Genomics & Informatics is indexed/tracked/covered by PubMed, PubMed Central, Scopus, KoreaMed, KoMCI, ScienceCentral, CrossRef, BIOSIS Previews, DOAJ, and Google Scholar.
Aims and scope

*Genomics & Informatics* is the official journal of the Korea Genome Organization (http://kogo.or.kr). Its abbreviated title is Genomics Inform. It was launched in 2003 by the Korea Genome Organization. It aims at making a substantial contribution to the understanding of any areas of genomics or informatics. Its scope includes novel data on the topics of gene discovery, comparative genome analyses, molecular and human evolution, informatics, genome structure and function, technological innovations and applications, statistical and mathematical methods, cutting-edge genetic and physical mapping, next generation sequencing and de novo assembly, and other topics that present data where sequence information is used to address biological concerns. Especially, Clinical genomics section is for a short report of all kinds of genome analysis data from clinical field such as cancer, diverse complex diseases and genetic diseases. It encourages submission of the cancer panel analysis data for a single cancer patient or a group of patients. It also encourages deposition of the genome data into designated database. Genome archives section is for a short manuscript announcing the genetic information of recently sequenced prokaryotic and eukaryotic genomes. These genome archives data can make the rationale for sequencing a specific organism.

It is published and distributed quarterly at the last dates of March, June, September, and December. All submitted manuscripts will be reviewed and selected for publication after single blind review process. All manuscripts must be submitted online through the e-submission system available from: http://Submit.genominfo.org. It is an online-only peer reviewed open access journal. A free full text both in the XML and PDF formats is available from the journal homepage (https://genominfo.org). It has been indexed by or searchable from PubMed, PubMed Central, Scopus, BIOSIS Previews, KoreaMed, KoMCI, Korea Citation Index, CrossRef metadata, DOAJ, and Google Scholar. This journal was supported by the Korean Federation of Science and Technology Societies Grant funded by the Korean Government.

- Manuscript Editing by InfoLumi Co., Seongnam, Korea.
- E-submission system by Inforang, Seoul, Korea
- PDF layout, XML production, and homepage management by M2PI Co., Seoul, Korea

Published by the Korea Genome Organization

Contact information
Park, Taesung, Editor-in-Chief

Editorial office of Genomics & Informatics
Room No. 806, 193 Mallijae-ro, Jung-gu, Seoul 04501, Korea
Tel: +82-2-558-9394, Fax: +82-2-558-9434, email: kogo3@kogo.or.kr, URL address: https://genominfo.org

Disclaimer: The publisher, editors, and reviewers do not assume any legal responsibility for errors, omissions, or claims, nor do they provide any warranty, expressed or implied, with respect to information published in Genomics & Informatics

© Copyright 2022, the Korea Genome Organization

It is an open access journal. The articles are distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Editorial

Editor’s introduction to this issue (G&I 20:2, 2022)
Taesung Park

Original articles

Bayesian analysis of longitudinal traits in the Korea Association Resource (KARE) cohort
Wonil Chung, Hyunji Hwang, Taesung Park

Identification of the associations between genes and quantitative traits using entropy-based kernel density estimation
Jaeyong Yee, Taesung Park, Mira Park

Associations between single-nucleotide polymorphisms of the interleukin-18 gene and breast cancer in Iraqi women
Bilal Fadil Zakariya, Asmaa M. Salih Almohaidi, Seçil Akıllı Şimşek, Areege Mustafa Kamal, Wijdan H. Al-Dabbagh, Safaa A. Al-Waysi

In silico genome wide identification and expression analysis of the WUSCHEL-related homeobox gene family in Medicago sativa
Tianhui Yang, Ting Gao, Chuang Wang, Xiaochun Wang, Caijin Chen, Mei Tian, Weidi Yang

Integrated bioinformatics analysis of validated and circulating miRNAs in ovarian cancer
Berkcan Dogan, Ece Gumusoglu, Ege Ulgen, Osman Ugur Sezerman, Tuba Gunel

Genome characterization and mutation analysis of human influenza A virus in Thailand
Somruthai Rattanaburi, Vorthon Sawaswong, Pattaraporn Nimsamer, Oraphan Mayuramart, Pavarot Sivapornnukul, Ariya Khanwut, Prangwalai Chanchaem, Kritsada Kongnomnan, Nungruthai Suntronwong, Yong Poovorawan, Sunchai Payungporn

