


RESEARCH

Open Access



Integrative eQTL-weighted hierarchical Cox models for SNP-set based time-to-event association studies

Haojie Lu^{1†}, Yongyue Wei^{2†}, Zhou Jiang¹, Jinhui Zhang¹, Ting Wang¹, Shuiping Huang^{1,3,4} and Ping Zeng^{1,3,4*} 

Abstract

Background: Integrating functional annotations into SNP-set association studies has been proven a powerful analysis strategy. Statistical methods for such integration have been developed for continuous and binary phenotypes; however, the SNP-set integrative approaches for time-to-event or survival outcomes are lacking.

Methods: We here propose IEHC, an integrative eQTL (expression quantitative trait loci) hierarchical Cox regression, for SNP-set based survival association analysis by modeling effect sizes of genetic variants as a function of eQTL via a hierarchical manner. Three p-values combination tests are developed to examine the joint effects of eQTL and genetic variants after a novel decorrelated modification of statistics for the two components. An omnibus test (IEHC-ACAT) is further adapted to aggregate the strengths of all available tests.

Results: Simulations demonstrated that the IEHC joint tests were more powerful if both eQTL and genetic variants contributed to association signal, while IEHC-ACAT was robust and often outperformed other approaches across various simulation scenarios. When applying IEHC to ten TCGA cancers by incorporating eQTL from relevant tissues of GTEx, we revealed that substantial correlations existed between the two types of effect sizes of genetic variants from TCGA and GTEx, and identified 21 (9 unique) cancer-associated genes which would otherwise be missed by approaches not incorporating eQTL.

Conclusion: IEHC represents a flexible, robust, and powerful approach to integrate functional omics information to enhance the power of identifying association signals for the survival risk of complex human cancers.

Keywords: Integrative analysis, SNP-set association study, Joint effect test, Hierarchical modeling, Cox model, Expression quantitative trait loci, Aggregated Cauchy association test

Background

A wide range of recent genome-wide association studies (GWASs) have revealed that germline variants (i.e., single nucleotide polymorphisms [SNPs]) are also an important inherited component of cancer risk [1–3], although

somatic mutations (e.g., copy number and DNA methylation alterations) play an essential role in the pathophysiology of many human cancers [4–6]. Conventionally, the association of SNPs is examined one at a time in cancer GWASs [1, 2]; however, the power for detecting such single SNP association signal remains limited because genetic variants generally have weak effect sizes [7–9], making the detection of cancer-associated SNPs difficult even with large samples. In addition, these identified genetic variants often explain only a very small fraction of cancer predisposition, leading to the so-called missing heritability [10–14], which also implies that a large

*Correspondence: zpstat@xzhmu.edu.cn

[†]Haojie Lu and Yongyue Wei wish it to be known that, in their opinion, the first two authors should be regarded as joint first authors

¹ Department of Biostatistics, School of Public Health, Xuzhou Medical University, Xuzhou 221004, Jiangsu, China

Full list of author information is available at the end of the article



amount of causal loci have yet been discovered and the endeavor to identify causative genes for cancers should continue.

As an effective alternative strategy, SNP-set analysis has been proposed in GWAS [15–21], where a set of SNPs defined a priori within a gene or other genetic units (e.g., pathway) are analyzed collectively to assess their joint influence on diseases or traits. Existing SNP-set approaches can be roughly grouped into two categories: (i) the burden test, in which the association with disease risk is evaluated for the overall effect of a weighted summation of variant alleles [22, 23]; and (ii) the variance component test, in which the association is examined for the variance of genetic variants under the framework of mixed-effects models [16, 24]. Due to the aggregation of multiple weak association signals and the reduced burden of multiple testing, SNP-set analysis is often more powerful than its counterpart of single SNP analysis. However, these SNP-set association approaches might be still underpowered when additional informative knowledge is available about the alternative. For example, if the association between a set of genetic variants and the survival risk of cancers is regulated through gene expression, the power improvement would be further achieved by integrating transcriptomic data into the test method. As it is widely demonstrated that disease-associated SNPs are more likely to be expression quantitative trait loci (eQTL) [25], it is thus conceivable that incorporating such knowledge would increase power for detecting association [26–28].

Several methods have been proposed for this goal within the mixed-effects model framework. For example, MiST was developed for continuous and binary phenotypes in rare variant association studies by modeling the effects of rare variants as a function of functional features while allowing for the heterogeneity of variant-specific effects [29]. This method was recently further generalized to integrate eQTL or other functional annotations [26, 30, 31]. Both simulations and real applications have exhibited the advantage of these integrative approaches compared with the general methods that do not incorporate functional characteristics of genetic variants. However, to our best knowledge, there is little relevant work with regards to integrative approaches for time-to-event association studies.

In the present study, we develop such a method within the hierarchical Cox model framework to jointly analyze multiple SNPs for association with censored survival outcomes (i.e., time-to-event phenotypes) [32, 33]. Specifically, we first group SNPs into SNP-sets based on a biologically meaningful unit (i.e., genes), and then test for the overall joint effects of all SNPs within the gene. To integrate eQTL, following prior work [26, 29], we suppose

the effect sizes of SNPs are partly explained by eQTL via a hierarchical modeling. As a result, our association analysis consists of two components: the first component stands for the fixed effect through the weighted burden score to reflect the impact of genetic variants on the survival risk explained by eQTL, while the second component examines the residual effects of genetic variants beyond eQTL. These residual effects are treated as random effects following an arbitrary distribution with mean zero and variance τ [32, 33]. Therefore, methodologically, testing the joint effect for a group of SNPs with the survival risk of a cancer of focus is equivalent to examining the fixed effect and random effects simultaneously.

Under our model context, a novel decorrelated modification is made so that two independent statistics (i.e., a burden test statistic and a variance component test statistic) are derived for each of the two components. Then, the joint effect test can be easily constructed based on these two uncorrected statistics via various p-values combination strategies. To this aim, we consider three data-driven approaches (e.g., the Fisher's combination, the optimally weighted combination, and the adaptively weighted combination) for combining them to capture the association signals from both sources. To further enhance power, we exploit the recently developed aggregated Cauchy association test (ACAT) to integrate the strengths of all the five types of test methods (i.e., three combination tests as well as the burden test and the variance component test; the latter was also called the kernel machine [KM] test) [34, 35]. We refer to our proposed approach and test framework described above as integrative eQTL hierarchical Cox model (IEHC). Extensive simulations demonstrate that the three combination tests have comparable power or are better than both the burden and variance component tests under some specific scenarios, while IEHC-ACAT enjoys consistently higher power across all simulation scenarios. We finally apply IEHC to ten TCGA cancers which have one explicit relevant tissue in The Genotype-Tissue Expression (GTEx) project and integrate eQTL into our method [36]. We identified a total of 21 (9 unique) cancer-associated genes which would otherwise be missed by the general SNP-set based survival association methods that do not consider eQTL.

