variations and somatic mutations. See Figure 7.

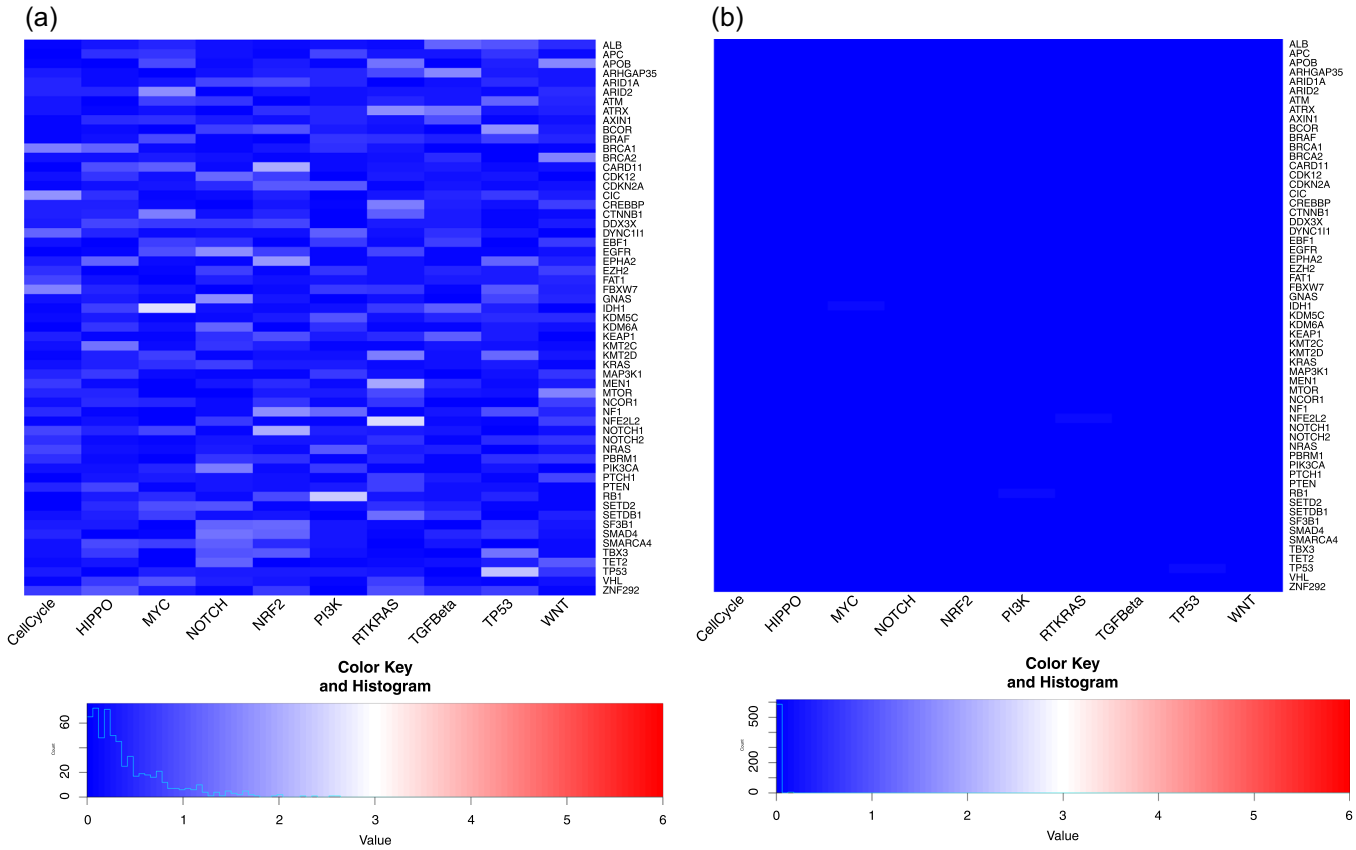

FIGURE 5 Heatmap of (a) $-\log_{10}(p\text{-value} + 10^{-6})$ and (b) $-\log_{10}(\text{FDR} + 10^{-6})$ from MAGMA gene-set analysis between germline variations and somatic mutations of driver genes. The color range of the heatmaps is from blue to red: darker red of the cells in the heatmap indicates a more significant association. The threshold corresponding to the *p*-value 0.05 is $-\log_{10}(0.05) \approx 1.3$. The threshold corresponding to FDR = 0.2 is $-\log_{10}(0.2) \approx 0.70$. FDR, false discovery rate.

