medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Gene-based Hardy–Weinberg equilibrium test using genotype count data identifies novel cancer-related genes

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

[1] Division of Bioinformatics, National Cancer Center Research Institute, Tokyo, Japan

[2] Center for Medical Genetics, Keio University School of Medicine, Tokyo, Japan

Correspondence: Jo Nishino

E-mail: jnishino@ncc.go.jp

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 distinctively requires disease cases without the need for controls 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

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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 genome-wide association study for six types of cancers in Japanese individuals and found that most of the genes detected are associated with cancers. In addition, we identified novel genes (*AGBL3* and *PSORS1C1*), novel variants in *CTSO* known to be associated with breast cancer prognosis and drug sensitivity, and novel genes as germline factors, which have associations in gene expression or methylation status with cancers in the combined analysis of six types of 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 x 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

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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, 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). Therefore, HWD can be used for genetic association, and this approach, unlike case-control studies, requires only cases and not controls. 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 increasing statistical power and the interpretability of results, similar to the gene-based case-control analysis. Multiple recessive mutations commonly exist within the same disease-causing gene, (18) and the gene-based HWE analysis is precisely targeted for such scenarios. However, until now, no

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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 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, the use of data in which mutations with a high probability of error have been removed prior to analysis is intended. The results obtained did not show any 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:

$$z = \frac{x_h - 2n\hat{p}\hat{q}}{2\hat{p}\hat{q}\sqrt{n}}, (1)$$

where x_h is observed number of heterozygosity in the sample, and $\hat{p}$ and $\hat{q}$ 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

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

study, we focused on loci with a minor allele frequency (MAF) ≥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 $0.5 \times \mathrm{sign}(x_h - 2n\hat{p}\hat{q})$ 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

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

where z_i is the HWT formula 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:

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

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

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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 α = 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.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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+β_1) : (1+β_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 β_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.

Power comparison: gene-HWT versus single-SNP HWT

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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 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

The genotype count data from GWASs for esophageal, lung, breast, gastric, colorectal, and prostate cancers in Japanese individuals were obtained from the website of the

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

National Bioscience Database Center (NBDC) Human Database (23). Each dataset had been quality-controlled and consisted of approximately 190 individuals. LD information was obtained from the 1000 Genomes Phase 3 dataset (21). The SNPs overlapping with genes (within 2 kb upstream or downstream of the transcripts) , and those with MAF ≥5% were selected. For the six types of cancers, gene-HWT was applied to analyze 11,813 to 13,482 genes and 92,690 to 174,270 SNPs (see Methods for details).

Eighty cancer type-gene pairs were identified by applying the gene-HWT with the criterion: FDR q-value < 0.2 for each cancer type (Supplementary Table 1). The 80 genes encompassed SNPs ranging from 1 to 103. To identify common causal genes in cancers, the results of six cancer studies were combined. In the combined analysis (see Methods for details), 11 genes were identified with q-values < 0.05 (Table 2). The combined z_{gene} values are all negative, suggesting that these 11 genes may have recessive mutations. Among these 11 genes, 8 genes (*CCDC32, POFUT2, PPP1CB, QRFPR, FSTL4, ACRV1, CTSO,* and *GPR180*) have been reported to be associated with gene expression, methylation, or germline mutations, and 3 genes, including *AGBL3* and *PSORS1C1*, were newly identified as candidate cancer-related genes. Additionally, among the genes detected using the threshold of FDR q-value < 0.2 for each cancer type (Supplementary Table 1), 8 genes were found in multiple cancers (Table 3). For genes except for *OR4N2*, z_{gene} values were negative, suggesting the presence of recessive mutations. *HLA-DMA* was identified in four cancer types, while *ZNF736* and *AL163636.2* were identified in three cancer types, and the remaining genes were observed in two cancer types.

DISCUSSION

The proposed method, gene-HWT, is the first method to detect HWD at the

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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., $Cov(z_i, z_j) = r_{i,j}^2$, 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 allows for combining 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 genome data sets, each with approximately 190 cases. By combining these results, we identified 11 significant genes at a threshold of q-value = 0.05. Of the 11 significant genes, 8 genes (*CCDC32, POFUT2, PPP1CB, QRFPR, FSTL4, ACRV1, CTSO,* and *GPR180*) were reported to be associated with various cancers (24-32). Furthermore, of those eight genes, cancer types reported in existing studies often showed the highest significance in our analysis. These findings suggest that gene-HWT captures true causal genes.

