BMC Genomics

RESEARCH

Open Access

Gene-based Hardy–Weinberg equilibrium test using genotype count data: application to six types of cancers

Jo Nishino[1*], Fuyuki Miya[2] and Mamoru Kato[1]

Abstract

Background An alternative approach to investigate associations between genetic variants and disease is to examine deviations from the Hardy–Weinberg equilibrium (HWE) in genotype frequencies within a case population, instead of case–control association analysis. The HWE analysis requires disease cases and demonstrates a notable ability in mapping recessive variants. Allelic heterogeneity is a common phenomenon in diseases. While gene-based case–control association analysis successfully incorporates this heterogeneity, there are no such approaches for HWE analysis. Therefore, we proposed a gene-based HWE test (gene-HWT) by aggregating single-nucleotide polymorphism (SNP)-level HWE test statistics in a gene to address allelic heterogeneity.

Results This method used only genotype count data and publicly available linkage disequilibrium information and has a very low computational cost. Extensive simulations demonstrated that gene-HWT effectively controls the type I error at a low significance level and outperforms SNP-level HWE test in power when there are multiple causal variants within a gene. Using gene-HWT, we analyzed genotype count data from a genome-wide association study of six cancer types in Japanese individuals and suggest DGKE and ANO3 as potential germline factors in colorectal cancer. Furthermore, FSTL4 was suggested through a combined analysis across the six cancer types, with particularly notable associations observed in colorectal and prostate cancers.

Conclusions These findings indicate the potential of gene-HWT to elucidate the genetic basis of complex diseases, including cancer.

Keywords Hardy–Weinberg equilibrium test, Gene-based analysis, Cancer-related genes, Allelic heterogeneity, Recessive variants, Genome-wide association study

Background

Case–control association analyses for individual single-nucleotide polymorphisms (SNPs; i.e., single-SNP case–control analysis), such as the chi-squared or Fisher's exact test on a 2×2 contingency table or logistic regression analysis, have been used to assess the genetic association between SNPs and disease states, leading to the detection of numerous disease-related SNPs [1]. This approach has been successfully extended to "gene-based" analysis [2–8]. Gene-based analysis has several advantages over single-SNP analysis. First, collectively considering multiple variants within a gene may increase the statistical power of the analysis if allelic heterogeneity is present (i.e., different variants at the same gene lead to the same or similar phenotypes). Second, focusing on genes instead of millions of SNPs reduces the burden of multiple tests,

*Correspondence:
Jo Nishino
jnishino@ncc.go.jp
[1] Division of Bioinformatics, National Cancer Center Research Institute, Tokyo, Japan
[2] Center for Medical Genetics, Keio University School of Medicine, Tokyo, Japan

which may also increase the power. Third, gene-based analysis addresses the allelic heterogeneity and allows for more consistent findings across different studies on similar diseases. Furthermore, studying genes, the functional units of the genome, can provide valuable insights into the underlying biology of a disease.

Instead of using a case–control analysis, using deviations in genotype frequencies from the Hardy–Weinberg equilibrium (HWE) within a case population, i.e., HWE analysis, is an alternative approach to investigate the association between SNPs and disease [9–12]. For a particular locus with two alleles A and a with frequencies $(1-p)$ and p, respectively, the HWE states that the genotype frequencies of AA, Aa, aa are $(1-p)^2$, $2p(1-p)$, and p^2, respectively, under conditions such as random mating, a large population, and no migration, mutation, or selection [13]. Because different genotypes in a disease-causing variant have different levels of susceptibility to the disease, the genotype frequencies within a case population may deviate from the expectations under the HWE, i.e., in Hardy–Weinberg disequilibrium (HWD). This method has been used for fine mapping of recessive variants and for providing additional evidence for case–control analysis of recessive variants [9, 12, 14–17].

Analogous to the gene-based case–control analysis, gene-based HWE analysis should be considered owing to its many advantages, including increased statistical power and improved interpretability of results. Multiple recessive mutations commonly exist within the same disease-causing gene [18], and gene-based HWE analysis is well-suited for such scenarios. However, until now, no such method has been proposed.

Therefore, we proposed a gene-based HWE test (gene-HWT), which advantageously uses genotype count data and publicly available linkage disequilibrium (LD) information, without requiring individual genotypes, from genome-wide association studies (GWAS) with increasingly large sample sizes [19]. Note that this proposed use of HWD is not intended to identify genotype errors as is commonly done [20]. Rather, data in which mutations with a high probability of error have been removed prior to analysis is used. The results showed no features attributable to errors.

Results

gene-HWT

We started with the statistic for the single-SNP Hardy–Weinberg equilibrium test (single-SNP HWT) and then introduced the proposed method, gene-HWT (an overview is illustrated in Fig. 1). For a particular locus with two alleles A and a with frequencies $q=1-p$ and p, respectively, the statistic for single-SNP HWT, z, in n diploid samples is calculated as follows:

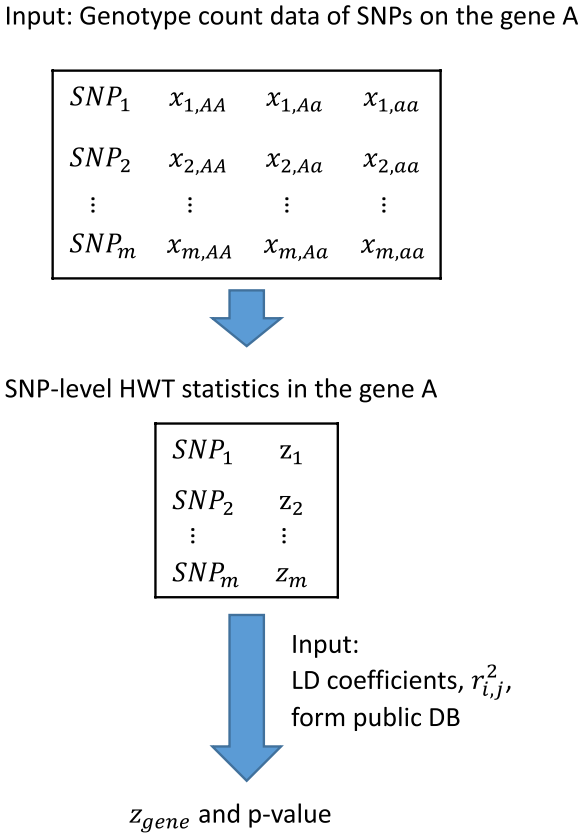

Fig. 1 Overview of the gene-HWT. The input is genotype count data of SNPs located on a specific gene. Within this gene, SNP-level HWT statistics , z_i are computed. To consider the correlations among z_i, linkage disequilibrium (LD) coefficients,$r^2_{i,j}$, are derived from a public database. The gene-HWT statistic, z_{gene}, and corresponding p-value are calculated