Methods

An overview of the IEHC model and the joint test

First, consider that there are S genotypes (denoted by G_i and coded as 0, 1 or 2 in terms of the number of effect allele) of SNPs located within a given gene and p covariates X_i (e.g., age, gender, and cancer stage) for n individuals; and S in general varies gene by gene. In addition, denote the observed survival time by t_i and

the true survival time by T_i with d_i indicating the censored status; that is, $d_i=1$ if $T_i=t_i$, whereas $d_i=0$ if $T_i < t_i$. Under the proportional hazards condition, we assume the hazard function $\lambda(t)$ of the survival time t_i is related to G_i and X_i through the classical Cox model [37]

$$\log\left(\frac{\lambda(t_i)}{\lambda_0(t_i)}\right) = G_i^T \alpha + X_i^T c$$

where λ_0 is an arbitrary baseline hazard function, $\alpha = (\alpha_1, \dots, \alpha_S)$ is an S -vector of effect sizes for SNPs and $c = (c_1, c_2, \dots, c_p)$ is a p -vector of fixed effect sizes for clinical covariates.

We here only provide an overview of IEHC, with technical details demonstrated in the Additional file 1. In brief, IEHC examines a group of SNPs in one gene at each time and integrates eQTL information by extending the Cox model above in a hierarchical manner

$$\log\left(\frac{\lambda(t_i)}{\lambda_0(t_i)}\right) = \sum_{j=1}^S G_{ij} \alpha_j + \sum_{k=1}^p X_{ik}^T c_k = \eta = G_i^T \alpha + X_i^T c$$

$$\alpha_j = \beta_j \times \theta + b_j$$

$$b_j \sim N(0, \tau)$$

Of note, plugging α_j into the first line leads to $\eta = (G_i^T \beta) \theta + G_i^T b + X_i^T c$. In the above, β_j is the known eQTL effect size of the j^{th} SNP and directly obtained in terms of summary statistics from the GTEx project [36, 38], θ is a scale of coefficient for eQTL and quantifies the association between the survival risk and the weighted burden score $G_i^T \beta$, and b_j is the normal residual variant-specific effect size that is not interpreted by eQTL alone. Then, the hypothesis of no association between a set of SNPs and the survival outcome is

$$H_0 : \theta = 0 \text{ and } b = 0 \Leftrightarrow H_0 : \theta = 0 \text{ and } \tau = 0$$

This is a joint test including both fixed effect and random effects: the first component examines the influence of genetic variants on the survival risk explained by eQTL (i.e., $\theta=0$); while the second component examines the impact of genetic variants beyond the effects of eQTL (i.e., $\tau=0$).

To implement the hypothesis testing while circumventing the potential correlation between statistics and improving the statistical computation, we propose the following two-stage strategy. Briefly, we derive the test statistic for θ under $H_0: \theta=0$ and $\tau=0$ as usual, while derive the score statistic for τ under $\tau=0$ but without the constraint of $\theta=0$. By doing this, we ensure that these two statistics are independent (see simulation results in Additional file 1). This strategy substantially

eases the construction of test statistics for the joint test and two asymptotically independent statistics are eventually derived: one for θ in the general Cox model (say U_θ) [37] and the other for the variance component parameter τ in the kernel machine (KM) Cox model (say U_τ) [32, 33]. We combine the two uncorrelated statistics via several aggregation approaches, including the Fisher’s combination (IEHC-Fisher) [39, 40], the optimally weighted linear combination (IEHC-optim) as well as the adaptively weighted linear combination (IEHC-adapt). For IEHC-optim we establish $T_\rho = \rho U_\theta + (1-\rho) U_\tau$, with $\rho \in [0, 1]$ controlling the contribution of the fixed-effect component. The final ρ in IEHC-optim is selected by optimizing T_ρ . On the other hand, IEHC-adapt is a data-adaptive generalization of the Fisher’s combination [39, 40], for which the test statistic takes the form $T = \rho_\theta Z_\theta + \rho_\tau Z_\tau$, where $Z_\theta = -2\log(p_\theta)$ and $Z_\tau = -2\log(p_\tau)$, based on which ρ_θ and ρ_τ are determined via an adaptive manner.

The IEHC test described above includes two special cases: the burden test for examining the fixed effect θ (with $\tau=0$) in the general Cox model and the KM test examining the variance component parameter τ (with $\theta=0$) in the KM Cox model. To further boost the power, we employ the recently developed aggregated Cauchy association test (ACAT) to combine the strengths of these five methods (i.e., the burden test, the KM test and three joint tests including IEHC-Fisher, IEHC-optim and IEHC-adapt) [34, 35]. The advantage of IEHC-ACAT is that it allows us to aggregate correlated p-values obtained from multiple various tests into a single well-calibrated p-value while maintaining the type I error control correctly. The detailed procedures for these approaches are relegated to Additional file 1. The code for IEHC is freely available at <https://github.com/biostatpzen/IEHC>.

Simulations for type I error control and power evaluation

We now perform simulations to evaluate the type I error control and power for IEHC. To mimic the truth, we undertook simulations based on realistic genotypes available from the Geuvadis program because the sample size in the real-life applications used in this paper matched closely that of Geuvadis [41]. First, we obtained 550 a group of correlated SNPs in a local genetic region from 465 individuals in Geuvadis. During the simulation we randomly selected S nearby SNPs (denoted by G_1), with S varying according to a uniform distribution ranging from 20 to 50 (i.e., S was on average equal to 35); among these selected genetic variants we further randomly set 0%, 30% or 50% of SNPs having zero effect sizes. We generated the gene expression level with the first 165

Table 1 Summary information of 10 TCGA cancers and the number of genes, sample sizes and SNPs of these cancers after combining the tissue in GTEx and quality control

Cancer	N_0	N_1	m	Age	Female/male	Censored rate (%)	Stage or grade (1/2/3/4/5)	Tissue in GTEx	N_2	k_0	k_1
ACC	97	75	4,473,001	47.6 ± 16.5	50/25	60.0	8/33/16/18/0	Adrenal gland	175	11,822	11,165
BRCA	1283	736	2,281,892	58.8 ± 13.0	736/0	85.9	138/408/174/11/5	Breast mammary tissue	251	13,068	7,793
COAD	570	201	4,216,239	66.0 ± 13.0	97/104	75.1	34/78/62/27/0	Colon transverse	246	12,779	11,586
LHC	469	166	3,369,784	62.7 ± 14.1	73/93	60.8	79/44/39/4/0	Liver	153	11,073	8,595
LUAD	877	384	3,881,580	65.9 ± 9.9	213/171	63.0	216/90/62/16/0	Lung	383	13,300	10,617
LUSC	765	344	3,279,532	67.1 ± 8.8	94/250	57.3	176/117/48/3/0	Lung	383	13,300	9,450
OV	758	455	1,527,607	60.2 ± 11.4	455/0	37.4	9/19/353/74/0	Ovary	122	12,623	7,851
PAAD	223	159	4,679,901	65.6 ± 10.8	69/90	45.9	20/130/4/5/0	Pancreas	220	11,173	10,728
STAD	544	249	3,384,736	64.8 ± 10.2	97/152	60.2	33/75/128/13/0	Stomach	237	12,045	9,247
UCEC	605	368	4,008,620	64.4 ± 10.8	368/0	83.2	239/33/79/17/0	Uterus	101	12,592	10,948

N_0 : the initial sample size in TCGA; N_1 : the sample size after quality control; m : the number of SNPs after combination; N_2 : the number of genes after combination; k_0 : the number of genes after quality control

individuals and sampled the effect sizes β from a normal distribution with a special variance so that the proportion of the explained variation (PVE) of the expression level would be 30% or 50%.