CTSO exhibited the most significant association with breast cancer (z_{gene} = –4.44) in a recessive manner in the combined analysis. This aligns with previous reports indicating that the homozygous G variant of *CTSO* rs10030044 is associated with a worse prognosis (i.e., in recessive form) in patients with hormone receptor-positive breast cancer receiving tamoxifen therapy (30). The SNP rs10030044 and the linked rs4256192 lead to

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

increased *CTSO* expression, resulting in decreased *BRCA1* expression and, consequently, tamoxifen resistance (33). Our results showed that *CTSO* could be one of the causative genes for not only breast cancer but also other types of cancer, particularly esophageal cancer, as indicated by its large negative z-score (z_{gene} = –2.2). For breast cancer, the strong effect was observed with the SNP rs10019161 (single-SNP HWT $P = 2.18 \times 10^{-9}$), which was also significant in lung cancer (P = 0.0054). In esophageal cancer, the strong effects were observed in rs10019975 (P = 0.0174) and rs7684248 (P = 0.0563). According to the LDmatrix (34) using data in the 1KG Project EAS population (21), these three SNPs rs10019161, rs10019975 and rs7684248 are not in LD with the previously reported rs10030044 and rs4256192 ($r^2 < 0.01$). This suggests that these three *SNPs* (which are in LD with each other with $r^2 > 0.4$) could potentially act as independent factors in the cancer of the two previously reported SNPs.

In the analysis of each of the six cancer types in Japanese individuals, the positive z-score for *TP53* detected only in colorectal cancer ($P = 1.45 \times 10^{-4}$), suggesting dominant variants, could be due to "dominant negative" features for *TP53* variants (35). Furthermore, *RAD51* (36), *APOBEC3A* (37), and *BRAP* (38) identified in gastric, lung, and esophageal cancers, respectively are well-known for their association with cancer. In the analysis of each of the six cancer types with a threshold of q-value = 0.2, out of the 8 detected genes, 6 genes (*HLA-DMA, NUP54, OR4N2, QRFPR, TBK1* and *ZNF736*) are associated with cancers (27, 39-47). *HLA-DMA* showed significant associations with four types of cancer in this study. HLA-DM (encoded by *HLA-DMA* and *-DMB*) is a non-classical MHC class II-like protein that acts as a peptide editor in the antigen presentation process and plays a crucial role in regulating the loading of antigenic peptides onto MHC class II molecules. In patients with chronic hepatitis C virus type 1 infection in

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

the Chinese Han population who underwent interferon/ribavirin therapy, individuals carrying the rs1063478 TT genotype of *HLA-DMA* had a higher likelihood of achieving sustained virological response (SVR) (39). In patients with lung adenocarcinoma, low *HLA-DMA* expression was associated with disease-specific survival and overall survival (OS) (40). In glioma, high *HLA-DMA* expression was associated with poor prognosis (41). These findings and the result of our analysis suggest that *HLA-DMA* may be associated with a variety of cancers.

New candidate cancer-related genes were identified in our analysis, such as *AGBL3*, *PSORS1C1,* and *ZNF680*. ATP/GTP Binding Protein-Like 3 (AGBL3), is a metallocarboxypeptidase that mediates deglutamylation of both tubulin and non-tubulin target proteins. Psoriasis Susceptibility 1 Candidate 1 (*PSORS1C1)* is one of the genes associated with psoriasis (48), an inflammatory skin disease, and since there is a known association between psoriasis and cancer (49), PSORS1C1 may also be a cancer-causing gene. In addition, many previous reports have focused on mRNA expression and methylation studies, and this is the first study to find an association between SNPs and cancer for genes such as *OFUT2, PPP1CB, QRFPR, FSTL4, ACRV1, GPR180, OR4N2, QRFPR,* and *ZNF736*.

The public data of the six cancer types used in this study had been quality controlled, and thus, HWD detected by gene-HWT was not attributable to genotyping errors. If the effect of error had been dominant in the detected genes, a tendency towards excess heterozygosity would have been observed (20); however, many detected genes showed excess homozygosity. This is not likely due to population structure or inbreeding (50, 51). 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

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

–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. Indeed, all genes detected, except for *ORN2*, exhibited negative z_{gene} values, indicating the presence of recessive mutations in those genes. 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, since gene-HWT enhances the detection of cumulative accumulation of homozygous or heterozygous excess within a gene, it might 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

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

cancers and other complex diseases.

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

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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 been quality controlled with sample call rate ≥ 0.98, SNP call rate ≥ 0.95 and p-value of original SNP-level HWT ≥ 1 x 10^{-6}, and consisted of approximately 190 individuals. LD information was obtained using LDmatrix function of R package LDlinkR (34) from EAS population in the 1000 Genomes Phase 3 dataset (21). The variants overlapping with genes (within 2 kb upstream or downstream of the transcripts), which were determined using SNPnexus (52), and those with MAF ≥5% were selected. For esophageal, lung, breast, gastric, colorectal, and prostate cancers, we applied gene-HWT to 13,482 genes with 174,270 variants, 12,051 genes with 100,496 variants, 11,813 genes with 94,412 variants, 11,992 genes with 99,748 variants, 11,855 genes with 92,809 variants, and 11,857 genes with 92,690 variants, respectively. q-value (53), an FDR-adjusted p-value, was calculated using the 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.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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

DECLARATIONS

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Genotypes from 1KG are available at

http://ftp.ensembl.org/pub/data_files/homo_sapiens/GRCh38/variation_genotype/

ALL.chr20_GRCh38.genotypes.20170504.vcf.gz.