$$z=\frac{x_h-2n\widehat{p}\widehat{q}}{2\widehat{p}\widehat{q}\sqrt{n}}, \tag{1}$$

where x_h is observed number of heterozygosity in the sample, and $\widehat{p}$ and $\widehat{q}$ represent the sample frequency of a and A, respectively [11]. Under HWE, z is expected to be 0 since the genotype frequency of Aa is expected to be $2npq$. HWT is performed based on the fact that z asymptotically follows a standard normal distribution under HWE. Note that the commonly used statistics for single-SNP HWT is the square of z,z^2 [11].

The test using z employs continuous approximation, which does not yield appropriate results when the number of minor alleles in the sample is low. Therefore, in this study, we focused on loci with a minor allele frequency (MAF) $\geq 5\%$ in the sample. In addition, Yates' continuity correction was applied to z when the expected number of homozygotes for minor allele was ≤ 5. The correction was performed by subtracting

$0.5 \times \text{sign}(x_h - 2n\widehat{pq})$ from the numerator of z, where sign() returns the sign of a real value.

For a gene with m loci, we proposed the statistic for gene-HWT, z_{gene}, as

$$z_{gene} = \frac{\sum_{i=1}^{m} z_i}{\sqrt{V(\sum_{i=1}^{m} z_i)}} = \frac{\sum_{i=1}^{m} z_i}{\sqrt{m + 2\sum_{i=1}^{m}\sum_{j=i+1}^{m} Cov(z_i, z_j)}}, \quad (2)$$

where z_i is the HWT statistic for i-th variant in the gene. z_{gene} is the sum of z_i divided by its standard deviation (Fig. 1), enhancing the detection of cumulative accumulation of homozygote or heterozygote excesses within a gene. This statistic includes the covariance between z_i and z_j, $Cov(z_i, z_j)$, due to LD, making the direct computation of z_{gene} challenging. Considering the representation of $Cov(z_i, z_j)$ in terms of LD coefficients, $r_{i,j}$, between the i-th and j-th variants, we successfully proved $Cov(z_i, z_j) = r_{i,j}^2$ as n is large in a Supplementary Note. Therefore, z_{gene} is as follows:

$$z_{gene} = \frac{\sum_{i=1}^{m} z_i}{\sqrt{m + 2\sum_{i=1}^{m}\sum_{j=i+1}^{m} r_{i,j}^2}} \quad (3)$$

The $r_{i,j}^2$ values were retrieved from a public database. Therefore, to calculate z_{gene}, only the genotype counts were required. The gene-HWT was performed using the standard normal distribution: since z_{gene} is the standardized sum of normal variables,z_i, z_{gene} asymptotically follows a standard normal distribution under the null hypothesis that all m variants in the gene are under HWE.

p-values under the null model and type I error rates

Under the null hypothesis (HWE), the behavior of the p-value and the type I error rates of gene-HWT were investigated by simulation. In each simulation, one gene was randomly selected, with replacement, from 388 genes that meet certain criteria on chromosome 20 from the 1000 Genomes Phase 3 [21] dataset (see Methods for details). Using Hapsim [22], n diplotypes are generated while preserving the real LD structure. The QQ-plots displayed p-values obtained from 20,000 simulations for each setting, representing approximately the total number of genes in the human genome (Fig. 2). The observed $-\log_{10}(P)$ values obtained from gene-HWT (Fig. 2, circles) exhibited a good fit to the theoretical straight line under HWE for all sample sizes, n = 200, 1000, and 3000. In contrast, when LD was not corrected, i.e., when using the statistic with LD set to 0 in [2], the observed –log10(P) values (Fig. 2, cross) were substantially inflated from the expected theoretical curve, leading to the inflation of type I error rates.

Type I error rates by one million simulations for each setting are presented in Table 1. When LD was not corrected, the type I error rate was much larger than the nominal significance level. Type I error for gene-HWT tended to be conservative when the sample size was small, especially when the nominal significance level was large. For example, the type I error rate was 3.0% under n = 200 and $\alpha = 5\%$. When the sample size was large, n = 1000 or 3000, especially with a small nominal significance level, the type I error rates of gene-HWT were very close to the nominal significance level. At the nominal significance level of 0.025%, corresponding to Bonferroni-corrected 5% significance level for 20,000 genes in the human genome, the type I error rates were 0.022% and 0.026% under n = 1000 and n = 3000, respectively. Therefore, even at small significance levels, such as those used in the genome scan, the gene-HWT type I error rate can be effectively controlled by appropriately adjusting the LD.

Power

A power analysis was conducted for gene-HWT under a multiplicative relative risk model, with 1–12 causative SNPs randomly assigned within a single gene. Diplotypes for genes on chromosome 20 were simulated in the same way as examining type I error rates. The genotype risk ratios for a causal SNP was defined as AA: Aa: $aa = 1$: $(1+\beta_1)$: $(1+\beta_2)$. The individual's relative risk was obtained by multiplying the risk ratios of each variant. The absolute risk was proportional to the relative risks, under the constraint of a prevalence (average risk) of 0.1 (see Methods for details).