Then, we calculated $\alpha = \beta \times \theta + \mathbf{b}$, with \mathbf{b} following a normal distribution with variance τ . Two independent covariates (i.e., X_1 was binary and X_2 was continuous) were also generated with each having an effect size of 0.50. We employed the inverse probability method to generate the survival time which followed a Weibull distribution with the shape parameter equal to 1 and the scale parameter equal to 0.01 [42]. The location parameter (denoted by μ) of this Weibull distribution was determined by α and the two covariates: $\mu = \exp(\eta)$ and $\eta = \mathbf{G}_2\alpha + 0.5X_1 + 0.5X_2$, with \mathbf{G}_2 representing the remaining genotypes of 300 samples in Geuvadis. The censored rate was fixed to be 50% in a random manner. Note that, this relatively high censored rate corresponded to the similar situation observed in the TCGA cancer dataset (see below). We set $\theta = 0$ and $\tau = 0$ to assess the type I error control and run 10^5 replications. To evaluate the power, we specified $\theta = 0, 0.1, 0.2, 0.3$ or 0.4 , and $\tau = 0, 0.02$, or 0.04 (here at least one of θ and τ was nonzero). The power simulation was repeated 10^3 times.

TCGA cancers and GTEx eQTL summary statistics

TCGA cancers and quality control

We applied the proposed method to multiple cancer data publicly available from TCGA [43]. We downloaded these datasets at <https://xenabrowser.net/> and focused on cancers having one explicitly relevant tissue in the GTEx project [36]. However, we did not include PRAD (prostate adenocarcinoma) and THCA (thyroid carcinoma) as nearly all the PRAD patients (99.3% = 146/147) and THCA patients (95.7% = 315/329) were alive during the follow-up. We also removed DLBC (lymphoid neoplasm diffuse large b-cell lymphoma), KICH (kidney chromophobe) and TGCT (testicular germ cell tumor) because of too small sample sizes (i.e., only 24 for DLBC, 57 for KICH and 69 for TGCT). Finally, we reserved ten cancers for further analysis (Table 1).

To avoid the issue of ethnic heterogeneity, we included only patients of European ancestry and selected the overall survival time and status in our analysis following prior work [44]. Several important clinical covariates were incorporated, such as age, gender, and pathologic tumor stage because only these clinical variables were available for the majority of TCGA patients. When the pathologic tumor stage is unavailable, we alternatively employed the clinical stage (i.e., OV). We further standardized each clinical covariate. In addition, for every cancer we only kept samples from primary tumor tissues and excluded patients with too many missing values in clinical covariates (Table 1).

TCGA genotypes, imputation, and quality control

For each cancer we first filtered out SNPs that had missingness rate > 0.95 across patients, genotype calling rate < 0.95 , minor allele frequency (MAF) > 0.01 , or Hardy–Weinberg equilibrium (HWE) p -value $< 10^{-4}$. Then, we undertook imputation by first phasing genotypes with SHAPEIT [45], then imputed SNPs based on the Haplotype Reference Consortium panel [46] on the Michigan Imputation Server using minimac3 [47]. The filtering procedure for imputed genotypes included HWE p -value $< 10^{-4}$, genotype call rate $< 95\%$, MAF < 0.01 and imputation score < 0.30 .

GTEx eQTL summary statistics and the combination with TCGA

At the same time, for these kept cancers we obtained eQTL summary statistics of the related tissue from GTEx [36] and performed a stringent quality control (Table 1): (i) reserved SNPs with MAF > 0.05 ; (ii) excluded non-biallelic SNPs and SNPs with strand-ambiguous alleles; (iii) excluded SNPs that had no rs labels as well as duplicated ones; (iv) kept only SNPs which were included within TCGA; (v) removed SNPs whose alleles did not match those in TCGA; (vi) aligned the effect allele of SNP between TCGA and GTEx.

For comparison we implemented the following six methods in both simulations and real-life applications within the context of Cox modeling $\log[\lambda(t)/\lambda_0(t)] = \eta$: (i) the burden test: to examine $H_0: \theta = 0$ in $\eta = (\mathbf{G}\beta) \times \theta + \mathbf{Xc}$ using the Wald test in the general Cox model; (ii) the KM test: to assess $H_0: \tau = 0$ in $\eta = \mathbf{G}\mathbf{b} + \mathbf{Xc}$ and $\mathbf{b} \sim N(0, \tau)$ using the kernel-machine based approach; (iii) IEHC-Fisher: to jointly test $H_0: \tau = 0$ and $\theta = 0$ in $\eta = (\mathbf{G}\beta) \times \theta + \mathbf{G}\mathbf{b} + \mathbf{Xc}$ and $\mathbf{b} \sim N(0, \tau)$ using the Fisher's combination method, or (iv) IEHC-adapt using the adaptive combination method, or (v) IEHC-optim using the optimal combination method; (vi) IEHC-ACAT: to aggregate the first five tests using the Cauchy combination method.

Results

Independence of the two statistics in the joint test and type I error control

First, in order to validate the independence of the two statistics (denoted by U_θ and U_τ) constructed in the joint test of IEHC, we computed the Pearson's correlation coefficient between them under the null of our simulation and find little evidence supporting the dependence of the two statistics (Additional file 1: Figure S1). For instance, across the 10^5 replications, the overall correlation between U_θ and U_τ is 1.75×10^{-3} (95% confidence interval: $-4.44 \times 10^{-3} - 7.95 \times 10^{-3}$, $P = 0.580$), confirming the validity of our proposed joint test framework within which we can combine two uncorrelated statistics in a statistically straightforward fashion. Next, the Q-Q

plots demonstrate all the tests, including the burden test, the KM test, IEHC-Fisher, IEHC-adapt, IEHC-optim as well as IEHC-ACAT, effectively control the type I error (Fig. 1). Particularly, we find IEHC-ACAT correctly maintains the type I error control even if the aggregated test methods (i.e., the first five) are highly correlated (Additional file 1: Figure S2). Furthermore, IEHC-Fisher is more powerful when the fixed effect explained by eQTL and random effects beyond eQTL exist simultaneously, but is less powerful when only one of the two types of effects is true or under the null that both θ and τ are zero, where the deflated p-values are observed (Fig. 1).