The network figure illustrates that the RTK.RAS pathway is significantly associated with nine genes' somatic mutations: *CDK12, ARID2, FBXW7, EGFR, VHL, KRAS, CTNNB1, MEN1*, and *BRAF*. In the RTK.RAS pathway, RTK represents Receptor Tyrosine Kinases and RAS is a kind of small GTPase. We found several literatures that support our findings on the RTK.RAS pathway germline associations with some genes' somatic mutations:

- *KRAS* is a gene in the RAS family (K-RAS, N-RAS, and H-RAS) (Regad, 2015), which is a part of RTK.RAS pathway.
- *BRAF* is a gene in the RAF family (A-Raf, B-Raf, and C-Raf) (Regad, 2015), which is along the downstream of RTK.RAS pathway (Imperial et al., 2017).
- *EGFR*, *KRAS*, and *BRAF* are all clinical targets for lung cancer patients (Imperial et al., 2017).

Interestingly, the *CTNNB1* somatic mutation is associated with all 10 pathways. The *CTNNB1* gene in the WNT pathway encodes β-catenin and this gene's mutation happens in many cancers (Gao et al., 2018). The alternation of β-catenin protein tends to seriously reprogram the nuclear transcriptional network (Gao et al., 2018). It was reported by Gillard et al. (2017) that the *CTNNB1* hotspot somatic mutation is associated with the WNT- or PI3K-pathway activation.

5.7 | Discussions on hotspots

There are some interesting findings in our association studies. The somatic mutation genes *CTNNB1* and *KRAS* are hotspots because they are significantly associated with almost all the driver genes' and pathways genes' germline variations. The *CTNNB1* gene is a protein-coding gene, which encodes catenin beta 1 protein. This gene is part of the WNT pathway (Maharjan et al., 2018). *KRAS* is important in cell signal events. For example, it controls cell proliferation because it acts as an on/off switch in cell signaling (Pantsar, 2019). Also, the *KRAS* gene gains substantial attention for target therapy in cancer (Huang et al., 2021).