The powers for gene-HWT are shown in Fig. 3. The recessive ($2\beta_1 = 0, \beta_2 > 0$) and dominant ($2\beta_1 = \beta_2 > 0$) models were as follows. A larger sample size increased the power. More causal SNPs led to greater detection of power. Even a small increase from 1 to 3 causal SNPs significantly increased the power of detection. For example, when $\beta_2 = 0.2$ in a recessive case, with n = 200, 1000, and 3000, the detection rates increased from 2.8% to 13% (4.64-fold), 8.7% to 36.9% (4.24-fold), and 18.5% to 69.4% (3.75-fold), respectively. In both recessive and dominant models with sufficient sample size (n = 3000), even for relatively weak effects with β_2 = 0.05, 0.1 and 0.2, a power of 70% was achieved with 12, 6, and 3 causal SNPs, respectively. Under the same value of β_2, the power for the recessive and dominant models were equivalent but deviated in opposite directions from HWE, with the recessive model showing "Homozygote excess" ($z < 0$) and the dominant model showing "Heterozygote excess" ($z > 0$) (Fig. 4). The semidominant model ($2\beta_1 = \beta_2 > 0$) had very low power.

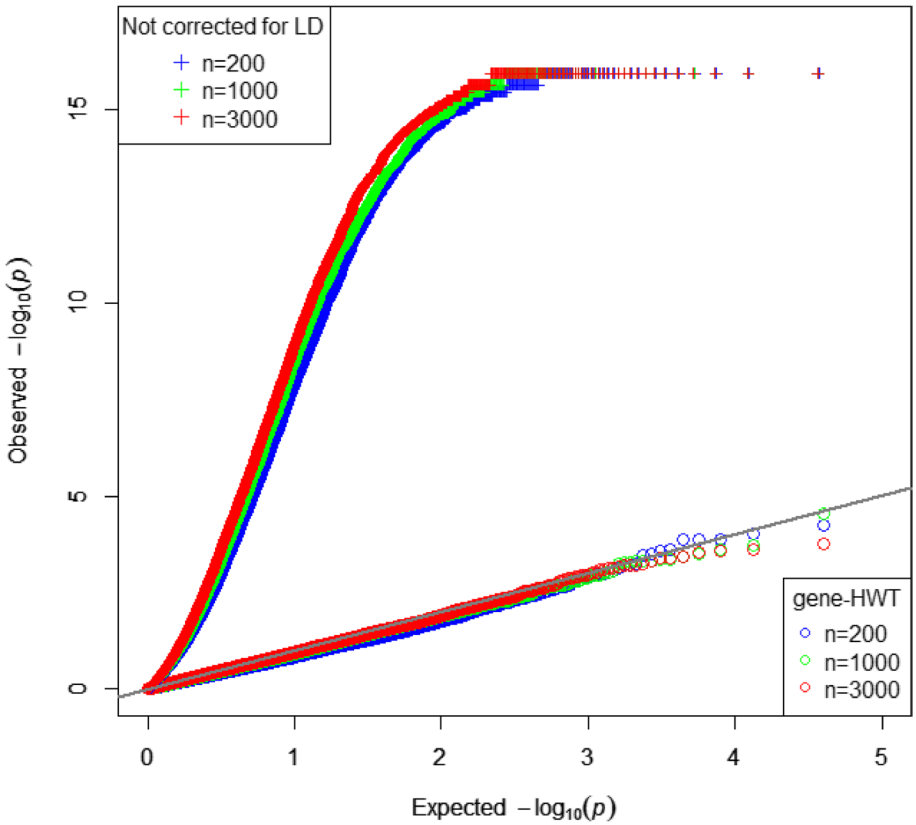

Fig. 2 QQ plot of *p*-values for gene-HWT and the test not corrected for LD under the null hypothesis. *p*-values for gene-HWT and the test not corrected for LD through simulations with sample size (n) = 200, 1000 and 3000 under the null hypothesis (HWE), utilizing real data from chromosome 20 of the EAS population in the 1000 Genomes Phase 3. The grey line represents the expected value under the null hypothesis

Table 1 Empirical type I error rates from 1,000,000 simulations

	$\alpha = 5\%$		$\alpha = 0.1\%$		$\alpha = 0.025\%$[a]		$\alpha = 0.01\%$	
	gene-HWT	**Not corrected for LD**	**gene-HWT**	**Not corrected for LD**	**gene-HWT**	**Not corrected for LD**	**gene-HWT**	**Not corrected for LD**
$n = 200$	3.0% (29,736)	56.1% (560,857)	0.058% (585)	36.335% (363,350)	**0.016% (161)**	32.124% (321,242)	0.008% (75)	29.773% (297,725)
$n = 1000$	3.9% (38,794)	59.3% (593,201)	0.082% (824)	39.685% (396,849)	**0.022% (224)**	35.347% (353,471)	0.010% (104)	32.936% (329,357)
$n = 3000$	4.5% (44,992)	61.1% (610,684)	0.097% (972)	41.529% (415,289)	**0.026% (257)**	37.209% (372,086)	0.010% (103)	34.734% (347,343)

Numbers in parentheses are numbers of rejections

LD Linkage disequilibrium, *HWT* Hardy–Weinberg equilibrium test

[a] Corresponds to significance level of 0.05 corrected for Bonferroni correction with 20,000 genes (0.05/20,000 = 0.025%)

Power comparison: gene-HWT versus single-SNP HWT

A comparison of the power of gene-HWT and single-SNP HWT, at the overall significant level of 0.05 both with multiple testing corrections, is shown in Supplementary Fig. 1. Specifically, for each parameter set, 1000 genes (= 1000 simulations) were set and in gene-HWT Bonferroni correction was applied for the testing of 1000 genes. In single-SNP test, Bonferroni correction was applied for the number of SNPs (on average, 78,537 SNPs across parameter sets) within each of the 1000 genes.