Simulation results for power evaluation

We now compare the power of these tests under the alternative. To save the space, here we only present the results under three scenarios: the PVE of the gene expression level explained by β (the effect sizes of eQTLs) was equal to 0.3 or 0.5, the effect size θ (the effect size of the eQTL-based genetic score) was set to 0 or 0.4, and τ (the variance of the direct effect sizes of genetic variants) was set to 0 or 0.04. The results for other scenarios are displayed in Additional file 1: Figures S3-S9. As for

the results shown in Fig. 2, we find the burden test is in general powerful when the association signal comes only from eQTL (i.e., $\theta=0.4$ and $\tau=0$), while is underpowered when the association signal comes only from SNPs (i.e., $\theta=0$ and $\tau=0.04$). The opposite results are observed for the KM test. Compared to the burden test and the KM test, the three joint tests (i.e., IEHC-Fisher, IEHC-adapt, and IEHC-optim) are often better when the association signal is contributed by both eQTL and SNPs (i.e., $\theta=0.4$ and $\tau=0.04$).

In addition, we find the relative performance of power between the joint tests (i.e., IEHC-Fisher, IEHC-adapt, and IEHC-optim) and the burden test as well as the KM test depends on the magnitude of θ and τ . More specifically, when θ and τ are not large enough, the burden test or (and) the KM test may behave better than the joint tests even the association signal is contributed by both the two components. For instance, the KM test is more powerful compared to the joint tests when $\theta=0.1$ and $\tau=0.04$ (Additional file 1: Figures S4); whereas the burden test has a higher power when $\theta=0.2$ and $\tau=0.02$ (Additional file 1: Figures S5). Finally, IEHC-ACAT, which integrates the five tests, consistently behaves better

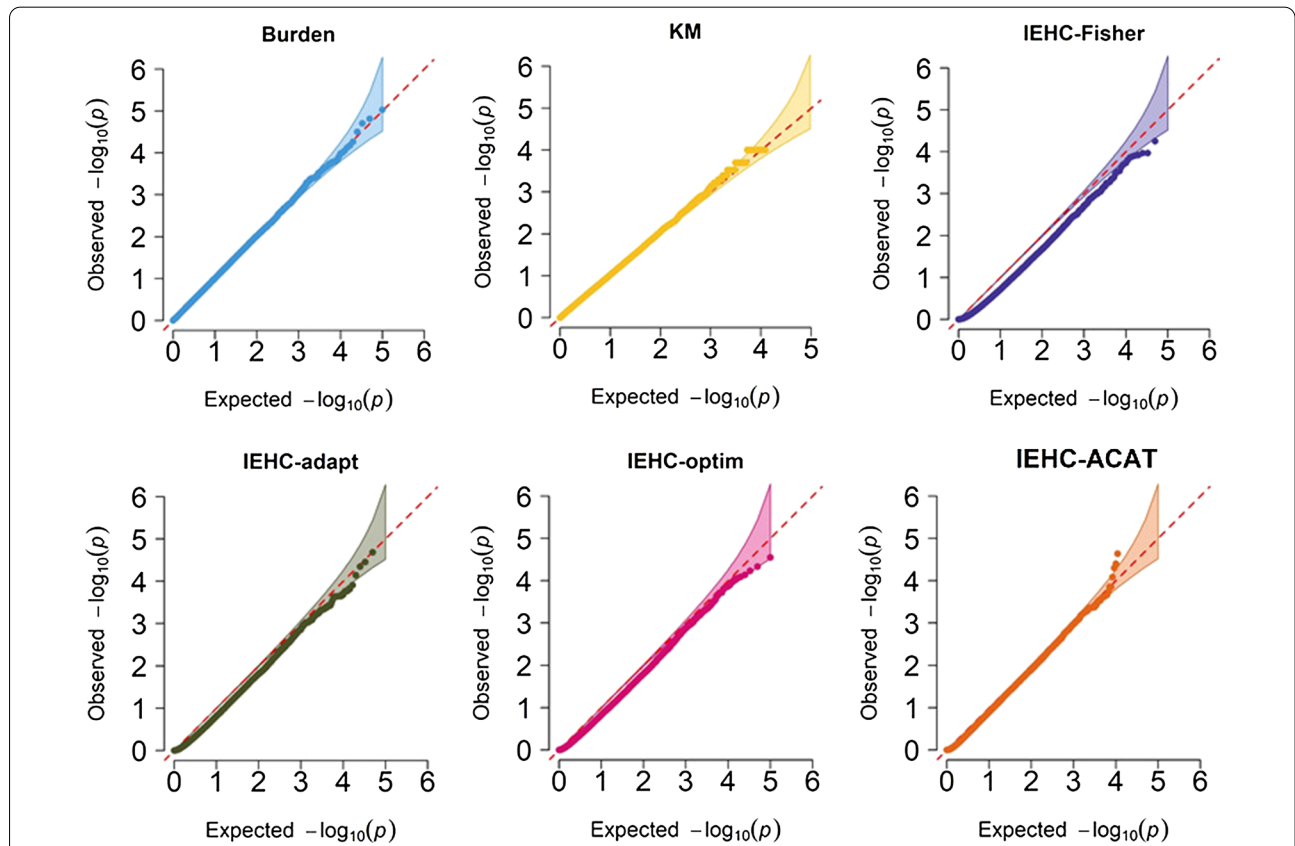
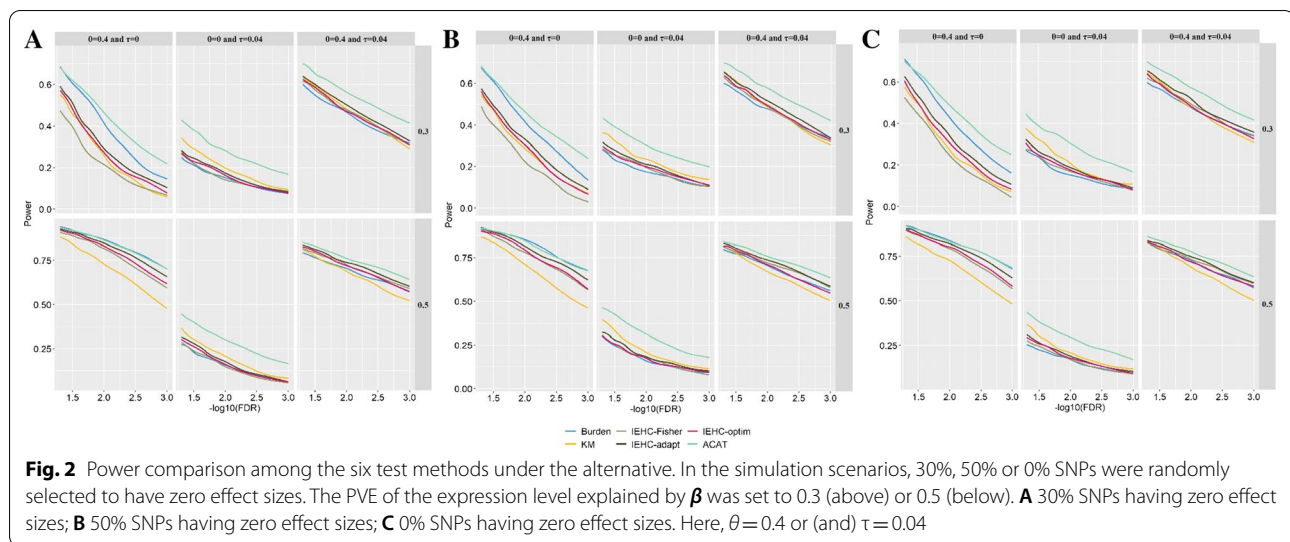


Fig. 1 The QQ plots evaluating the type I error for the burden test, the KM test, IEHC-Fisher, IEHC-adapt, IEHC-optim as well as IEHC-ACAT under the null that both θ and τ are zero



across various simulation scenarios (Fig. 2 and Additional file 1: Figures S3–S9).