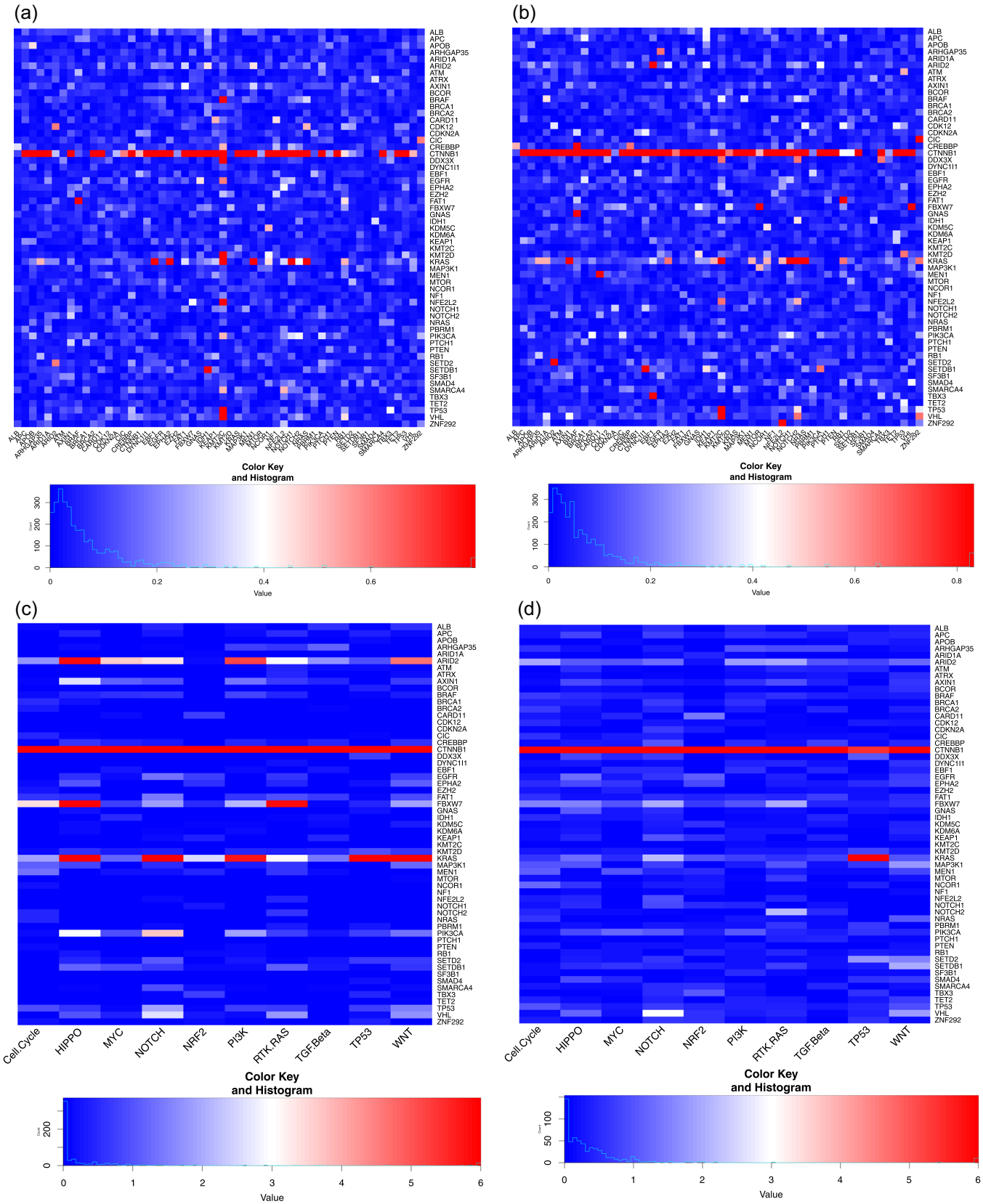

FIGURE 6 Heatmap of $-\log_{10}(\text{FDR})$ from (a) SSU test (b) UminP test and $-\log_{10}(\text{FDR}+10^{-6})$ from (c) HYST test (d) Gates-Simes test between germline variations and somatic mutations of driver genes. The color range of the heatmaps is from blue to red: darker red of the cells in the heatmap indicates a more significant association. The threshold corresponding to FDR = 0.2 is $-\log_{10}(0.2) \approx 0.70$. FDR, false discovery rate; HYST, hybrid set-based testGates-Simes, gene-based association test using extended Simes procedure; SSU, sum of squared score; UminP, univariate minimum *p*-value.