Compared with the single-SNP test, gene-HWT generally exhibits higher power (Supplementary Fig. 1). Particularly, in cases of intermediate detection power, gene-HWT exhibits a detection power approximately

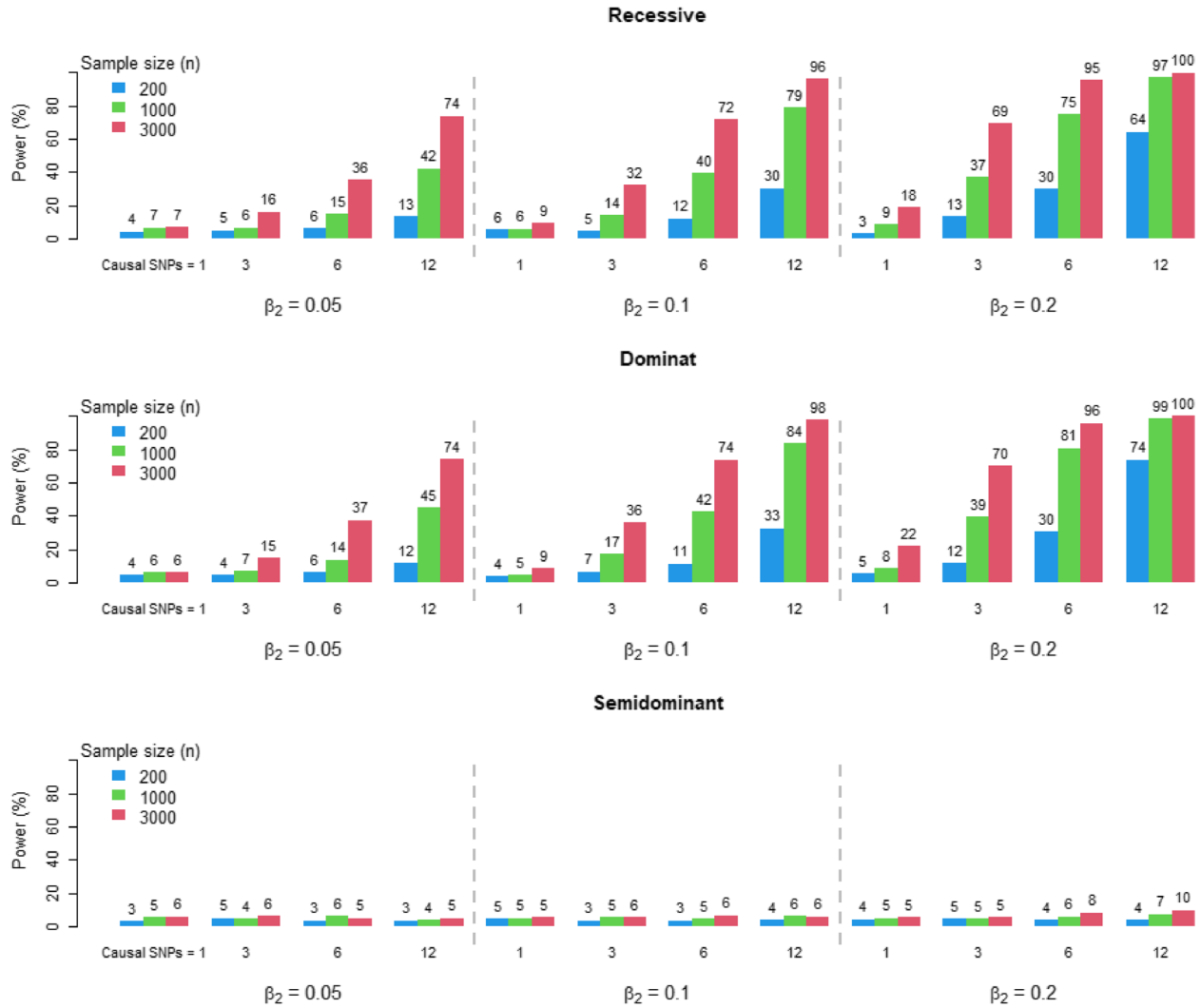

Fig. 3 Power of gene-HWT. The results shown are based on simulations with sample size (n) = 200, 1000 and 3000 using data from chromosome 20 of EAS population in the 1000 Genomes Phase 3. The genotype risk ratios for a specific causal SNP are defined as AA: Aa: aa = 1: $(1+\beta_1)$: $(1+\beta_2)$. Simulations were performed for the recessive model ($\beta_1 = 0$, $\beta_2 > 0$), dominant model ($\beta_1 = \beta_2 > 0$), and semidominant model ($2\,\beta_1 = \beta_2 > 0$)

1.2 to 1.8 times greater than that of the single-SNP testing. For example, when n = 200, causal SNP = 12, and β_2 = 0.2 in the dominant model, the power of single-SNP test was 14.4%, while that of gene-HWT was 25.6% (1.78-fold increase). In the recessive model, the power of single-SNP test was 15.9%, whereas that of gene-HWT was 22.8% (1.43-fold increase).

The power of detection was compared between single-SNP test and gene-HWT using the standard genome-wide significance levels as shown in Supplementary Fig. 2. In gene-HWT, the number of genes was set to 20,000, corresponding to Bonferroni-corrected significance level of $P < 2.5 \times 10^{-6}$. For single-SNP test, we assumed 1 million SNPs, corresponding to Bonferroni-corrected significance level of $P < 5 \times 10^{-8}$. The results aligned with the previous comparison (Supplementary Fig. 1), demonstrating that gene-HWT generally displays a higher power of detection than single-SNP test.