Correlation between *cis*-SNP marginal effect sizes of each gene in TCGA and GTEx

In the real application, we first quantify the association of *cis*-SNP effect sizes between TCGA and GTEx. To do so, for each SNP in TCGA we generated its marginal effect size using the general Cox model while adjusting for available cancer-specific covariates (e.g., age and tumor stage), and then conducted a simple linear regression with the two sets of estimated SNP effect sizes for each gene of these cancers. Note that, the SNP effect sizes of GTEx can be directly accessed through public portal (<https://www.gtexportal.org/>). Such regression analysis renders us a rough insight to interrogate the relationship of the two types of SNP effect sizes.

We discover that these two types of SNP effect sizes are substantially correlated for a great deal of genes for each cancer (Table 2). For example, we find that on average ~72.8% (ranging from 67.6% for BRCA to 76.4% for ACC) of regression coefficients are significant (false discover rate [FDR] < 0.05). Notably, for a given cancer the regression coefficients may be positive for some genes while negative for others (Fig. 3A). Particularly, among a total of 118 genes whose regression coefficients are significant across all the ten cancers, we still find the regression coefficients are either positive or negative across diverse cancers (Fig. 3B), indicating distinct genetic influences of SNPs on the regulation of gene expression and the survival risk of cancers. More importantly, a small fraction of (~3.4% on average) determination coefficients (R^2) are larger than 10%, implying that the *cis*-SNP effect sizes of some certain genes in TCGA cancers can be

indeed explained by eQTL of relevant tissue in the GTEx (Table 2).

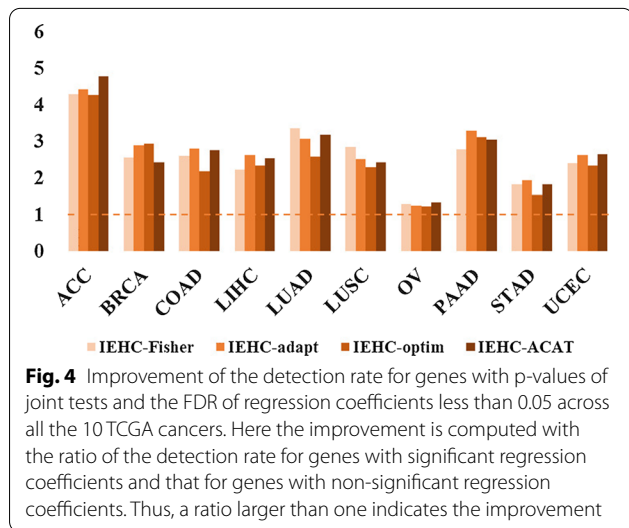
Taken together, although the average strength of the relationship between the two types of SNP effect sizes across genes may be relatively moderate, it nevertheless suggests potential genetic overlap especially at some certain genes. It is therefore worthy of integrating the eQTL of GTEx into the SNP-set based survival association analysis of TCGA cancers to boost the power.

Associated genes identified with the IEHC method

We here demonstrate that incorporating the eQTL information of GTEx into the SNP-set association analysis has the potential to enhance the power. We also exhibit that integrating all available tests by IEHC-ACAT can further increase the power. For each cancer and each type of joint tests (i.e., IEHC-Fisher, IEHC-optim, IEHC-adapt, and IEHC-ACAT), we classify all the genes into four various groups in terms of the regression coefficients (i.e., FDR < 0.05) and the results of joint tests ($P < 0.05$) (Additional file 1: Tables S1–S4 and Figures S10, S11). Taking ACC for example, there are a total of 8533 (= 259 + 8274) genes whose regression coefficients are significant (FDR < 0.05) and 2,630 (= 19 + 2611) genes whose regression coefficients are non-significant (FDR > 0.05); among these genes, 259 (~3.04% = 259/8533) and 19 (~0.72% = 19/2630) genes have a p-value less than 0.05 in terms of IEHC-Fisher, indicating that IEHC-Fisher has a fourfold higher likelihood (~4.22 = 3.04/0.72) to discover association signals ($p = 4.61 \times 10^{-11}$; Additional file 1: Table S1). The basic logic is that a smaller p-value would be generated in the joint tests if the eQTL of GTEx is predictive to the effect size of SNP in TCGA. Therefore, we expect that the detection rate of potentially associated

genes (determined by $p < 0.05$) would be greater among genes with significant regression coefficients compared to those with non-significant regression coefficients. Formally, we employ the chi-square test to examine the difference in the detection rates (e.g., 3.04% vs. 0.72%) and observe a pronounced improvement of the detection rate for the four joint tests across nearly all cancers except OV (Fig. 4), in line with our expectation and suggesting the improvement of power when integrating the eQTL information of GTEx.

Finally, the number of associated genes identified with various test approaches is summarized in Table 3. Note that, the KM test cannot identify any associations, whereas a total of 21 (9 unique) genes are discovered for four cancers after incorporating the eQTL information of GTEx (Table 4). Specifically, IEHC-ACAT and the burden test detect 5 genes, followed by IEHC-adapt (4 genes), IEHC-optim (4 genes) and IEHC-Fisher (3 genes). We find that some genes are specifically discovered by some methods (e.g., *COL9A1*, *MSANTD2*, and *LMBRD1* by the burden test), although several genes are simultaneously identified by multiple tests (e.g., *RP11-1391J7.1* by the burden test, IEHC-adapt, IEHC-optim, and IEHC-ACAT), suggesting the various power



of these test approaches across diverse genes, in line with the observation found in the simulations. Among the nine unique genes, the SNP effect sizes have a moderate correlation between TCGA and GTEx (Table 4 and Additional file 1: Figure S12).