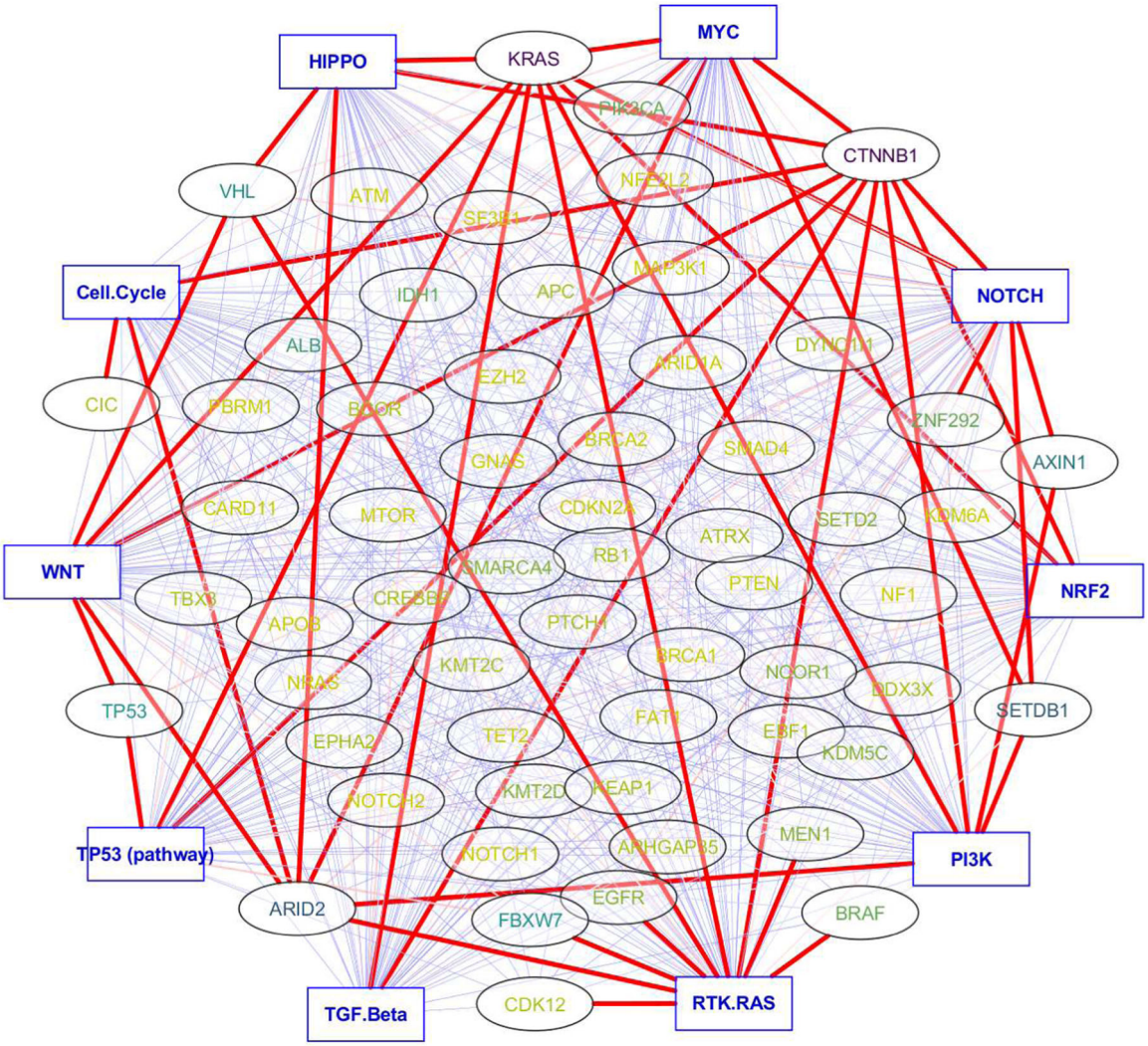

FIGURE 7 Network showing the association results from aSPUpath, where the labels in rectangles are for pathways and the labels in ellipses are for somatic mutation genes. The color of the label of each gene indicates how many pathways the gene has associations with. The darker the color is, the more associations it has. The thickness of each network edge indicates the significance of the association. The thicker the edge is, the stronger the association is. aSPUpath, data-adaptive pathway-based test.

Our findings of the interactions between germline variations and somatic mutations are supported by multiple literatures. For example, for the *CTNNB1* gene somatic mutation, it was shown by Yedid et al. (2016) that a germline mutation in the *APC* (Adenomatous polyposis coli) gene is associated with enhanced β-catenin activity (*CTNNB1* gene). In addition, both *APC* and *CTNNB1* genes are in the WNT pathway. The TP53 germline variation status is shown to be associated with the *CTNNB1* mutation (Pfaff et al., 2010). It was indicated by Mamidi et al. (2019b) that the *CTNNB1* somatic mutation is associated with multiple germline variations (including *RAS*) and other somatic mutation genes (including *PIK3CA*) in prostate cancer. It was reported that *CTNNB1* hotspot mutation is associated with WNT- or PI3K-pathway activation (Gillard et al., 2017). The inactivation of the gene *CDK12* is associated with CTNNB1 activating mutation in prostate cancer (Wu et al., 2018). As for the *KRAS* gene, the *KRAS* somatic mutation is shown to be associated with germline 10q22.3-q23.2 deletion in a patient with juvenile myelomonocytic leukemia (Yao et al., 2018). The somatic mutation of *KRAS* is more frequent in individuals with specific MHC-I genotypes (Ramroop et al., 2019).

6 | DISCUSSION

In this paper, we have adapted and developed two powerful data-adaptive large-sample score tests to study the interactions between germline variations and somatic mutations in human cancers. One is the data-adaptive

aSPU test, which aggregates SNPs into genes. The other is the pathway-based method aSPUpath, which aggregates genes into pathways. To accommodate different data situations, different weights and parameters are used to produce different sets of *p*-values, which are then combined to yield data-adaptive aSPU and aSPUpath tests for the interactions between somatic mutations and germline variations. The methods can be applied to multiple SNPs/genes/pathways and are more powerful than commonly used germline–somatic association test methods such as Fisher's exact test, SSU, UminP, HYST, and Gates-Simes. Simulation results show that both aSPU and aSPUpath have enhanced statistical powers compared with some conventionally used association models and maintain similar type I errors.