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.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

The R code for implementing gene-HWT is available at

https://github.com/jonishino/gene-HWT.git

Competing interests

The authors declare that they have no competing interests.

Funding

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

Authors' contributions

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

Acknowledgement

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.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

departures from Hardy-Weinberg equilibrium. Am J Hum Genet. 2005;76(6):967-86.

13. Hartl DL, Clark AG, Clark AG. Principles of population genetics: 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

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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. Kafaie S, Xu L, Hu T. Statistical methods with exhaustive search in the identification of gene-gene interactions for colorectal cancer. Genet Epidemiol. 2021;45(2):222-34.

25. Qiu P, Chen X, Xiao C, Zhang M, Wang H, Wang C, et al. Emerging glyco-risk prediction model to forecast response to immune checkpoint inhibitors in colorectal cancer. J Cancer Res Clin Oncol. 2023;149(9):6411-34.

26. Hu L, Xu H, Wang X, Wu B, Chen F, Chen W, et al. The expression and clinical prognostic value of protein phosphatase 1 catalytic subunit beta in pancreatic cancer. Bioengineered. 2021;12(1):2763-78.

27. Kawan MA, Kyrou I, Ramanjaneya M, Williams K, Jeyaneethi J, Randeva HS, et al. Involvement of the glutamine RF□amide peptide and its cognate receptor GPR103 in prostate cancer. Oncol Rep. 2019;41(2):1140-50.

28. Liu Y, Zhu D, Xing H, Hou Y, Sun Y. A 6□gene risk score system constructed for predicting the clinical prognosis of pancreatic adenocarcinoma patients. Oncol Rep. 2019;41(3):1521-30.

29. Wang X, Kaczor-Urbanowicz KE, Wong DT. Salivary biomarkers in cancer

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

detection. Med Oncol. 2017;34(1):7.

30. Hato Y, Kondo N, Yoshimoto N, Endo Y, Asano T, Dong Y, et al. Prognostic impact of a single-nucleotide polymorphism near the CTSO gene in hormone receptor-positive breast cancer patients. Int J Clin Oncol. 2016;21(3):539-47.

31. Balazova L, Balaz M, Horvath C, Horvath A, Moser C, Kovanicova Z, et al. GPR180 is a component of TGFbeta signalling that promotes thermogenic adipocyte function and mediates the metabolic effects of the adipocyte-secreted factor CTHRC1. Nat Commun. 2021;12(1):7144.

32. Honda S, Minato M, Suzuki H, Fujiyoshi M, Miyagi H, Haruta M, et al. Clinical prognostic value of DNA methylation in hepatoblastoma: Four novel tumor suppressor candidates. Cancer Sci. 2016;107(6):812-9.

33. Cairns J, Ingle JN, Wickerham LD, Weinshilboum R, Liu M, Wang L. SNPs near the cysteine proteinase cathepsin O gene (CTSO) determine tamoxifen sensitivity in ERalpha-positive breast cancer through regulation of BRCA1. PLoS Genet. 2017;13(10):e1007031.

34. Myers TA, Chanock SJ, Machiela MJ. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. Front Genet. 2020;11:157.

35. Willis A, Jung EJ, Wakefield T, Chen X. Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. Oncogene. 2004;23(13):2330-8.

36. Grundy MK, Buckanovich RJ, Bernstein KA. Regulation and pharmacological targeting of RAD51 in cancer. NAR Cancer. 2020;2(3):zcaa024.

37. Middlebrooks CD, Banday AR, Matsuda K, Udquim KI, Onabajo OO, Paquin A, et

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

al. Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. Nat Genet. 2016;48(11):1330-8.

38. Li S, Ku CY, Farmer AA, Cong YS, Chen CF, Lee WH. Identification of a novel cytoplasmic protein that specifically binds to nuclear localization signal motifs. J Biol Chem. 1998;273(11):6183-9.

39. Chen H, Yao Y, Wang Y, Zhou H, Xu T, Liu J, et al. Polymorphisms of HLA-DM on Treatment Response to Interferon/Ribavirin in Patients with Chronic Hepatitis C Virus Type 1 Infection. Int J Environ Res Public Health. 2016;13(10).

40. Yu QIN, Chen C, Zhang H, Chen JIN, Shen J, Yan JUN. Prognosis and immunological role of HLA-DMA in lung adenocarcinoma. Biocell. 2023;47(6):1279-92.

41. Yin X, Wu Q, Hao Z, Chen L. Identification of novel prognostic targets in glioblastoma using bioinformatics analysis. Biomed Eng Online. 2022;21(1):26.