With regards to these discovered genes, there are previous studies which provided evidence supporting their associations with the cancers. For instance, it was discovered the methylation level of *COL9A1* reduced more evidently in tumors compared to that in the blood or healthy breast tissue, suggesting the association between *COL9A1* and the risk of breast cancer [48]. Dysregulation of *EGFR* expression and signaling was previously well documented to contribute to the progression and metastasis of breast cancer while *MSANTD2* played a crucial role in decreased epidermal growth factor endocytosis [49]. It was recently shown *LMBRD1* was significantly over-expressed in BRCA1 mutated cell line compared to BRCA1 wild-type cell line [50]. As another example, *COMMD1* was under-expressed in ovarian cancer, and the lack of detectable *COMMD1* protein expression was more frequent in ovarian cancer; *COMMD1* was also shown to be related to the cisplatin sensitivity in ovarian cancer [51]. In addition, we observe that four genes (i.e., *COL9A1*, *MSANTD2*, *LMBRD1* and *RP11-1391J7.1*) were differentially expressed between normal samples and tumor samples (Additional file 1: Figure S13), and that *COL9A1* was differentially expressed among different tumor stages (Additional file 1: Figure S14). In summary, these identified genes may represent potentially promising candidate biomarkers for cancer prediction, clinical treatment, and survival prognosis evaluation.

Discussion

Recent technological advances in high-throughput platforms have greatly expanded the breadth of available omics datasets, including gene expression at the transcriptome level [36]. These abundant data resources facilitate to elucidate the interpretation of genetic variation in relation to survival risk and generate insightful perspective into the genetic underpinning of many complex

Table 3 The number of significant genes identified by different test approaches in the 10 TCGA cancers (FDR < 0.1)

Cancer	Burden	KM	IEHC-Fisher	IEHC-adapt	IEHC-optim	IEHC-ACAT
BRCA	4	0	0	1	1	1
COAD	1	0	0	0	0	0
OV	0	0	2	2	3	3
STAD	0	0	1	1	0	1
Total	5	0	3	4	4	5

We here ignore those cancers for which non associated genes are discovered by any methods

Table 4 Summary information for associated genes for the four cancers identified by different tests

Cancer	Gene	chr (pos)	S	r	Burden	KM	IEHC-Fisher	IEHC-adapt	IEHC-optim	IEHC-ACAT
BRCA	<i>COL9A1</i>	6 (70,924,764–71,012,786)	4,121	-0.32	8.94×10^{-2}	8.43×10^{-1}	1.41×10^{-1}	1.54×10^{-1}	2.18×10^{-1}	1.56×10^{-1}
BRCA	<i>MSANTD2</i>	11 (124,636,394–124,670,569)	4,563	-0.16	8.94×10^{-2}	8.43×10^{-1}	4.64×10^{-1}	2.03×10^{-1}	2.20×10^{-1}	2.33×10^{-1}
BRCA	<i>LMBRD1</i>	6 (70,385,694–70,507,003)	4,324	0.26	8.94×10^{-2}	8.43×10^{-1}	1.67×10^{-1}	1.54×10^{-1}	2.18×10^{-1}	1.56×10^{-1}
BRCA	<i>RP11-1391J7.1</i>	11 (856,880–859,795)	4,309	0.34	1.77×10^{-2}	8.43×10^{-1}	1.25×10^{-1}	3.85×10^{-2}	5.91×10^{-2}	4.65×10^{-2}
COAD	<i>CTA-407F11.6</i>	22 (26,043,228–26,045,199)	4,290	0.23	9.42×10^{-2}	8.41×10^{-1}	6.67×10^{-1}	2.09×10^{-1}	3.04×10^{-1}	2.48×10^{-1}
OV	<i>KIAA1841</i>	2 (61,293,006–61,391,960)	3,162	0.01	9.52×10^{-1}	2.62×10^{-1}	2.28×10^{-4}	4.15×10^{-5}	5.75×10^{-2}	1.75×10^{-4}
OV	<i>FAM161A</i>	2 (62,051,989–62,081,278)	3,064	0.17	8.63×10^{-1}	2.62×10^{-1}	1.70×10^{-4}	4.15×10^{-5}	1.68×10^{-3}	1.75×10^{-4}
OV	<i>COMMD1</i>	2 (62,115,859–62,374,382)	3,047	0.16	9.93×10^{-1}	2.26×10^{-1}	6.75×10^{-1}	5.38×10^{-1}	1.32×10^{-2}	4.18×10^{-2}
STAD	<i>CTC-338M12.5</i>	5 (180,618,924–180,621,429)	1,051	0.06	9.96×10^{-1}	5.88×10^{-1}	6.55×10^{-4}	9.82×10^{-5}	4.63×10^{-1}	4.27×10^{-4}

The italic values represent genes identified by different tests (FDR < 0.1 is marked as italic)

S: the number of SNPs for each gene; r: estimated correlation coefficients for SNP effect sizes between the TCGA and GTEx datasets. The FDR level is set to 0.1

human cancers [1–3]. However, how to effectively leverage the useful omics information is still an open problem. Therefore, there is a great demand for powerful analysis tools to fully harness the utility of these datasets. To fill such knowledge gap in the literature, herein we have proposed a novel genetic integrative Cox approach, called IEHC, to undertake the association analysis particularly for survival (time-to-event) phenotypes.

By characterizing effect sizes of SNPs between GTEx and TCGA, we found that there existed a substantial correlation across genes between the two types of effect sizes, indicating that we had the potential to improve the power if incorporating the GTEx eQTL into the survival SNP-set association studies. Methodologically, under the hierarchical model framework, IEHC has an appealing property that it models the effect sizes of SNPs as a function of variant characteristics (i.e., eQTL) to leverage information across loci while allowing for individual heterogeneous variant effects [26, 29]. Moreover, IEHC can be further interpreted within the framework of transcriptome-wide association studies (TWAS) [30, 31, 52]. In brief, the weighted burden score in IEHC (i.e., $G_i^T \beta$) is viewed as an imputed expression level with the weights of SNPs estimated from external tissue-related transcriptome reference datasets (i.e., GTEx), and the association between imputed expressions and cancers is examined for that gene while controlling for the direct effects of SNPs (i.e., $G_i^T \mathbf{b}$). Because TWAS is effectively viewed as performing a two-sample causal inference [53, 54]; consequently, in this sense, IEHC has the ability to identify putative causal genes for cancers under certain regularity conditions [53–55].

Compared to the permutation test which is often computationally intensive, the proposed joint tests in IEHC are much more efficient because only two independent statistics are involved, both of which can be implemented with existing software and can be further combined via three kinds of p-values combination strategies. In addition, two previously used tests, including the burden test and the KM test, can be considered as special cases of the joint test. Furthermore, in IEHC we utilized ACAT to combine all these test methods. IEHC-ACAT enjoys the attractive strength that it takes the summary of a set of p-values as the test statistic and evaluates the significance analytically without the knowledge of correlation structure [34, 35]; thus, it is extraordinarily flexible and computationally fast. As a result, IEHC-ACAT allows us to aggregate dependent p-values obtained from these tests into a single well-calibrated p-value that can achieve the maximal power while maintaining the type I error correctly [34, 35, 56].