We incorporate the two data-adaptive test methods into the comprehensive analysis of large-scale ICGC data set for the association studies between germline variations and somatic mutations. Raw germline SNPs and somatic mutation data are processed through a sequence of screening and filtering techniques. Various association results are discovered between driver gene somatic mutations and germline variations. For example, aSPU results show that the *p*-values from a rate of 7.38% associations are less than the significance level of 0.05. The germline variation from the *CTNNB1* gene is associated with almost all of the driver somatic mutations we tested. We also detect other hotspots, such as *KRAS* and *KMR2C*. Our results indicate that the data-adaptive methods detect many more significant association signals between germline variations and somatic mutations than with Fisher's exact test (a rate of 2.88% association signals) and MAGMA gene analysis (a rate of 5.09%). The aSPUpath test is performed on the interactions between 10 key pathways' germline variations and the driver genes' somatic mutations, showing a rate of 12.05% significant associations, which is much higher than that with MAGMA gene-set analysis (a rate of 5.13%). In addition to *CTNNB1* and *KRAS*, we find that pathway RTK.RAS germline variations are associated with a large proportion of genes with somatic mutation.

In practice, one may think that the variation of pathways such as the inclusion or the exclusion of certain genes affects the results of our association tests. Regarding this, we performed an on/off test to see whether a significant association between a germline pathway and a somatic mutation is driven by a certain gene. For example, we did a test where we removed the most significant gene *MDM4* from the TP53 pathway. The germline TP53 pathway was still significantly associated with the *KRAS* somatic mutation ($p = 0.003$). On the pathway level, genes work collectively. Our pathway-based association tests consider the association between a germline pathway and somatic mutations. Thus, in general the variations of the pathway should not have significant impacts on this kind of association tests.

Our research in this paper provides valuable statistical tools and models for cancer studies, cancer prediction, and cancer therapy. The statistical tests give new insights into the associations among multiple aspects of multiple cancer types. For example, if certain germline variations are observed in a patient, our association studies can be used to tell the chance of also having some somatic mutations. According to Knudson's "two-hit" theory, we can predict if the patient is likely to have a certain cancer or not. Thus, our research is useful for cancer risk prediction and can help medical researchers focus on specific genes in cancer studies. The work also helps researchers better understand the molecular mechanisms of specific cancer genes and brings new insights into the development of novel cancer therapy.

ACKNOWLEDGMENTS

We appreciate the comments and suggestions from the two anonymous referees which helped greatly improve the manuscript. We thank Il-Youp Kwak for sharing R codes related to the aSPU package. We also thank Mei-Ju Chen for her helpful discussions on preprocessing the ICGC data.

DATA AVAILABILITY STATEMENT

The ICGC data are provided at https://www.nature.com/articles/s41467-020-16785-6#data-availability. The data that support the findings of this study are available from International Cancer Genome Consortium. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from https://www.nature.com/articles/s41467-020-16785-6#data-availability with the permission of International Cancer Genome Consortium. Relevant codes are available from the following link: https://github.com/chenstatistics/germsom.

ORCID

Zhongyuan Chen http://orcid.org/0009-0001-3209-0894
Peng Wei http://orcid.org/0000-0001-7758-6116

REFERENCES

Barfield, R., Qu, C., Steinfelder, R. S., Zeng, C., Harrison, T. A., Brezina, S., Buchanan, D. D., Campbell, P. T., Casey, G., Gallinger, S., Giannakis, M., Gruber, S. B., Gsur, A., Hsu, L., Huyghe, J. R., Moreno, V., Newcomb, P. A., Ogino, S., Phipps, A. I., ... Peters, U. (2022). Association between germline variants and somatic mutations in colorectal cancer. *Scientific Reports*, *12*, 10207.

Carter, H., Marty, R., Hofree, M., Gross, A. M., Jensen, J., Fisch, K. M., Wu, X., DeBoever, C., Van Nostrand, E. L., Song, Y., Wheeler, E., Kreisberg, J. F., Lippman, S. M., Yeo, G. W., Gutkind, J. S., & Ideker, T. (2017). Interaction landscape of inherited polymorphisms with somatic events in cancer. *Cancer Discovery*, *7*(4), 410–423.

Chan, K., & Gordenin, D. A. (2015). Clusters of multiple mutations: Incidence and molecular mechanisms. *The Annual Review of Genetics*, *49*, 243–267.

Chen, Z., Liang, H., & Wei, P. (2023). Efficient randomized low-rank parameter pre-selection for pathway-based test methods. Preprint.