Analysis of genome-wide data in six cancer types

Genotype count data from GWASs for esophageal, lung, breast, gastric, colorectal, and prostate cancers in Japanese individuals were obtained from the website of the National Bioscience Database Center (NBDC) Human Database [23]. Each dataset had been quality-controlled and consisted of data from approximately 190 individuals. LD information was obtained from the 1000 Genomes Phase 3 dataset [21]. SNPs overlapping with genes (within 2 kb upstream or downstream of the

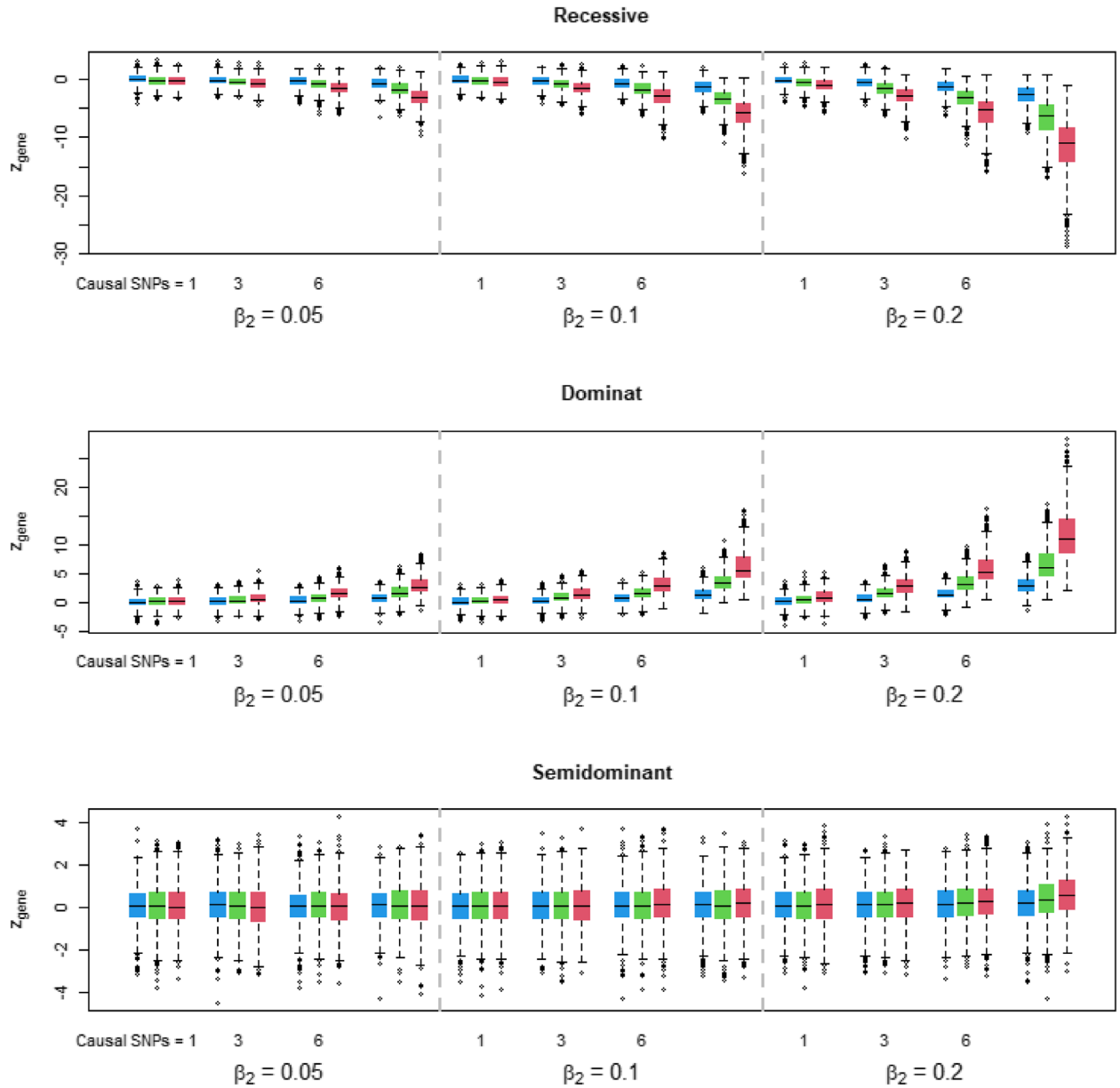

Fig. 4 z-values of gene-HWT (z_{gene}). z-values of gene-HWT obtained from the same simulations as in Fig. 3

transcripts) and those with MAF ≥ 5% were selected. To ensure the robustness of the analysis and reduce the impact of potential genotyping errors, we applied a stringent filter. Variants with single-SNP HWT p-value $< 10^{-4}$ were excluded based on datasets from NBDC used as controls, which included healthy individuals and patients with 11 non-cancerous diseases. Following the QC filters, gene-HWT was applied to analyze 11,780–13,455 genes and 92,195–173,753 SNPs across six types of cancers (see Methods for details).

Table 2 Genes identified by gene-HWT in analyses for six cancer types at q-value < 0.2

Cancer	Gene	# of SNPs	gene-HWT		
			Z	*p*-value	q-value
esophageal	BRAP	2	6.0	2.2E-9	3.0E-5
colorectal	DGKE	2	-4.9	1.1E-6	1.3E-2
colorectal	ANO3	31	-4.6	3.8E-6	2.2E-2

By applying the gene-HWT with an FDR q-value < 0.05, three genes—BRAP, DGKE, and ANO3—were identified (Table 2). BRAP, with 2 SNPs, was detected in esophageal cancer, while DGKE and ANO3, harboring 2 and 31 SNPs respectively, were identified in colorectal cancer. Among these three genes, two showed negative z-values, suggesting the potential involvement of recessive mutations. To identify common causal genes across cancers, results from six cancer studies were combined. In this analysis (see Methods for details), one gene, FSTL4, was identified with a q-value < 0.05 (Table 3), showing a negative z-value (-4.57), a p-value of 4.80×10^{-6}, and a q-value of 0.0346. Among individual cancer types, notable negative z-values were observed in colorectal (-3.71) and prostate cancers (-3.28). To further ensure robustness, SNPs analyzed in genes identified in each cancer type and the combined analysis were checked for HWD in the same NBDC control datasets described above. The HWD z-values did not indicate any noticeable bias, suggesting that genotyping errors were unlikely to affect the results (Supplementary Fig. 3).