Mathematical modeling of the impact of Omicron variant on the COVID-19 situation in South Korea
Jooha Oh, Catherine Apio, Taesung Park

Comparison of survival prediction models for pancreatic cancer: Cox model versus machine learning models
Hyunsuk Kim, Taesung Park, Jinyoung Jang, Seungyeoun Lee

Clinical genomics

A novel mutation in GJC2 associated with hypomyelinating leukodystrophy type 2 disorder
Sajad Rafiee Komachali, Mozghan Sheikholeslami, Mansoor Salehi
Editor's introduction to this issue (G&I 20:2, 2022)

Taesung Park*
Department of Statistics, Seoul National University, Seoul 08826, Korea

In this issue, there are seven original articles and one publication in the category of clinical genomics. In this editorial, I would like to focus on two original articles about genome-wide association studies (GWAS). There have been great successes in large-scale GWAS to identify novel single-nucleotide polymorphisms (SNPs) associated with traits of interest. By combining independent, large-scale GWAS, meta-analyses have revealed additional SNPs. These SNPs have now been utilized for Mendelian randomization analyses for causal inference. There continue to be challenging methodological issues in GWAS, such as the analysis of longitudinal genetic data, gene-gene interactions, and gene-environment interactions.

Dr. Wonil Chung (Soongsil University, Korea), and his collaborators considered a Bayesian mixed model for longitudinal genetic data. In their earlier work, they demonstrated that their Bayesian method showed a higher statistical power than cross-sectional analysis for detecting SNP-time interactions. Through an analysis of Korea Association Resource (KARE) data for various longitudinal obesity traits, they demonstrated how to apply their Bayesian method in a more effective way. They conducted a two-stage analysis. In the first stage, they performed GWAS analyses of cross-sectional traits and applied a meta-analysis to identify candidate SNPs. In the second stage, they applied the Bayesian method to a subset of SNPs selected in the first stage. The main objective of the Bayesian method was to discover SNPs associated with traits of interest and SNP-time interactions. The application of their Bayesian method to KARE data successfully identified several novel SNPs associated with longitudinal obesity-related traits and significant SNP-time interactions.

Dr. Mira Park (Eulji University, Korea) and her collaborators proposed an entropy-based gene-gene interaction analysis. In genetic association studies, using entropy-based mutual information is advantageous in that it does not depend on parametrization. For binary traits, both entropy and conditional entropy can be easily derived. For quantitative traits, however, these values cannot easily be obtained because quantitative traits require an exact evaluation of entropy by estimating the probability density function. Dr. Park and her collaborators proposed a method of combining the kernel density estimation and the entropy estimation with the probability density function. Through extensive simulation studies, they showed that the proposed method performed better or comparably well in detecting gene-gene interactions compared to existing methods, such as multifactor dimensionality reduction.

New research topics have continued to emerge in GWAS. For example, polygenic risk scores have been popularly used in building prediction models using SNPs. I believe that the Bayesian method of Dr. Chung can be easily extended to estimate more accurate polygenic risk scores. Furthermore, Dr. Park’s work on gene-gene interactions can also be utilized in deriving polygenic risk scores that have both marginal genetic effects and interaction effects.
Bayesian analysis of longitudinal traits in the Korea Association Resource (KARE) cohort

Wonil Chung¹,²*, Hyunji Hwang¹, Taesung Park³

¹Department of Statistics and Actuarial Science, Soongsil University, Seoul 06978, Korea
²Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
³Department of Statistics, Seoul National University, Seoul 08826, Korea

Various methodologies for the genetic analysis of longitudinal data have been proposed and applied to data from large-scale genome-wide association studies (GWAS) to identify single nucleotide polymorphisms (SNPs) associated with traits of interest and to detect SNP-time interactions. We recently proposed a grid-based Bayesian mixed model for longitudinal genetic data and showed that our Bayesian method increased the statistical power compared to the corresponding univariate method and well detected SNP-time interactions. In this paper, we further analyze longitudinal obesity-related traits such as body mass index, hip circumference, waist circumference, and waist-hip ratio from Korea Association Resource data to evaluate the proposed Bayesian method. We first conducted GWAS analyses of cross-sectional traits and combined the results of GWAS analyses through a meta-analysis based on a trajectory model and a random-effects model. We then applied our Bayesian method to a subset of SNPs selected by meta-analysis to further discover SNPs associated with traits of interest and SNP-time interactions. The proposed Bayesian method identified several novel SNPs associated with longitudinal obesity-related traits, and almost 25% of the identified SNPs had significant p-values for SNP-time interactions.