Chen, Z. S., Wen, W., Beeghly-Fadiel, A., Shu, X.-o., Díez-Obrero, V., Long, J., Bao, J., Wang, J., Liu, Q., Cai, Q., Moreno, V., Zheng, W., & Guo, X. (2019). Identifying putative susceptibility genes and evaluating their associations with somatic mutations in human cancers. *American Journal of Human Genetics*, *105*, 477–492. https://doi.org/10.1016/j.ajhg.2019.07.006

Churchill, G. A., & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, *130*, 963–971.

Dworkin, A. M., Ridd, K., Bautista, D., Allain, D. C., Iwenofu, O. H., Roy, R., Bastian, B. C., & Toland, A. E. (2010). Germline variation controls the architecture of somatic alterations in tumors. *PLoS Genetics*, *6*(9), e1001136.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association*, *91*, 674–688.

Gao, C., Wang, Y., Broaddus, R., Sun, L., Xue, F., & Zhang, W. (2018). Exon 3 mutations of CTNNB1 drive tumorigenesis: A review. *Oncotarget*, *9*(4), 5492–5508.

Gillard, M., Lack, J., Pontier, A., Gandla, D., Hatcher, D., Sowalsky, A. G., Rodriguez-Nieves, J., Griend, D. V., Paner, G., & VanderWeele, D. (2017). Integrative genomic analysis of coincident cancer foci implicates CTNNB1 and PTEN alterations in ductal prostate cancer. *European Urology Focus*, *5*(3), 433–442.

Gui, H., Li, M., Sham, P. C., & Cherny, S. S. (2011). Comparisons of seven algorithms for pathway analysis using the WTCCC Crohn's disease dataset. *BMC Research Notes*, *4*, 386.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., ... Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, *48*, 245–252. https://doi.org/10.1038/ng.3506

Huang, L., Guo, Z., Wang, F., & Fu, L. (2021). KRAS mutation: From undruggable to druggable in cancer. *Signal Transduction and Targeted Therapy*, *6*, 386.

Imperial, R., Toor, O. M., Hussain, A., Subramanian, J., Masood, A. (2017). Comprehensive pancancer genomic analysis reveals (RTK)-RAS-RAF-MEK as a key dysregulated pathway in cancer: Its clinical implications. *Cancer Biology*, *54*, 14–28.

Jonsson, G., Naylor, T. L., Vallon-Christersson, J., Staaf, J., Huang, J., Renee Ward, M., Greshock, J. D., Luts, L., Olsson, H., Rahman, N., Stratton, M., Ringnér, M., Borg, A., & Weber, B. L. (2005). Distinct genomic profiles in hereditary breast tumors identified by array-based comparative genomic hybridization. *Cancer Research*, *65*, 7612–7621.

Knudson, Jr., A. G. (1971). Mutation and cancer: Statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, *68*, 820–823. https://doi.org/10.1073/pnas.68.4.820

Kwak, I.-Y., et al. (2021). *aSPU: Adaptive sum of powered score test*. https://cran.r-project.org/web/packages/aSPU/index.html

LaFramboise, T., Dewal, N., Wilkins, K., Pe'er, I., & Freedman, M. L. (2010). Allelic selection of amplicons in glioblastoma revealed by combining somatic and germline analysis. *PLoS Genet*, *6*(9), e1001086.

Landi, M. T., Bauer, J., Pfeiffer, R. M., Elder, D. E., Hulley, B., Minghetti, P., Calista, D., Kanetsky, P. A., Pinkel, D., & Bastian, B. C. (2006). MC1R germline variants confer risk for BRAF-mutant melanoma. *Science*, *313*(5786), 521–522.

Leeuw C., Mooij, J., Heskes, T., & Posthuma, D. (2015). MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Computational Biology*, *11*(4), e1004219. https://doi.org/10.1371/journal.pcbi.1004219

Li, M. X., Gui, H. S., Kwan, J. S., & Sham, P. C. (2011). GATES: A rapid and powerful gene-based association test using extended Simes procedure. *American Journal of Human Genetics*, 88, 283–293.

Li, M. X, Kwan, J. S., Sham, P. C. (2012). HYST: A hybrid set-based test for genome-wide association studies, with application to protein–protein interaction-based association analysis. *American Journal of Human Genetics*, *91*, 478–488.

Ma, Y., & Wei, P. (2019). FunSPU: A versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing data. *PLoS Genetics*, *15*(4), e1008081.