Discussion

The proposed method, gene-HWT, is the first method to detect HWD at the gene-level by aggregating HWD in genetic variants (SNPs) within or close to the gene, while adjusting the LD among variants. This test uses only genotype count data, without the individual genotype data, and publicly available LD information. The derived simple relationship between the covariance of the HWT statistic of a pair of variants and the LD coefficient, i.e., $Cov(z_i, z_j) = r^2_{i,j}$, allows for the immediate calculation of the gene-HWT statistic from the single-SNP HWT statistics in a gene of interest without computationally intensive permutation or simulation. gene-HWT effectively controls the type I error rate at a very low significance level, enabling genome scanning, and exhibits significantly increased power compared to single-SNP tests as the number of causal variants within the gene increases. The method also enables results for each gene from different studies, i.e., studies for different cancer types as shown in this study, with allelic heterogeneity, which might lead to further increases in power.

We applied the gene-HWT to six cancer GWAS data sets, each with approximately 190 cases. In this study, BRAP was identified in esophageal cancer. BRAP has been reported to regulate the cell cycle and DNA repair [24]. GWAS studies have identified SNPs in BRAP associated with an increased risk of esophageal squamous cell carcinoma through interactions with alcohol and smoking [25]. Additionally, BRAP activates MAPK signaling in gastric cancer [26] and is linked to poor prognosis in liver cancer due to immune modulation [27]. DGKE and ANO3 were identified in colorectal cancer. DGKE has been reported to undergo promoter hypermethylation in colorectal cancer, suppressing its expression and potentially impairing its tumor suppressor function [28]. ANO3 is known to be involved in tumor progression and maintenance of the tumor microenvironment in breast cancers [29]. Notably, this study is the first to suggest a potential association between germline variants of DGKE and ANO3 and cancer. Through the combined analysis across six cancer types, FSTL4 was identified, with particularly strong associations observed in colorectal and prostate cancers. FSTL4 has been shown to regulate gene expression through interactions with DNA methylation and miRNAs, potentially contributing to the progression of cancers [30]. High expression of FSTL4 is associated with distant recurrence in breast cancer [31] and contributes to tumor microenvironment modulation in lung cancers [32].

The public data for the six cancer types used in this study had already undergone quality control by the data providers. For each dataset, the criteria included a sample call rate ≥ 0.98, SNP call rate ≥ 0.95, and SNP-level HWT p-value ≥ 1×10^{-6}. Additionally, to address the potential impact of genotyping errors, we implemented our own stringent filtering process. The control datasets consisted of 3,071 individuals, including healthy individuals and patients with 11 non-cancerous diseases, and included data genotyped on platforms partially shared with those used for the six cancer datasets. SNPs that deviated from HWE ($p < 10^{-4}$) in the control datasets were excluded from gene-HWT analysis. This very stringent threshold ensured the removal of platform-specific genotyping errors, further enhancing the reliability of the results. To further validate the findings, we examined the HWD of SNPs in the identified genes using the same NBDC control datasets. The HWD z-values showed no noticeable bias, suggesting that genotyping errors were unlikely to influence the results (Supplementary Fig. 3).

Table 3 FSTL4 gene identified in combined analysis at q-value < 0.05, with results for six cancer types

Cancer	# of SNPs	gene-HWT		
		Z	*p*-value	q-value
esophageal	101	-0.72	4.72E-01	-
lung	41	-1.57	1.16E-01	-
breast	43	-0.7	4.84E-01	-
gastric	47	-1.23	2.19E-01	-
colorectal	48	-3.71	2.07E-04	-
prostate	49	-3.28	1.04E-03	-
combined	**-**	**-4.57**	**4.80E-06**	**3.46E-02**

This is not likely due to population structure or inbreeding [33, 34]. If the effect had been large, the z-value should be negative overall, but the median SNP-level z-values for the six cancers were close to zero: −0.021, −0.026, −0.015, −0.056, and −0.029 for esophageal, lung, breast, stomach, colon, and prostate cancers, respectively. Of course, in the case of non-negligible effects of population structure or inbreeding, gene-HWT may produce erroneous results. Thus, the development of a gene-based HWE test that considers population structure and inbreeding is a future challenge.

The proposed method has certain limitations. It targets common variants with MAF ≥ 5%. As a result, many variants would be excluded from consideration. BRCA1- and BRCA2-associated hereditary breast and ovarian cancer (HBOC) follow a dominant inheritance pattern. Such dominant variants exert their effects heterozygously, making it difficult for them to be highly maintained in the population through natural selection. The reason for excluding variants with MAF < 5% is that the single-SNP HWT may not work well with rare mutations owing to breakdown of continuous approximation, and naturally, gene-HWT would also fail for these cases. Moreover, because gene-HWT improves the detection of cumulative accumulation of homozygous or heterozygous excess within a gene, it may be difficult to detect genes with both recessive and dominant mutations using gene-HWT.

Conclusions

In summary, we proposed a novel method for detection of gene-based HWD, which uses only the genotype counts and publicly available LD information. It is common for specific genes to have multiple disease-causing mutations, and our approach can aggregate their cumulative effects to enhance the detection power. We successfully demonstrated the application of this method on cancer genomic data, showing its effectiveness. Together, these findings highlight the potential utility of gene-HWT in elucidating the genetic basis of cancers and other complex diseases. The R code implementing the gene-HWT is publicly available at https://github.com/jonishino/gene-HWT.git. The analysis can be performed with genotype count data for variants, and the script supports automatic retrieval of LD information if it is not already available. The repository includes sample data, detailed usage instructions, and is distributed under the GPL v2.0.

Methods

Simulation for type I error rates of gene-HWT

Type I error rates for the proposed gene-HWT were investigated by simulations under the null hypothesis (HWE) using real data for mimicking realistic LD structure. Specifically, phased genotype data from chromosome 20 in the East Asian (EAS) population, comprising 504 individuals from the 1000 Genomes Phase 3 [21], were utilized. Only SNPs with MAF ≥ 5% were selected. To reduce the computational burden and to specify a maximum of 12 causal SNPs for subsequent power analysis, genes with 12 to 200 SNPs on chromosome 20 were selected, resulting in the use of 388 genes. In each simulation, one gene was randomly selected from 388 genes obtained, and using Hapsim [22], 2n haplotypes were generated while preserving the LD structure obtained from real data within the gene. Then, 2n haplotypes were randomly combined to create n diplotype and finally, gene-HWT was applied.