Keywords: Bayesian mixed model, KARE data, longitudinal data, obesity-related traits

Introduction

Genome-wide association studies (GWAS) have been extensively performed to identify single nucleotide polymorphisms (SNPs) associated with cross-sectional traits, but GWAS on longitudinal traits have been underexplored [1–5]. Longitudinal traits dynamically change over time, controlled by both genetic and environmental factors. Multiple measurements at various time points under varying environmental conditions are collected as longitudinal traits; thus, conducting a GWAS for longitudinal traits requires considering the covariance among observations at all-time points. Multivariate linear mixed models have been widely adopted for longitudinal GWAS, and efficient multivariate algorithms for longitudinal GWAS have been developed to reduce the computational complexity for SNP-inferred kinship matrix [3,6]. Furthermore, several Bayesian methods such as time-varying-coefficient regression [7], a simple regression-based method [8], and a Gaussian process-based model [9] have been developed for analyzing such longitudinal traits.
Recently, for the analysis of longitudinal genetic data, we developed a Bayesian mixed model with a built-in variable selection feature based on a grid-based covariance estimation approach \cite{10}. The proposed Bayesian method modeled multiple candidate SNPs simultaneously and allowed SNP-SNP and SNP-time/environment interactions, enabling us to detect SNPs with time-varying effects that were of great scientific and medical interest. The proposed grid-based approach modeled the covariance structure nonparametrically; not only was this parsimonious in estimating the covariance matrix, but it also enabled the flexible approximation of any type of covariance structure by employing a reasonable number of grid points. The number of grid points was set in advance, but the deviance information criterion (DIC) \cite{11} and simplified Bayesian predictive information criterion (BPIC) \cite{12} can be utilized to select the optimal number of grid points. Through simulation studies, we showed that the proposed Bayesian method using all-time points outperformed the ordinary Bayesian method with one or all-time points, and statistical power increased as the data had more samples, smaller number of SNPs, a lower proportion of causal SNPs, and larger trait-heritability.

In this paper, we analyze longitudinal obesity-related traits such as body mass index (BMI), hip circumference (HIP), waist circumference (WST), and waist-hip ratio (WHR) from Korea Association Resource (KARE) data to further evaluate the proposed Bayesian method. These traits are well-known markers of obesity across all ages and continuously change over time. We first conduct cross-sectional GWAS analysis at each time point and combine the results of the GWAS analysis of all-time points using a meta-analysis based on a trajectory model \cite{13} and a random-effects model \cite{14}. We then apply our Bayesian mixed model to further discover SNPs associated with the traits and SNP-time/environment interactions. The paper is organized as follows. In Methods section, we describe the KARE data, GWAS analysis, meta-analysis, and summarize our grid-based Bayesian mixed model for longitudinal genetic data. In the Results section, we present the analysis results for longitudinal obesity-related traits such as BMI, HIP, WST, and WHR from KARE data and evaluate the proposed Bayesian method. We conclude the paper with a discussion on Bayesian longitudinal analysis and future work.

Methods

Study participants

The Korean Genome and Epidemiology Study (KoGES) was a large prospective cohort study that was initiated to solve public health issues and prepare for personalized and preventive health care in Korea \cite{15}. The purpose of the KoGES was to build a genome epidemiological study platform for various researchers and examine the genetic and environmental etiology of complex traits and common diseases such as type 2 diabetes, hypertension, obesity, and metabolic syndrome in Korea. Data collection was initiated in 2001, and follow-up examinations for the participants have been conducted every 2 years. In particular, the availability of GWAS data and repeatedly measured traits in the KoGES Ansan-Ansung study facilitated the identification of genetic variants associated with various disease traits. The GWAS for the KoGES was known as the KARE cohort, which was launched to perform GWAS to discover the underlying genetic variants associated with diverse complex traits and diseases. The KARE data included 10,038 unrelated individuals aged 40–69 years, assembled through the KoGES Ansan-Ansung study, representing urban and countryside populations, respectively \cite{16}. Our analyses utilized the KARE dataset from the baseline (2001–2002) to the sixth (2013–2014) follow-up.