Extensive simulations revealed the relative performance of these joint test methods in IEHC and

highlighted the strength of IEHC-ACAT. In agreement with the results of simulation, in the real application to ten TCGA cancers, we found that integrating eQTL can in general enhance the power and discover more genes that might be related to the survival risk of cancer. Particularly, IEHC-ACAT identified the highest number of associated genes among these competitive methods. In contrast, the KM test, which did not consider the eQTL of GTEx, cannot identify any association signals, suggesting the usefulness of integrating external informative variant annotations. Besides the attractive property in methodology, IEHC is also biologically interpretable when integrating transcriptomic information. First, it has been revealed that molecular features measured at the transcription level generally affect clinical outcomes more directly than those measured at other omic levels. Thus, the gene expression level would have the best predictive power for cancer prognostic evaluation compared to other genomic measurements [44]. Second, as it is widely demonstrated that SNPs associated with complex phenotypes are more likely to be eQTL [25], implying that gene expression may mediate the influence of genetic variants on the cancer risk. Therefore, eQTL can bridge the gap between cancers and many identified causal SNPs which have unknown function roles.

It needs to be emphasized that for the current IEHC model we only considered one type of variant characteristics (i.e., eQTL) but ignored other relevant information (e.g., protein quantitative trait loci). Therefore, the power of IEHC-Fisher, IEHC-adapt, IEHC-optim, and IEHC-ACAT may be further improved if more useful variant annotations would be employed in IEHC. The hierarchical modeling in IEHC offers an effective and general manner to incorporate more functional annotations as they become available. However, when many functional annotations can be applied, some of them may not be useful for determining associated genes. Therefore, the selection of informative annotations during the association analysis is necessary, which may be an interesting avenue for future investigations. Furthermore, although there include many cancer types in TCGA, their effective sample sizes are still relatively small, and the censored proportions are high [43, 44], which inevitably undermines the power of any methods and may lead to the failure of identifying some associated genes with survival. In addition, we only considered the linear kernel in these joint tests of IEHC when assessing the direct effects of SNPs (i.e., $H_0: \tau = 0$). The linear kernel may be sub-optimal if the relationship between SNPs and the survival risk is non-linear. Intuitively, the power of IEHC would depend on how well the chosen kernel captures the true relationship between SNPs and the survival risk, which

can differ in the numbers, effect sizes, and effect directions of causal variants across diverse genes. For example, if only a very small fraction of SNPs may be causal, then the sparse kernel should be a better choice; if SNPs have mutual interaction effect sizes, then the product kernel consisting of main effects and interaction terms is preferred. However, in practice the relation is rarely known in advance, selecting an optimal kernel may be very challenging [20, 57–61]. Therefore, adaptive IEHC model and test methods for multiple candidate kernel functions are warranted to study in the future [20].

Conclusion

Overall, IEHC represents a flexible, robust, and powerful approach to integrate functionally omic datasets to improve the power of identifying associated genes for the survival risk of complex human cancers.

Abbreviations

GWAS: Genome-wide association study; ACAT: Aggregated Cauchy association test; eQTL: Expression quantitative trait loci; SNP: Single nucleotide polymorphisms; GTEx: Genotype-Tissue Expression; FDR: False discover rate; TWAS: Transcriptome-wide association study; IEHC: Integrative eQTL hierarchical Cox model.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12967-021-03090-z>.

Additional file 1.

Acknowledgements

We thank TCGA and GTEx for the sharing of datasets analyzed in our work; these datasets can be available at <https://xenabrowser.net/> and <https://www.gtexportal.org/>. We are also very grateful to the editor and two referees for their insightful and constructive comments, which substantially improved our original manuscript. The data analyses in the present study were carried out with the high-performance computing cluster that was supported by the special central finance project of local universities for Xuzhou Medical University.

Authors' contributions

PZ conceived the idea for the study. PZ, SH and YW obtained the data, PZ and HL performed the data analyses. PZ, ZJ, JZ, TW, and HL interpreted the results of the data analyses. PZ and HL drafted the manuscript, and all authors approved the manuscript and provided relevant suggestions. All authors read and approved the final manuscript.

Funding

The research of Ping Zeng was supported in part by the National Natural Science Foundation of China (82173630 and 81402765), the Youth Foundation of Humanity and Social Science funded by Ministry of Education of China (18YJC910002), the Natural Science Foundation of Jiangsu Province of China (BK20181472), the China Postdoctoral Science Foundation (2018M630607 and 2019T120465), the QingLan Research Project of Jiangsu Province for Outstanding Young Teachers, the Six-Talent Peaks Project in Jiangsu Province of China (WSN-087), the Training Project for Youth Teams of Science and Technology Innovation at Xuzhou Medical University (TD202008), the Postdoctoral Science Foundation of Xuzhou Medical University, and the Statistical Science Research Project from National Bureau of Statistics of China (2014LY112). The research of Shuiping Huang was supported in part by the Social Development Project of Xuzhou City (KC19017). The research of Ting Wang was supported in part

by the Social Development Project of Xuzhou City (KC20062). The research of Yongyue Wei was supported in part by the National Natural Science Foundation of China (81973142 and 81402764).

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its additional file.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Biostatistics, School of Public Health, Xuzhou Medical University, Xuzhou 221004, Jiangsu, China. ²Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing 211166, Jiangsu, China. ³Center for Medical Statistics and Data Analysis, Xuzhou Medical University, Xuzhou 221004, Jiangsu, China. ⁴Key Laboratory of Human Genetics and Environmental Medicine, Xuzhou Medical University, Xuzhou 221004, Jiangsu, China.

Received: 9 May 2021 Accepted: 26 September 2021

Published online: 09 October 2021

References

- Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, Lemaçon A, Soucy P, Glubb D, Rostamianfar A, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*. 2017;551:92–4.
- Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, Dadaev T, Leongamornlert D, Anokian E, Cieza-Borrella C, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet*. 2018;50:928–36.
- Huang K-I, Mashl RJ, Wu Y, Ritter DJ, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell*. 2018;173:355–370.
- Baylin SB. DNA methylation and gene silencing in cancer. *Nat Clin Pract Oncol*. 2005;2:54–11.
- Robertson KD. DNA methylation and human disease. *Nat Rev Genet*. 2005;6:597–610.
- Jones PA. DNA methylation and cancer. *Oncogene*. 2002;21:5358.
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009;106:9362–7.
- Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet*. 2017;101:5–22.
- Zeng P, Zhao Y, Qian C, Zhang L, Zhang R, Gou J, Liu J, Liu L, Chen F. Statistical analysis for genome-wide association study. *J Biomed Res*. 2015;29:285–97.
- Girirajan S. Missing heritability and where to find it. *Genome Biol*. 2017;18:89.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461:747–53.
- Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*. 2010;11:446–50.
- Gusev A, Bhatia G, Zaitlen N, Vilhjalmsón BJ, Diogo D, Stahl EA, Gregersen PK, Worthington J, Klareskog L, Raychaudhuri S. Quantifying missing heritability at known GWAS loci. *PLoS Genet*. 2013;9:e1003993.