Simulation for power analysis of gene-HWT

A power analysis was conducted using simulations based on a disease causation model involving 1–12 causal SNPs in a single gene. The process of creating diplotypes was identical to that of simulation for type I error rates. The causal SNPs were randomly determined in SNPs within the genes. The genotype risk ratio for each causal mutation is defined as AA: Aa: $aa = 1$: $(1+\beta_1)$: $(1+\beta_2)$. The individual's risk ratio was determined by multiplying the risk ratios for each variant. The individual's absolute risk was determined while considering the constraint of prevalence = 0.1. In one simulation, a sufficiently large population with 'n/prevalence' diploids was created in advance, and then n individuals were selected based on each individual's absolute risk. Finally, gene-HWT was applied to the diplotypes in the patient population.

Analysis of genotype count data in six cancers

The genome-wide genotype count data for the six cancer types were obtained from the website of the National Bioscience Database Center (NBDC) Human Database (http://humandbs.biosciencedbc.jp/). Each dataset had already undergone quality control by the data providers. For each dataset, the quality control criteria included a sample call rate of ≥ 0.98, an SNP call rate of ≥ 0.95, and an original SNP-level HWT p-value of $\geq 1 \times 10^{-6}$. Each dataset consisted of data from approximately 190 individuals.

To further improve the reliability of the genetic associations and address potential genotyping errors, we applied our own additional stringent filtering criterion based on single-SNP HWT, using datasets obtained from the NBDC as controls.

1. Healthy individuals: 934 individuals genotyped using the Illumina HumanHap550v3 Genotyping BeadChip platform, consistent with the platform used for esophageal cancer analysis.

2. Patients with 11 diseases (heart failure, myocardial infarction, unstable angina, stable angina, arrhythmia, peripheral arterial disease, cerebral aneurysm, cerebral infarction, bronchial asthma, pulmonary emphysema, interstitial lung disease): 2,137 individuals in total genotyped using the Perlegen Sciences high-density oligonucleotide arrays platform, the same platform used for the analysis of cancers other than esophageal cancer.

Variants with HWT p-value $< 1 \times 10^{-4}$, calculated by combining HWT z-scores across control diseases using Stouffer's method (the approach described later for a different application), or not found in these datasets were excluded from the gene-HWT analysis.

LD information was obtained using LDmatrix function of R package LDlinkR [35] from the EAS population in the 1000 Genomes Phase 3 dataset [21]. Variants overlapping with genes (within 2 kb upstream or downstream of the transcripts), which were identified using SNPnexus [36], and those with a MAF $\geq 5\%$ were selected.

For esophageal, lung, breast, gastric, colorectal, and prostate cancers, we applied gene-HWT to 13,455 genes with 173,753 variants, 11,941 genes with 97,801 variants, 11,780 genes with 94,031 variants, 11,945 genes with 98,917 variants, 11,820 genes with 92,273 variants, and 11,815 genes with 92,195 variants, respectively. The q-value [37], an FDR-adjusted p-value, was calculated using q-value package in R.

We combined z_{gene} values from the six cancers using Stouffer's method. Specifically, the combined z-score, $z_{gene(comb)}$, was computed by summing up the individual z-scores, $z_{gene(i)}$, and dividing by the square root of the total number of studies, $k (= 6)$:

$$z_{gene(comb)} = \frac{\sum z_{gene(i)}}{\sqrt{k}}.$$

p-values (and subsequently q-values) were calculated based on the fact that, under the null hypothesis, $z_{gene(comb)}$ follows the standard normal distribution.

Statistics and bioinformatics tools

The following tools were used:

qvalue package in R: http://www.bioconductor.org/packages/release/bioc/html/qvalue.html

SNPnexus: https://www.snp-nexus.org/v4/

Hapsim package in R: https://cran.r-project.org/web/packages/hapsim/index.html

LDlinkR package in R: https://cran.r-project.org/web/packages/LDlinkR/index.html

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11321-6.

Supplementary Material 1.

Supplementary Material 2.

Authors' contributions

J.N. conceptualized and developed the methodology. J.N. performed the simulations and data analysis and wrote the manuscript. M.K. and F.M. contributed to interpretation of the results and discussions. All authors read and approved the final manuscript.

Funding

This work was supported by JSPS KAKENHI (Grant Number JP23K05871).

Data availability

Genotype counts data of six cancer types used for this research are available at the website of the NBDC Human Database / the Japan Science and Technology Agency (JST) (http://humandbs.biosciencedbc.jp/) through the following six accession numbers: hum0014.v2.jsnp.cc.v1, hum0014.v2.jsnp.pc.v1, hum0014.v2.jsnp.sc.v1, hum0014.v2.jsnp.bc.v1, hum0014.v2.jsnp.lc.v1, and hum0014.v2.jsnp.182ec.v1. Genotype counts data used as controls, including healthy individuals and patients with 11 non-cancerous diseases, are also available at the NBDC Human Database through the following accession numbers: hum0014.v2.jsnp.934ctrl.v1, hum0014.v2.jsnp.hf.v1, hum0014.v2.jsnp.mi.v1, hum0014.v2.jsnp.ua.v1, hum0014.v2.jsnp.sa.v1, hum0014.v2.jsnp.ar.v1, hum0014.v2.jsnp.aso.v1, hum0014.v2.jsnp.ca.v1, hum0014.v2.jsnp.ci.v1, hum0014.v2.jsnp.ba.v1, hum0014.v2.jsnp.pe.v1, and hum0014.v2.jsnp.ip.v1. The R code for implementing gene-HWT is available at. https://github.com/jonishino/gene-HWT.git

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 12 April 2024 Accepted: 4 February 2025
Published online: 10 February 2025

References

1. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018;27(2):e1608.
2. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A versatile gene-based test for genome-wide association studies. Am J Hum Genet. 2010;87(1):139–45.
3. Li MX, Gui HS, Kwan JS, Sham PC. GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am J Hum Genet. 2011;88(3):283–93.
4. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11(4):e1004219.
5. Bakshi A, Zhu Z, Vinkhuyzen AA, Hill WD, McRae AF, Visscher PM, et al. Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. Sci Rep. 2016;6:32894.