Genotype data

The 10,038 participants in the KARE were genotyped at ~500,000 SNPs using the Affymetrix Genomewide Human SNP array 5.0. For SNP quality control (QC), we removed SNPs with minor allele frequency (MAF) < 0.05, genotype calling rates < 95%, and Hardy-Weinberg equilibrium p-values < 10^{-6}, resulting in 4,518,929 SNPs. For sample QC, we only preserved participants with consistent sex and calling rates > 90%, resulting in 9,331 individuals. The SNP QC and sample QC were performed using PLINK (v1.90) \cite{17}. We then imputed the genotyped SNP data using the Beagle 5.0 software \cite{18} after the SNP and sample QCs. We restricted the analysis to Hapmap3 SNPs with MAF > 0.05 and imputation R^2 > 0.80, which consisted of 3,848,960 SNPs.

GWAS analysis

In the GWAS for cross-sectional obesity-related traits such as BMI, HIP, WST, and WHR, we performed linear mixed model (LMM) association analysis implemented in the BOLT-LMM (v2.3) software \cite{19} using genotyped and imputed genetic variants from the KARE data \cite{20}. The current analysis adjusted for age, age squared, sex, drinking, smoking status, and 10 genotype principal components (PCs). In order to obtain independent SNPs (i.e., a smaller number of clumps of correlated SNPs), we reduced the genome-wide scan using the PLINK clumping function based on empirical estimates of linkage disequilibrium (LD) between genetic variants loci (with four main flags: --clump-p1 5e-8 --clump-p2 1e-5 --clump-r2 0.1 --clump-kb 1000). The clumping
procedure found index SNPs ($p < 5 \times 10^{-8}$) that were independent from each other and formed clumps of other SNPs that were located within 1,000 kb from each index SNP and in LD with the index SNP based on $r^2 > 0.1$. Clumps annotated to the same gene were further combined to the clump with the most significant index SNP.

**Meta-analysis using trajectory model**

To discover SNPs associated with the overall trend and trajectory of phenotypes, we prepared two different longitudinal outcomes (the average and trajectory), which were obtained from longitudinal BMI, HIP, WST, and WHR measurements (baseline to sixth follow-up). To assess the average outcome, we calculated the average BMI, HIP, WST, and WHR measurements over the six time points and then used the average values as the phenotype for GWAS analysis. To assess the trajectory outcome, we utilized the methodology in Gouveia et al. [13] designed to study cognitive trajectories. The LMM function implemented in the R package lme4 [21] was used to obtain the per-individual trajectory for all obesity-related traits. Specifically, we fitted the following mixed model for subject $i$ at time $t$: $y_i = (\beta_0 + x_{it} \beta_1 + \ldots + x_{itq} \beta_{1q}) + \nu_i + e_i$ (where $y_i$ is a phenotypic value; $x_{itq}$ is a random effect allowing variation around the intercept $\beta_{ij}$; $\nu_i$ is a random effect allowing variation around the slope for age covariate $\beta_{ij}$; $x_{itq} (h = 1, \ldots, q)$ are $q$ covariates including age ($h=1$) with fixed effects $\beta_{ij}$ and $e_i$ is a random error. We assumed that the random intercept $\nu_i$ and the random slope $e_i$ were independent of each other and normally distributed. The models were adjusted for age squared, sex, drinking, and smoking status. We then used the predicted random slope ($\nu_i$) from the model as the phenotype for a GWAS analysis of trajectory outcomes. All GWAS results for average and trajectory outcomes were obtained using BOLT-LMM software. In the GWAS for trajectory outcomes, the models were adjusted for only 10 genotype PCs to avoid over-adjustment.