42. Rodriguez-Berriguete G, Granata G, Puliyadi R, Tiwana G, Prevo R, Wilson RS, et al. Nucleoporin 54 contributes to homologous recombination repair and post-replicative DNA integrity. Nucleic Acids Res. 2018;46(15):7731-46.

43. Wang F, Zhang J, Tang H, Pang Y, Ke X, Peng W, et al. Nup54-induced CARM1 nuclear importation promotes gastric cancer cell proliferation and tumorigenesis through transcriptional activation and methylation of Notch2. Oncogene. 2022;41(2):246-59.

44. Wang H, Shen L, Li Y, Lv J. Integrated characterisation of cancer genes identifies key molecular biomarkers in stomach adenocarcinoma. J Clin Pathol. 2020;73(9):579-86.

45. Cho HJ, Koo J. Odorant G protein-coupled receptors as potential therapeutic targets for adult diffuse gliomas: a systematic analysis and review. BMB Rep. 2021;54(12):601-7.

46. Runde AP, Mack R, S JP, Zhang J. The role of TBK1 in cancer pathogenesis and

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

anticancer immunity. J Exp Clin Cancer Res. 2022;41(1):135.

47. Herreros-Pomares A, Llorens C, Soriano B, Bagan L, Moreno A, Calabuig-Farinas S, et al. Differentially methylated genes in proliferative verrucous leukoplakia reveal potential malignant biomarkers for oral squamous cell carcinoma. Oral Oncol. 2021;116:105191.

48. Holm SJ, Carlen LM, Mallbris L, Stahle-Backdahl M, O'Brien KP. Polymorphisms in the SEEK1 and SPR1 genes on 6p21.3 associate with psoriasis in the Swedish population. Exp Dermatol. 2003;12(4):435-44.

49. Peng H, Wu X, Wen Y, Li C, Lin J, Li J, et al. Association between systemic sclerosis and risk of lung cancer: results from a pool of cohort studies and Mendelian randomization analysis. Autoimmun Rev. 2020;19(10):102633.

50. 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.

51. Kwong AM, Blackwell TW, LeFaive J, de Andrade M, Barnard J, Barnes KC, et al. Robust, flexible, and scalable tests for Hardy-Weinberg equilibrium across diverse ancestries. Genetics. 2021;218(1).

52. 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-W92.

53. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100(16):9440-5.

FIGURE LEGENDS

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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 are computed. To consider the correlations among $z'_i s$, linkage disequilibrium (LD) coefficients, $r_{i,j}^2$, are derived from a public database. The gene-based HWT statistics, z_{gene}, and corresponding p-value are calculated.

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.

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+β_1) : (1+β_2). Simulations were performed for the recessive model (β_1=0, β_2>0), dominant model (β_1=β_2>0), and semidominant model (2β_1=β_2>0).

Fig. 4. z-values of gene-HWT (z_{gene}).

z-values of GHWT obtained from the same simulations as in Figure 2.

TABLES

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

	α = 5%		*α* = 0.1%		***α* = 0.025%***		α = 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.

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

LD, linkage disequilibrium; HWT, Hardy–Weinberg equilibrium test.

Table 2 Genes identified by combined gene-HWT results for six cancer types at *q*-value < 0.05.

gene-HWT z-value, z_{gene} (# of SNPs)	Combined results	Relevant literature

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Gene	Esophageal	Lung	Breast	Gastric	Colorectal	Prostate	Z_{gene} (combined)	*p*-value (x10^7)	*q*-value (x10^4)	highlights
CCDC32	-1.4 (1)	-2.5 (1)	-2.8 (1)	-2.4 (1)	-***3.4*** (1)	-3.0 (1)	-6.4	0.002	0.02	SNP-SNP interactions within *CCDC32* are observed in ***colorectal*** cancer GWAS (24).
AC007998.2	-1.1 (7)	-1.3 (2)	-2.6 (2)	-3.5 (2)	-2.1 (2)	-4.2 (2)	-6.0	0.020	0.10	None
POFUT2	-0.5 (3)	-0.9 (2)	-3.5 (1)	-1.5 (2)	-***4.6*** (1)	-3.2 (1)	-5.8	0.075	0.26	Reduced expression of *POFUT2* is associated with poor prognosis in ***colorectal*** cancer (25).
PPP1CB	0.9 (5)	-1.7 (2)	-3.1 (2)	-4.4 (3)	-1.1 (2)	-3.0 (2)	-5.1	4.0	10.3	Expression of *PPP1CB* is associated with poor prognosis in pancreatic cancer (26).
QRFPR	0.0 (9)	-4.2 (9)	-0.5 (9)	-1.8 (9)	-0.9 (11)	-***3.8*** (11)	-4.6	43.2	89.0	Significant upregulation of *QRFP* expression is observed in ***prostate cancer*** tissue samples (27).
FSTL4	-0.7 (101)	-1.4 (42)	-0.7 (43)	-1.2 (47)	-3.7 (48)	-3.3 (49)	-4.5	64.5	110.7	A 6-gene score using expression, featuring *FSTL4*, predicts pancreatic adenocarcinoma prognosis (28).
ACRV1	0.6 (1)	-2.1 (3)	-2.7 (3)	-1.3 (3)	-2.2 (2)	-3.2 (2)	-4.4	110.7	162.7	Four messenger RNAs, including