Maharjan, R., Backman, S., Akerström, T., Hellman, P., & Björklund, B. (2018). Comprehensive analysis of CTNNB1 in adrenocortical carcinomas: Identification of novel mutations and correlation to survival. *Scientific Reports*, *8*, 8610.

Maldonado, J. L., Fridlyand, J., Patel, H., Jain, A. N., Busam, K., Kageshita, T., Ono, T., Albertson, D. G., Pinkel, D., & Bastian, B. C. (2003). Determinants of BRAF mutations in primary melanomas. *Journal of the National Cancer Institute*, *95*, 1878–1890.

Mamidi, T. K. K., Wu, J., & Hicks, C. (2019a). Integrating germline and somatic variation information using genomic data for the discovery of biomarkers in prostate cancer. *BMC Cancer*, *19*, 229.

Mamidi, T. K. K., Wu, J., & Hicks, C. (2019b). Interactions between germline and somatic mutated genes in aggressive prostate cancer. *Prostate Cancer*, *2019*, 4047680. PMID: 31007957; PMCID: PMC6441536. https://doi.org/10.1155/2019/4047680

Middlebrooks, C. D., Rouf Banday, A., Matsuda, K., Udquim, K.-I., Onabajo, O. O., Paquin, A., Figueroa, J. D., Zhu, B., Koutros, S., Kubo, M., Shuin, T., Freedman, N. D., Kogevinas, M., Malats, N., Chanock, S. J., Garcia-Closas, M., Silverman, D. T., Rothman, N., & Prokunina-Olsson, L. (2016). Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nature Genetics*, *48*(11), 1330–1338.

Nik-Zainal, S., Wedge, D. C., Alexandrov, L. B., Petljak, M., Butler, A. P., Bolli, N., Davies, H. R., Knappskog, S., Martin, S., Papaemmanuil, E., Ramakrishna, M., Shlien, A., Simonic, I., Xue, Y., Tyler-Smith, C., Campbell, P. J., & Stratton, M. R. (2014). Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-

dependent mutations in breast cancer. *Nature Genetics*, *46*(5), 487–491.

Pan, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology*, *33*, 497–507.

Pan, W., Kim, J., Zhang, Y., Shen, X., & Wei, P. (2014). A powerful and adaptive association test for rare variants. *Genetics*, *197*, 1081–1095.

Pan, W., Kwak, I., & Wei, P. (2015). A powerful pathway-based adaptive test for genetic association with common or rare variants. *American Journal of Human Genetics*, *97*, 86–98.

Pan, W., & Shen, X. (2011). Adaptive tests for association analysis of rare variants. *Genet Epidemiology*, *35*, 381–388.

Pantsar, T. (2019). The current understanding of KRAS protein structure and dynamics. *Computational and Structural Biotechnology Journal*, *18*, 189–198.

Pfaff, E., Remke, M., Sturm, D., Benne, A., Witt, H., von Bueren, T. M. O., Wittmann, A., Schöttler, A., Jorch, N., Graf, N., Kulozik, A. E., Witt, O., Scheurlen, W., von Deimling, A., Rutkowski, S., Taylor, M. D., Tabori, U., Lichter, P., Korshunov, A., & Pfister, S. M. (2010). *TP53* mutation is frequently associated with *CTNNB1* mutation or *MYCN* amplification and is compatible with long-term survival in medulloblastoma. *Journal of Clinical Oncology*, *28*(35), 5188–5196. https://doi.org/10.1200/JCO.2010.31.1670

Ramos, A. H., Lichtenstein, L., Gupta, M., Lawrence, M. S., Pugh, T. J., Saksena, G., Meyerson, M., & Getz, G. (2015). Oncotator: Cancer variant annotation tool. *Human Mutation*, *36*(4), E2423–E2429. https://doi.org/10.1002/humu.22771

Ramroop, J. R., Gerber, M. M., & Toland, A. E. (2019). Germline variants impact on somatic events during tumorigenesis. *Trend in Genetics*, *35*(7), 515–526. https://doi.org/10.1016/j.tig.2019.04.005

Regad, T. (2015). Targeting RTK signaling pathways in cancer. *Cancers*, *7*, 1758–1784.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., Dimitriadoy, S., Liu, D. L., Kantheti, H. S., Saghafinia, S., Chakravarty, D., Daian, F., Gao, Q., Bailey, M. H., Liang, W.-W., Foltz, S. M., Shmulevich, I., Ding, L., Heins, Z., … Schultz, N. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell*, *173*, 321–337.