**Meta-analysis using a random-effects model**

We utilized two different meta-analysis methods based on a fixed-effects model (Meta-Fixed) and a random-effects model (Meta-Random) to combine GWAS results for cross-sectional traits across all time points (baseline to sixth follow-up). The Meta-Fixed method, as exemplified by the inverse-variance weighted method [22] and weighted sum-of-z-scores method [22], assumed that the magnitude of the true effects was common or fixed in every GWAS study and did not model heterogeneity. $p$-values for all GWAS analyses were converted to z-scores and then combined with different weights for each study according to their sample sizes. Meanwhile, the Meta-Random method proposed by DerSimonian and Laird [23] and Lee et al. [14] assumed that the true effect size of each study was sampled from an underlying distribution and modeled heterogeneity, explicitly. The Meta-Random method had two modifications: (1) accounting for correlation among test statistics based on the Lin and Sullivan method [24], and (2) focusing on heterogeneous effects conditioned on the Meta-Fixed method. The Meta-Random method can obtain higher power to detect heterogeneous effects than the Meta-Fixed method and it allows the testing of heterogeneous effect sizes between individual test statistics.

**Grid-based Bayesian mixed model**

To further discover SNPs associated with traits of interest and SNPs interacting with time/environmental factors, we considered the following Bayesian mixed model:

\[
y_i = \mu + x_{it} \beta + x_{it} \beta_{x} + x_{itq} \beta_{xq} + x_{it} \beta_{x} + \nu_i + e_i \quad (i = 1, \ldots, n),
\]

where $y_i = (y_{i1}, \ldots, y_{in})^T$ is an $n \times 1$ phenotype vector of individual $i$; $\mu = \mu_1$, is an $n \times 1$ overall mean vector; $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$ is the design matrix corresponding to $q$ time/environmental covariates, $p$ SNPs, two-way interactions among $p$ SNPs (resulting in total of $p(p-1)/2$ terms), and $pq$ SNP-time/SNP-environment interactions; $\beta = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14})^T$ is a vector of genetic effects, time/environmental effects, epistasis effects, and SNP-time/environment interactions; $\nu$ is an $n \times 1$ vector of random errors with $e_i \sim N (0, \sigma^2 I_n)$. For a given trait, we had $n$ individuals, where individual $i$ has phenotypic values measured at $n$ time points $(i = 1, \ldots, n)$, and the total number of observations was $N = \sum_i n_i$. To model the correlation among the repeated measurements of the same individual, we partitioned the observed time interval by $k$ pre-specified grid points, $t = (t_1, \ldots, t_k)^T$, and defined $v$ as a $k \times 1$ vector of random effects at the grid time points with $v_i \sim N(0, D)$, where $D$ is a $k \times k$ covariance matrix. We also defined the incidence matrix $p$, as $p_i = (p_{i1}, \ldots, p_{ik})^T$. If all subjects had $k$ observations measured exactly on the $k$ grid time points, then $p_i$ became an identity matrix. We applied a linear interpolation procedure to any observation that did not fall on any one of the grid time points. When the $j$th measurement of individual $i$ fell at time $t$, which was in between the grid points $t$ and $t_{i+1}$, $t_i < t < t_{i+1}$, we set $p_i = (0, \frac{t_i - t}{t_{i+1} - t_i}, \frac{t - t_i}{t_{i+1} - t_i}, 0, \ldots, 0)$. When $t = t_i$, we get $p_i = (0, 1, 0, \ldots, 0)$. For Bayesian estimation of the mixed-effects model (1), we ap-
plied the modified Cholesky decomposition of Chen and Dunson [25] to the $k \times k$ covariance matrix $D$, resulting in the decomposition, $D = \Delta Y \Psi \Delta$, where $\Delta$ is a nonnegative $k \times k$ diagonal matrix and $\Psi$ is a $k \times k$ lower triangular with 1’s in the diagonal elements. We re-parameterized model (1) as

$$y_i = \mu + x_i \beta + p_i \Delta \Psi b_i + e_i (i = 1, \ldots, n).$$

where $b_i = (b_{i1}, \ldots, b_{ik})^T$ such that $b_{ij} \sim N(0, 1)$ and $b_{ij} \perp b_{ij}' (j \neq j')$, $j = 1, \ldots, k$.