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

										ACRV1, distinguish pancreatic cancer patients from non-cancer subjects (chronic pancreatitis patients and healthy controls) (29).
AGBL3	-0.3 (11)	-1.7 (3)	-3.2 (1)	-2.2 (2)	-1.8 (1)	-1.5 (2)	-4.3	166.2	211.8	None
CTSO	-2.2 (7)	-1.3 (3)	-***4.4*** (2)	-1.4 (2)	-1.1 (2)	-0.0 (2)	-4.3	185.2	211.8	The GG homozygote of CTSO rs10030044 is associated with poor prognosis in hormone receptor-positive ***breast cancer*** patients receiving tamoxifen therapy (30).
PSORS1C1	-1.5 (16)	-3.4 (4)	-1.4 (2)	-1.9 (3)	-1.2 (3)	-0.8 (3)	-4.1	358.5	368.9	None
GPR180	-0.5 (4)	-0.7 (4)	-1.8 (4)	-1.3 (4)	-2.5 (3)	-3.2 (2)	-4.1	486.1	454.8	*GPR180*, a component of TGFβ signaling (31), is likely a tumor suppressor gene; increased methylation is associated with poor prognosis (32).

If cancer types reported in existing studies showed the highest significance in our analysis, those cancer types and z_{gene} values have been highlighted in bold.

HWT, Hardy–Weinberg equilibrium test; SNP, single-nucleotide polymorphism.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3 Genes detected in multiple cancer types with FDR q-value < 0.2 criteria.

Gene	Cancer	# of SNPs	gene-HWT			Relevant literature highlights
			Z_{gene}	*p*-value (10^5)	*q*-value (10^2)	
AL163636.2	lung	1	-4.18	2.89	4.36	None
	colorectal	1	-4.33	1.47	2.90	
	prostate	1	-4.21	2.56	9.31	
HLA-DMA	lung	1	-4.65	0.33	2.40	In patients with chronic hepatitis C virus type 1 infection, the rs1063478 TT genotype of *HLA-DMA* is more likely to achieve SVR with interferon/ribavirin therapy (39). Expression of *HLA-DMA* is associated with the prognosis of both lung adenocarcinoma (40) and glioma patients (41).
	breast	1	-4.18	2.91	5.50	
	colorectal	1	-3.79	15.00	11.10	
	prostate	1	-4.10	4.18	9.91	
NUP54	***gastric***	1	***-5.14***	0.03	0.33	Nup54 contributes to homologous recombination repair and DNA integrity (42), while promoting ***gastric cancer*** cell growth and tumorigenesis through CARM1 nuclear importation (43).
	prostate	1	-4.62	0.38	4.50	
OR4N2	lung	5	4.53	0.60	2.40	High expression of *OR4N2* is associated with increased mortality in stomach adenocarcinoma (44). *OR4N2* is also highly expressed in a subtype of glioma, and the expression may be related to glioma prognosis (45).
	breast	2	4.82	0.14	1.71	
QRFPR	lung	9	-4.24	2.21	4.36	Significant upregulation of *QRFP* expression is observed in ***prostate cancer*** tissue samples (27).
	prostate	11	***-3.79***	15.04	17.83	
TBK1	gastric	1	-3.68	23.66	18.91	*TBK1* regulates the proliferation and survival of malignant cells in many types of cancer and controls antitumor immunity and inflammation by regulating cytokine production in dendritic cells and macrophages, making it a potential molecular anticancer target (46).
	prostate	1	-4.03	5.70	11.26	
ZNF680	lung	2	-3.65	25.77	17.24	None
	breast	1	-4.28	1.85	4.69	
	prostate	1	-4.40	1.07	6.36	
ZNF736	lung	1	-3.71	20.89	15.87	Downregulated methylation of *ZNF736* is observed in the PVL group,

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

breast	1	-4.65	0.33	1.92	compared to that in the control group (47).

If cancer types reported in existing studies showed the highest significance in our analysis, those cancer types and z_{gene} values have been highlighted in bold.

SNP, single-nucleotide polymorphism.

medRxiv preprint doi: https://doi.org/10.1101/2024.03.20.24304529; this version posted March 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuit
It is made available under a CC-BY-NC-ND 4.0 International license .