Stefansson, O. A., Jonasson, J. G., Johannsson, O. T., Olafsdottir, K., Steinarsdottir, M., Valgeirsdottir, S., & Eyfjord, J. E. (2009). Genomic profiling of breast tumours in relation to BRCA abnormalities and phenotypes. *Breast Cancer Research*, *11*, R47.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*, 82–93. https://doi.org/10.1038/s41586-020-1969-6

Tippett, L. H. C. (1931). *The methods of statistics*. Williams and Norgate Ltd.

Vali-Pour, M., Lehner, B., Supek, F. (2022). The impact of rare germline variants on human somatic mutation processes. *Nature Communications*, *13*, 3724

Vosoughi, A., Zhang, T., Shohdy, K. S., Vlachostergios, P. J., Wilkes, D. C., Bhinder, B., Tagawa, S. T., Nanus, D. M., Molina, A. M., Beltran, H., Sternberg, C. N., Motanagh, S., Robinson, B. D., Xiang, J., Fan, X., Chung, W. K., Rubin, M. A., Elemento, O., Sboner, A., … Faltas, B. M. (2020). Common germline–somatic variant interactions in advanced urothelial cancer. *Nature Communications*, *11*, 6195.

Wang, T., & Elston, R. C. (2007). Improved power by use of a weighted score test for linkage disequilibrium mapping. *American Journal of Human Genetics*, *80*, 353–360.

Waszak, S. M., Tiao, G., Zhu, B., Rausch, T., Muyas, F., Rodríguez-Martín, B., Rabionet, R., Yakneen, S., Escaramis, G., Li, Y., Saini, N., Roberts, S. A., Demidov, G. M., Pitkänen, E., Delaneau, O., Heredia-Genestar, J. M., Weischenfeldt, J., Shringarpure, S. S., Chen, J., … the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network. (2017). Germline determinants of the somatic mutation landscape in 2,642 cancer genomes. bioRxiv. https://doi.org/10.1101/20833

Wei, P., Cao, Y., Zhang, Y., Xu, Z., Kwak, I.-Y., Boerwinkle, E., & Pan, W. (2016). On robust association testing for quantitative traits and rare variants. *G3 (Bethesda)*, *6*, 3941–3950. https://doi.org/10.1534/g3.116.035485

Wu, Y. M., Cieślik, M., Lonigro, R. J., Vats, P., Reimers, M. A., Cao, X., Ning, Y., Wang, L., Kunju, L. P., de Sarkar, N., Heath, E. I., Chou, J., Feng, F. Y., Nelson, P. S., de Bono, J. S., Zou, W., Montgomery, B., Alva, A., Chinnaiyan, A. M., … PCF/SU2C International Prostate Cancer Dream Team. (2018). Inactivation of CDK12 delineates a distinct immunogenic class of advanced prostate cancer. *Cell*, *173*(7), 1770–1782.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, *89*, 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

Xu, Z., Wu, C., Wei, P., & Pan, W. (2017). A powerful framework for integrating eQTL and GWAS summary data. *Genetics*, *207*, 893–902. https://doi.org/10.1534/genetics.117.300270

Yang, T., Chen, H., Tang, H., Li, D., & Wei, P. (2019). A powerful and data-adaptive test for rare-variant-based gene-environment interaction analysis. *Statistics in Medicine*, *38*, 1230–1244. https://doi.org/10.1002/sim.8037

Yao, R., Yu, T., Xu, Y., Li, G., Yin, L., Zhou, Y., Wang, J., Yan, Z. (2018). Concurrent somatic KRAS mutation and germline 10q22.3-q23.2 deletion in a patient with juvenile myelomonocytic leukemia, developmental delay, and multiple malformations: A case report. *BMC Medical Genomics*, *11*(1), 60. https://doi.org/10.1186/s12920-018-0377-3.

Yedid, N., Kalma, Y., Malcov, M., Amit, A., Kariv, R., Caspi, M., Rosin-Arbesfeld, R., & Ben-Yosef, D. (2016). The effect of a germline mutation in the APC gene on β -catenin in human embryonic stem cells. *BMC Cancer*, *16*, 952.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Chen, Z., Liang, H., & Wei, P. (2023). Data-adaptive and pathway-based tests for association studies between somatic mutations and germline variations in human cancers. *Genetic Epidemiology*, 1–20. https://doi.org/10.1002/gepi.22537