6. Wang M, Huang J, Liu Y, Ma L, Potash JB, Han S. COMBAT: A Combined Association Test for Genes Using Summary Statistics. Genetics. 2017;207(3):883–91.
7. Li A, Liu S, Bakshi A, Jiang L, Chen W, Zheng Z, et al. mBAT-combo: A more powerful test to detect gene-trait associations from GWAS data. Am J Hum Genet. 2023;110(1):30–43.
8. Berrandou TE, Balding D, Speed D. LDAK-GBAT: Fast and powerful gene-based association testing using summary statistics. Am J Hum Genet. 2023;110(1):23–9.
9. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, et al. A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet. 1996;13(4):399–408.
10. Nielsen DM, Ehm MG, Weir BS. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Hum Genet. 1998;63(5):1531–40.
11. Lee WC. Searching for disease-susceptibility loci by testing for Hardy-Weinberg disequilibrium in a gene bank of affected individuals. Am J Epidemiol. 2003;158(5):397–400.
12. Wittke-Thompson JK, Pluzhnikov A, Cox NJ. Rational inferences about departures from Hardy-Weinberg equilibrium. Am J Hum Genet. 2005;76(6):967–86.
13. Hartl DL, Clark AG. Principles of population genetics, 4th ed: Sinauer associates Sunderland, MA; 1997.
14. Luo X, Kranzler HR, Zuo L, Wang S, Blumberg HP, Gelernter J. CHRM2 gene predisposes to alcohol dependence, drug dependence and affective disorders: results from an extended case-control structured association study. Hum Mol Genet. 2005;14(16):2421–34.
15. Luo X, Kranzler HR, Zuo L, Lappalainen J, Yang BZ, Gelernter J. ADH4 gene variation is associated with alcohol dependence and drug dependence in European Americans: results from HWD tests and case-control association studies. Neuropsychopharmacology. 2006;31(5):1085–95.
16. Luo X, Kranzler HR, Zuo L, Wang S, Schork NJ, Gelernter J. Diplotype trend regression analysis of the ADH gene cluster and the ALDH2 gene: multiple significant associations with alcohol dependence. Am J Hum Genet. 2006;78(6):973–87.
17. Gangwar R, Ahirwar D, Mandhani A, Mittal RD. Do DNA repair genes OGG1, XRCC3 and XRCC7 have an impact on susceptibility to bladder cancer in the North Indian population? Mutat Res. 2009;680(1–2):56–63.
18. Heyne HO, Karjalainen J, Karczewski KJ, Lemmela SM, Zhou W, FinnGen, et al. Mono- and biallelic variant effects on disease at biobank scale. Nature. 2023;613(7944):519–25.
19. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: Realizing the promise. Am J Hum Genet. 2023;110(2):179–94.
20. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet. 2011;Chapter 1:Unit1 19.
21. Consortium. GP. A global reference for human genetic variation. Nature. 2015;526(7571):68-74
22. Montana G. HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. Bioinformatics. 2005;21(23):4309–11.
23. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y. JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res. 2002;30(1):158–62.
24. Volland C, Schott P, Didie M, Manner J, Unsold B, Toischer K, et al. Control of p21Cip by BRCA1-associated protein is critical for cardiomyocyte cell cycle progression and survival. Cardiovasc Res. 2020;116(3):592–604.
25. Cui R, Kamatani Y, Takahashi A, Usami M, Hosono N, Kawaguchi T, et al. Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. Gastroenterology. 2009;137(5):1768–75.
26. Wei X, Liu X, Liu H, He X, Zhuang H, Tang Y, et al. BRCA1-associated protein induced proliferation and migration of gastric cancer cells through MAPK pathway. Surg Oncol. 2020;35:191–9.
27. Ju Q, Li XM, Zhang H, Zhao YJ. BRCA1-Associated Protein Is a Potential Prognostic Biomarker and Is Correlated With Immune Infiltration in Liver Hepatocellular Carcinoma: A Pan-Cancer Analysis. Front Mol Biosci. 2020;7:573619.
28. Kai M, Yamamoto E, Sato A, Yamano HO, Niinuma T, Kitajima H, et al. Epigenetic silencing of diacylglycerol kinase gamma in colorectal cancer. Mol Carcinog. 2017;56(7):1743–52.
29. Yun JW, Yang L, Park HY, Lee CW, Cha H, Shin HT, et al. Dysregulation of cancer genes by recurrent intergenic fusions. Genome Biol. 2020;21(1):166.
30. Zhao J, Chen HQ, Yang HF, Li XY, Liu WB. Gene expression network related to DNA methylation and miRNA regulation during the process of aflatoxin B1-induced malignant transformation of L02 cells. J Appl Toxicol. 2022;42(3):475–89.
31. Mittempergher L, Saghatchian M, Wolf DM, Michiels S, Canisius S, Dessen P, et al. A gene signature for late distant metastasis in breast cancer identifies a potential mechanism of late recurrences. Mol Oncol. 2013;7(5):987–99.
32. Parfenova OK, Kukes VG, Grishin, DV. Follistatin-like proteins: structure, functions and biomedical importance. Biomedicines. 2021;9:999.
33. Meisner J, Albrechtsen A. Testing for Hardy-Weinberg equilibrium in structured populations using genotype or low-depth next generation sequencing data. Mol Ecol Resour. 2019;19(5):1144–52.
34. Kwong AM, et al. Robust, flexible, and scalable tests for Hardy–Weinberg equilibrium across diverse ancestries. Genetics. 2021;218(1).
35. Myers TA, Chanock SJ, Machiela MJ. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. Front Genet. 2020;11:157.
36. Oscanoa J, Sivapalan L, Gadaleta E, Dayem Ullah AZ, Lemoine NR, Chelala C. SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). Nucleic Acids Res. 2020;48(W1):W185–92.
37. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100(16):9440–5.

Publisher's Note