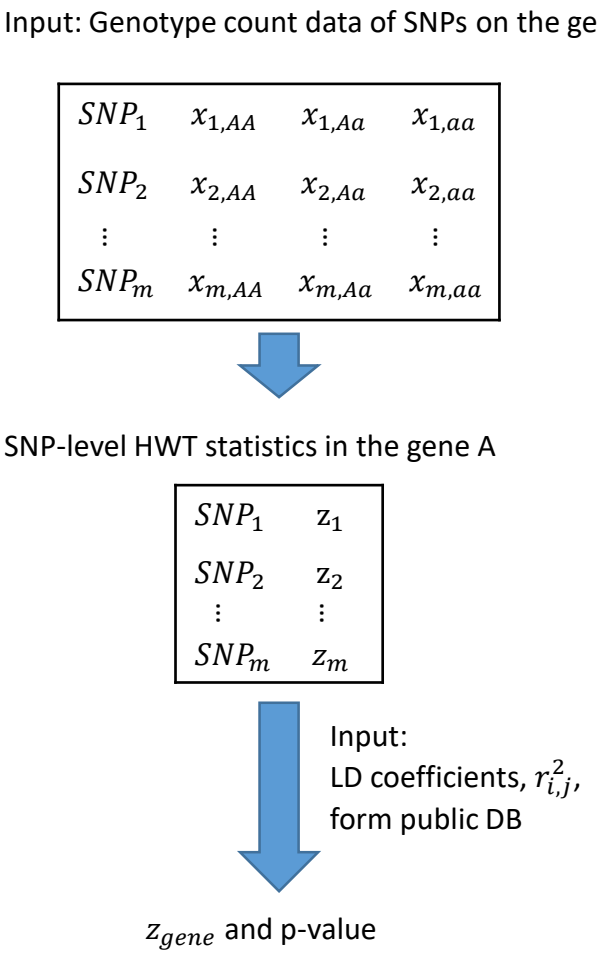

Figure 1. Overview of the gene-HWT.
The input is genotype count data of SNPs located on a specific gene. Within this gene, SNP-level HWT statistics are computed. To consider the correlations among $z'_i s$, linkage disequilibrium (LD) coefficients, $r_{i,j}^2$, are derived from a public database. The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

(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuit
It is made available under a CC-BY-NC-ND 4.0 International license .