14. Young AL. Solving the missing heritability problem. *PLoS Genet*. 2019;15:e1008222.
15. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *Am J Hum Genet*. 2010;86:929–42.
16. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am J Hum Genet*. 2011;89:82–93.
17. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani David C, Wurfel Mark M, Lin X. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *Am J Hum Genet*. 2012;91:224–37.
18. Schifano ED, Epstein MP, Bielak LF, Hjun MA, Kardia SLR, Peyser PA, Lin X. SNP Set Association Analysis for Familial Data. *Genet Epidemiol*. 2012;36:797–810.
19. Wang X, Lee S, Zhu X, Redline S, Lin X. GEE-Based SNP Set Association Test for Continuous and Discrete Traits in Family-Based Association Studies. *Genet Epidemiol*. 2013;37:778–86.
20. Wu MC, Maity A, Lee S, Simmons EM, Harmon QE, Lin X, Engel SM, Moll-drem JJ, Armistead PM. Kernel Machine SNP-Set Testing Under Multiple Candidate Kernels. *Genet Epidemiol*. 2013;37:267–75.
21. Lee S, Abecasis Gonçalo R, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am J Hum Genet*. 2014;95:5–23.
22. Morgenthaler S, Thilly W. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res*. 2007;615:28–56.
23. Li B, Leal SS. Novel methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008;83:311–21.
24. Zeng P, Zhao Y, Liu J, Liu L, Zhang L, Wang T, Huang S, Chen F. Likelihood ratio tests in rare variant detection for continuous phenotypes. *Ann Hum Genet*. 2014;78:320–32.
25. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*. 2010;6:e1000888.
26. Su YR, Di C, Bien S, Huang L, Dong X, Abecasis G, Berndt S, Bezieau S, Brenner H, Caan B, et al. A Mixed-Effects Model for Powerful Association Tests in Integrative Functional Genomics. *Am J Hum Genet*. 2018;102:904–19.
27. Wu C, Pan W. Integrating eQTL data with GWAS summary statistics in pathway-based analysis with application to schizophrenia. *Genet Epidemiol*. 2018;42:303–16.
28. Xue H, Pan W, for the Alzheimer's Disease Neuroimaging I. Some statistical consideration in transcriptome-wide association studies. *Genet Epidemiol*. 2020;44:221–232.
29. Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol*. 2013;37:334–44.
30. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Elyer AE, Denny JC, Consortium GT, Nicolae DL, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47:1091–8.
31. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, Jansen R, de Geus EJC, Boomsma DI, Wright FA, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 2016;48:245–52.
32. Lin X, Cai T, Wu MC, Zhou Q, Liu G, Christiani DC, Lin X. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genet Epidemiol*. 2011;35:620–31.
33. Cai T, Tonini G, Lin X. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics*. 2011;67:975–86.
34. Liu Y, Xie J. Cauchy Combination Test: A Powerful Test With Analytic p-Value Calculation Under Arbitrary Dependency Structures. *J Am Stat Assoc*. 2020;115:393–402.
35. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet*. 2019;104:410–21.
36. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*. 2017;550:204–213.
37. Cox DR. Regression Models and Life-Tables. *J Roy Stat Soc: Ser B (Methodol)*. 1972;34:187–220.
38. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45:580–585.
39. Koziol JA, Perlman MD. Combining independent chi-squared tests. *J Am Stat Assoc*. 1978;73:753–63.
40. Fisher RA: Statistical Methods for Research Workers, 5th Edn. Biological monographs and manuals. Edinburgh: Oliver and Boyd Ltd; 1934.
41. Lappalainen T, Sammeth M, Friedländer MR, Pa TH, Monlong J, Rivas MA, Gonzálezpota M, Kurbatova N, Griebel T, Ferreira PG. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013;501:506–11.
42. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med*. 2005;24:1713–23.
43. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*. 2018;173:291–304.
44. Yu X, Wang T, Huang S, Zeng P. How can gene expression information improve prognostic prediction in TCGA cancers: an empirical comparison study on regularization and mixed-effect survival models. *Front Genetics*. 2020;11:8.
45. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*. 2013;10:5–6.
46. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48:1279–83.
47. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48:1284–7.
48. Piotrowski A, Benetkiewicz M, Menzel U, de Ståhl TD, Mantripragada K, Grigelionis G, Buckley PG, Jankowski M, Hoffman J, Bala D. Microarray-based survey of CpG islands identifies concurrent hyper- and hypomethylation patterns in tissues derived from patients with breast cancer. *Genes Chromosom Cancer*. 2006;45:656–67.
49. Runkle KB, Meyerkord CL, Desai NV, Takahashi Y, Wang H-G. Bif-1 suppresses breast cancer cell migration by promoting EGFR endocytic degradation. *Cancer Biol Ther*. 2012;13:956–66.
50. Privat M, Rudewicz J, Sonnier N, Tamisier C, Ponelle-Chachuat F, Bignon Y-J. Antioxydation and cell migration genes are identified as potential therapeutic targets in basal-like and BRCA1 mutated breast cancer cell lines. *Int J Med Sci*. 2018;15:46.
51. Fedoseienko A, Wieringa HW, Wisman GBA, Duiker E, Reyners AK, Hofker MH, van der Zee AG, van de Sluis B, van Vugt MA. Nuclear COMMD1 is associated with cisplatin sensitivity in ovarian cancer. *PLoS ONE*. 2016;11:e0165385.
52. Zeng P, Dai J, Jin S, Zhou X. Aggregating multiple expression prediction models improves the power of transcriptome-wide association studies. *Hum Mol Genet*. 2021;30:939–51.
53. Zhu H, Zhou X. Transcriptome-wide association studies: a view from Mendelian randomization. *Quant Biol*. 2020;9:78.
54. Yuan Z, Zhu H, Zeng P, Yang S, Sun S, Yang C, Liu J, Zhou X. Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nat Commun*. 2020;11:3861.
55. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K, et al. Opportunities and challenges for transcriptome-wide association studies. *Nat Genet*. 2019;51:592–9.
56. Xiao L, Yuan Z, Jin S, Wang T, Huang S, Zeng P. Multiple-tissue integrative transcriptome-wide association studies discovered new genes associated with amyotrophic lateral sclerosis. *Front Genetics*. 2020;11:587243.
57. Urrutia E, Lee S, Maity A, Zhao N, Shen J, Li Y, Wu MC. Rare variant testing across methods and thresholds using the multi-kernel sequence kernel association test (MK-SKAT). *Stat Interface*. 2015;8:495–505.
58. Wang X, Xing EP, Schaid DJ. Kernel methods for large-scale genomic data analysis. *Brief Bioinform*. 2014;16:183–92.
59. Yang H, Cao H, He T, Wang T, Cui Y. Multilevel heterogeneous omics data integration with kernel fusion. *Brief Bioinform*. 2020;21:156–70.

60. Yang H, Li S, Cao H, Zhang C, Cui Y. Predicting disease trait with genomic data: a composite kernel approach. *Brief Bioinform.* 2016;18:591–601.
61. He T, Li S, Zhong P-S, Cui Y. An optimal kernel-based U-statistic method for quantitative gene-set association analysis. *Genet Epidemiol.* 2019;43:137–49.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