As priors for $\Delta$ and $\Psi$, we defined two vectors $\delta = (\delta_i : l = 1, \ldots, k)^T$ and $\psi = (\psi_m : m = 2, \ldots, k; l = 1, \ldots, m - 1)^T$. The prior distribution for $\delta$ is $P(\delta) = \Pi_{l=1}^k P(\delta_l) = \Pi_{l=1}^k N(\delta_l | m_{l0}, \delta_0)$ where $N(\delta_l | m_{l0}, \delta_0)$ was the density of a half normal distribution that was a $N(\delta_l | m_{l0}, \delta_0^*)$ density truncated below by zero. The prior distribution for $\psi$ was $P(\psi) = N(\psi_m | R_m)$ where $\psi_m$ and $R_m$ are pre-specified hyperparameters. The joint prior distribution of $\theta = (b_{11}, \ldots, b_{kk})^T$ was $P(\theta) = N(0, \Theta_{ll})$. The prior for the ath genetic effect was a normal distribution, $P(\beta_a | \gamma_\alpha, \sigma_a^2) \sim N(0, \gamma_\alpha \sigma_a^2)$ and the prior for the variance $\sigma_a^2$ was a scaled inverse $\chi^2$ distribution. The prior for the overall mean $\mu$ was given by $P(\mu) \sim N(0, \tau_0^2)$. The prior for the residual variance $\sigma^2$ was chosen as a scaled inverse $\chi^2$ distribution, $P(\sigma^2) \sim \text{inv-}\chi^2(\nu, s^2)$. The joint posterior distribution was proportional to the product of the likelihood and the prior distributions of all unknown parameters. We utilized Metropolis-Hastings and Gibbs sampling algorithms, and alternately updated each unknown parameter or set of unknown parameters conditional on all the other parameters and the observed data.

The posterior samples can be used to approximate the posterior distribution of the parameters. The posterior inclusion probability of each SNP was calculated using its inclusion proportion in the Markov chain Monte Carlo samples. The Bayes factor (BF) can be calculated to quantify the evidence for the inclusion of a specific parameter or set of unknown parameters conditional on all the other parameters and the observed data.

Results

GWAS analysis

We conducted a GWAS analysis of KARE data on longitudinal obesity-related phenotypes such as BMI, HIP, WST, and WHR. Seven covariates (age, age squared, sex, area, drinking, smoking status and 10 genotype PCs) were included in the analysis. Among 9,331 individuals, we selected 4,621 individuals who had no missing values for either phenotypes or covariates. The number of measurements for each individual was six, and thus the total number of observations was 27,726. Hapmap3 SNPs with MAF > 0.05 and imputation $R^2 > 0.8$ were retained for the GWAS analyses, resulting in 3,848,960 SNPs.

GWAS analyses for cross-sectional obesity-related phenotypes were performed using LMM association analysis implemented in the BOLT-LMM. To assess the cross-sectional outcomes, we analyzed all measurements (baseline to sixth follow-up) separately for each phenotype. Fig. 1 displays Manhattan plots for the baseline BMI, HIP, WST, and WHR measurements and Supplementary Figs. 1–5 in the supplementary materials show the Manhattan plots for the second, third, fourth, fifth, and sixth measurements. Supplementary Figs. 6–11 show the corresponding Quantile-Quantile (Q-Q) plots. For BMI, one SNP on chromosome 6 (baseline) reached GWAS significance ($p < 5 \times 10^{-8}$), and for HIP, five SNPs on chromosome 11 (baseline), one SNP on chromosome 12 (third follow-up), two SNPs on chromosome 12 (fourth follow-up), and one SNP on chromosome 6 (sixth follow-up) exceeded the threshold for GWAS significance in Supplementary Table 1. For WST, 11 SNPs on chromosome 11 (baseline) and for WHR, 16 SNPs on chromosome 12 (second follow-up), one SNP on chromosome 12 (fifth follow-up), one SNP on chromosome 18 (second follow-up), and one SNP on chromosome 18 (fifth follow-up) reached GWAS significance (see Supplementary Table 1).

Meta-analysis

In order to identify SNPs associated with the overall trend and trajectory of the phenotype, we utilized two types of longitudinal outcomes (average and trajectory), obtained for BMI, HIP, WST, and WHR measurements across consecutive examinations of each individual. For the average outcomes, we calculated the average of all measurements from baseline to sixth follow-up. For the trajectory outcomes, we estimated the per-individual trajectory for all obesity-related traits by fitting a mixed model in the R package...
The model contained a random intercept and a slope for age, as well as other covariates such as age squared, sex, area, and drinking and smoking status. The predicted random slopes of all individuals for age were used as the phenotype for GWAS on trajectory outcomes. We performed BOLT-LMM analyses with these average and trajectory outcomes for four obesity-related traits. Supplementary Fig. 12 shows Manhattan plots for the average outcomes, and Supplementary Fig. 13 shows Manhattan plots for the trajectory outcomes. Supplementary Figs. 14 and 15 show the corresponding Q-Q plots. No GWAS significant SNPs were found for either the average or trajectory outcomes, suggesting that a simple average and simple trajectory of longitudinal outcomes cannot effectively combine the cross-sectional traits; therefore, more sophisticated approaches are necessary.