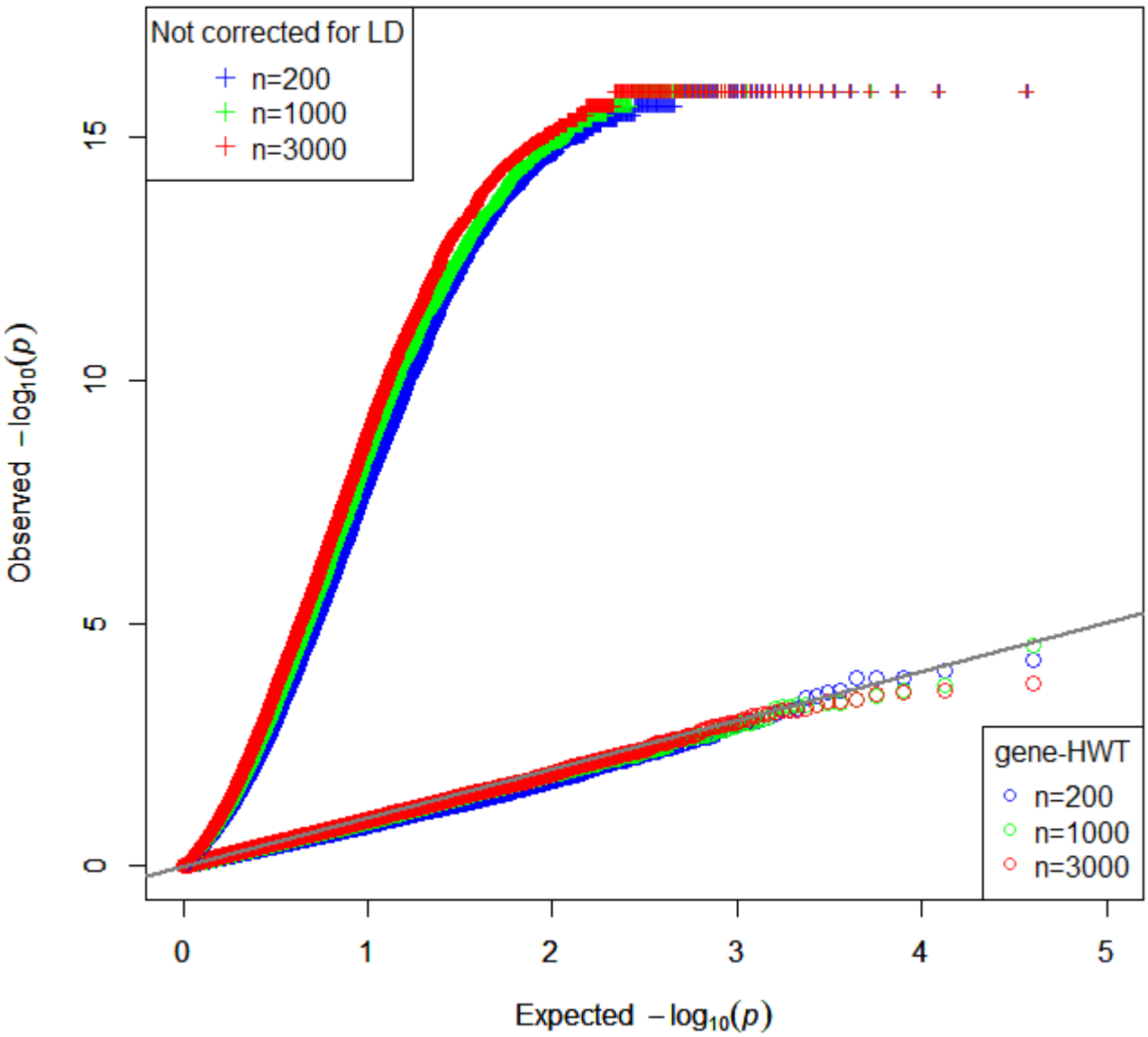

Figure 2. QQ plot of P-values for gene-HWT and the test not corrected for LD under the null hypothesis. P values for gene-HWT and the test not corrected for LD through simulations with sample size (n)=200, 1000 and 3000 under the null hypothesis (HWE), utilizing real data from chromosome 20 of the EAS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

(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuit
It is made available under a CC-BY-NC-ND 4.0 International license .

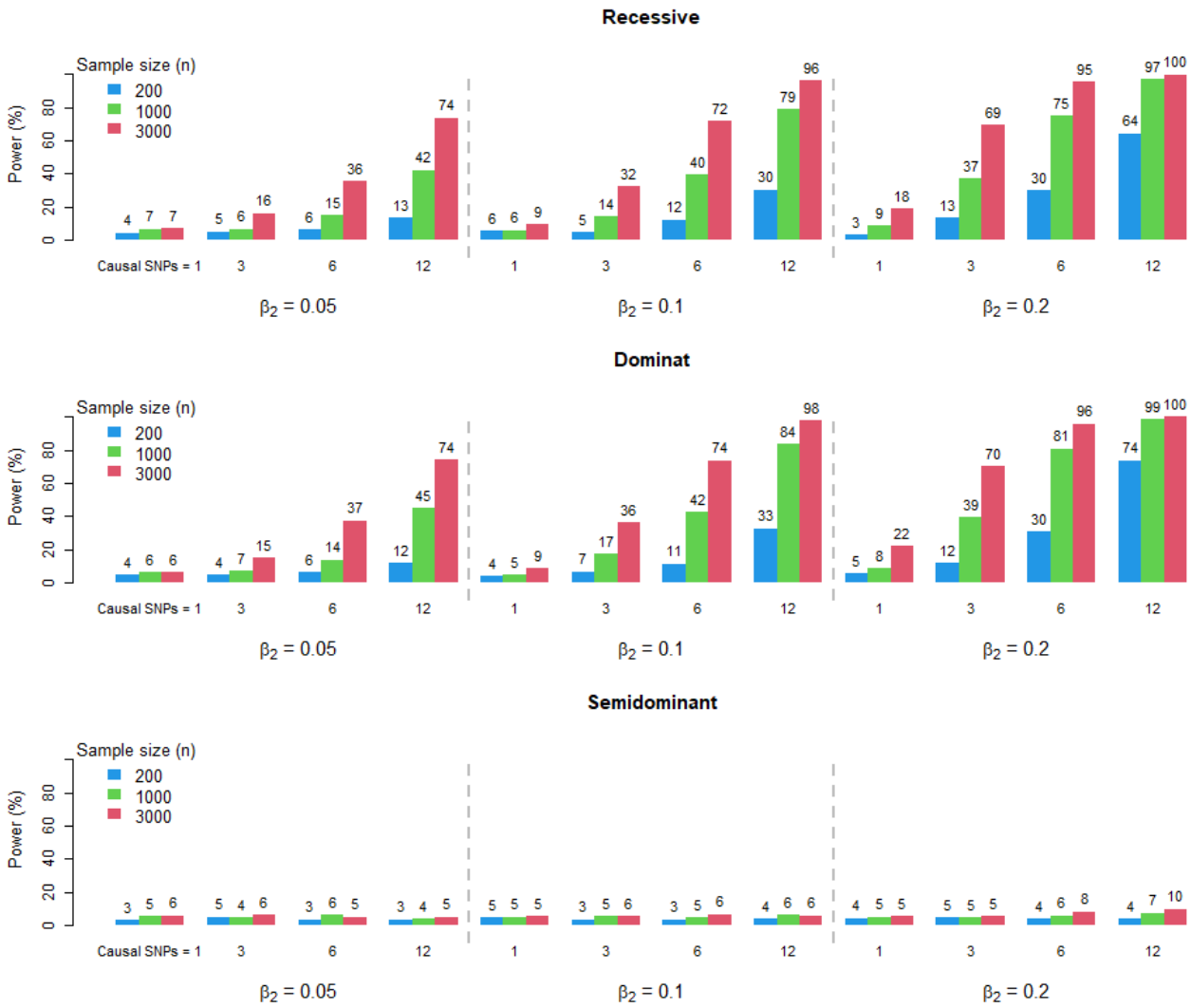

Figure 3. Power of gene-HWT. The results shown are based on simulations with sample size (n)=200, 1000 and 3000 using data from chromosome 20 of EAS population in the 1000 Genomes Phase 3. The genotype risk ratios for a specific causal SNP are defined as AA : Aa : aa = 1 : (1+β_1) : (1+β_2). Simulations were performed for the recessive model (β_1=0, β_2>0), dominant model (β_1=β_2>0), and semidominant model (2β_1=β_2>0).

(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuit
It is made available under a CC-BY-NC-ND 4.0 International license .

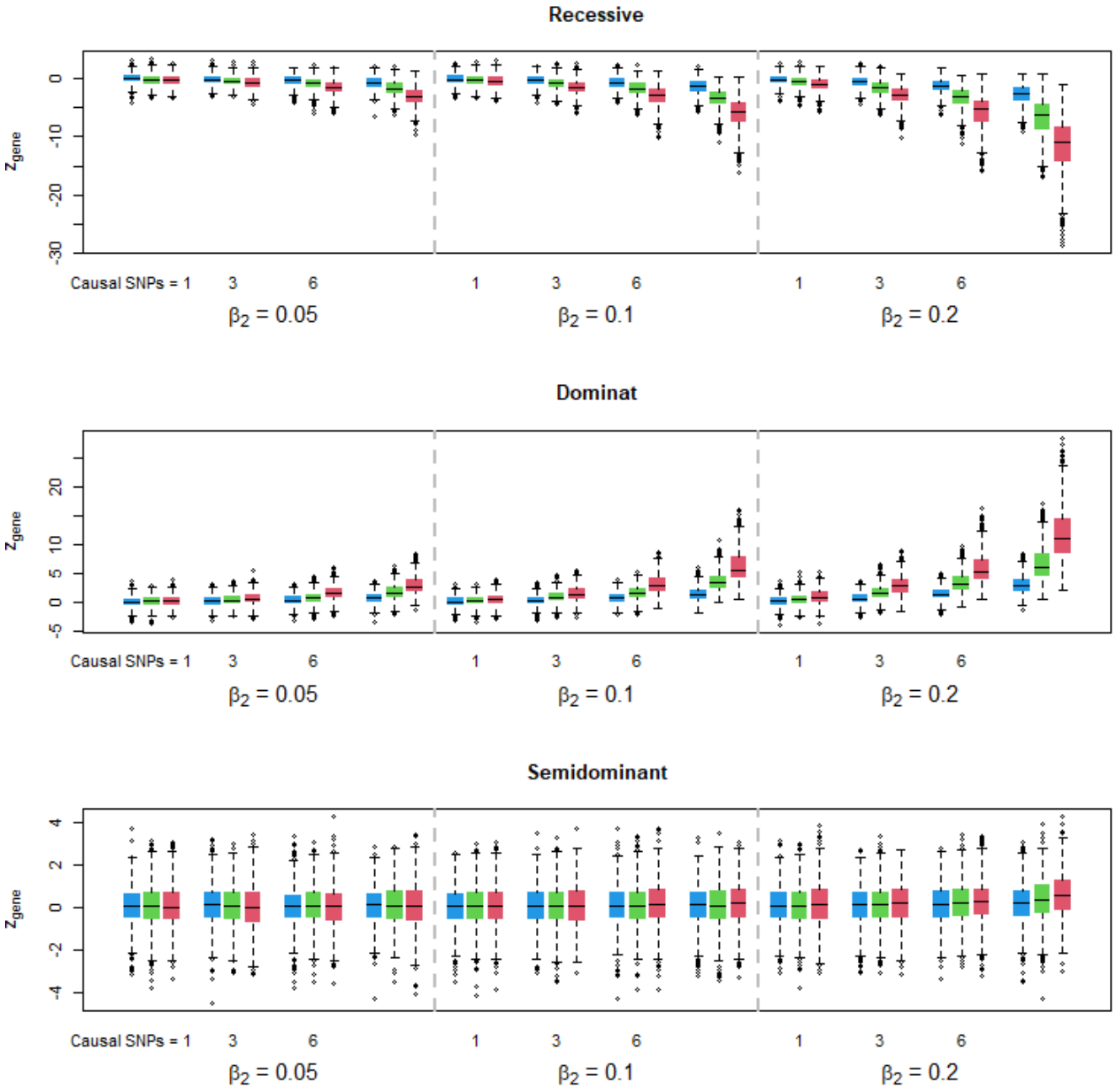

Figure 4. z-values of gene-HWT (z_{gene}). z-values of GHWT obtained from the same simulations as in Figure 2.