To effectively combine all GWAS results for cross-sectional traits, we performed two different meta-analyses (Meta-Fixed and Meta-Random; see the Methods section) using RE2C software [14]. Because the Meta-Random method accounted for correlation among test statistics for all cross-sectional traits, it would be more appropriate than Meta-Fixed method for our analysis. Supplementary Fig. 16 shows Manhattan plots for the Meta-Fixed method, and Fig. 2 shows Manhattan plots for the Meta-Random method. Supplementary Figs. 17 and 18 show the corresponding Q-Q plots. From the results of Meta-Fixed method, for BMI, two SNPs on chromosome 5 and seven SNPs on chromosome 16 reached GWAS significance, and for WHR, five SNPs on chromosome 11 exceeded the threshold for GWAS significance.
some 12 reached GWAS significance (Supplementary Table 2). From the results of the Meta-Random method, for BMI, two SNPs on chromosome 5 and seven SNPs on chromosome 16 reached GWAS significance, and for WHR, six SNPs on chromosome 12 reached GWAS significance (Supplementary Table 3). Supplementary Fig. 19 shows the scatter plots of $-\log_{10}(p$-value) in GWAS for four longitudinal traits between the Meta-Random method and Meta-Fixed method. As expected, the p-values with the Meta-Random method were smaller than those obtained with the Meta-Fixed method, but there was reasonable concordance between them. In Fig. 3, it can be seen that p-values of the Meta-Random method were more significant than those of the baseline analysis in general.

**Bayesian analysis**

For our Bayesian analysis, a subset of SNPs was selected using LD clumping analysis based on the results of the Meta-Random method to only include SNPs that are not highly correlated with each other (with correlation < 0.5 to avoid multicollinearity). For our Bayesian analysis, we picked a list of the 4,124 SNPs from the LD clumping analyses (option: --clump -kb 250 --p1 0.001 --p2 0.01 --r2 0.5) using PLINK (v1.90) based on summary statistics from the Meta-Random analysis for the four obesity-related traits. Again, our Bayesian analysis included age, age squared, sex, area, drinking, smoking status, and 10 genotype PCs as covariates. The number of grid time points was set to 3 for the analysis based on the DIC and simplified BPIC scores with a different number of
grid points (results not shown). Fig. 4 shows the one-dimensional genome-wide profiles of 2log(BF) for the combined effects (main, epistasis, and SNP-age interactions). Based on the criteria suggested in Jeffreys [26] and Yandell et al. [27], we found 112 SNPs with strong signals (i.e., log$_2$(BF) > 6.8) for BMI, 20 SNPs for HIP, 10 SNPs for WST, and 6 SNPs for WHR in Supplementary Table 4. We compared the results of our Bayesian analysis with p-values from the baseline, average, trajectory, and Meta-Random methods and discovered various newly detected novel loci and reasonable concordance among different approaches.

Also, our Bayesian method can discover SNPs interacting with environmental covariates including age, we needed to investigate...
Fig. 4. Genome-wide profiles of $2\log(BF)$ for all combined effects with body mass index (BMI), hip circumference (HIP), waist circumference (WST), and waist-hip ratio (WHR) measurements from Gridbayes. The two dashed horizontal lines represent the genome-wide thresholds for moderate (Bayes factor $[BF]=10$) strong ($BF=30$) genome-wide associations. We found 112 single nucleotide polymorphisms (SNPs) with strong signals for BMI, 20 SNPs for HIP, 10 SNPs for WST, and six SNPs for WHR.

whether the identified SNPs interacted with age covariate. For SNP-age interaction, we fitted the LMM with each of the four traits (BMI, HIP, WST, and WHR) as a response variable and eight covariates (SNP, age, age squared, sex, area, drinking and smoking status, SNP-age interaction) as predictors: $y_{ij} = (\beta_0 + u_{0i}) + \sum_{h=1}^{7} \beta_h x_{hij} + e_{ij}$ ($i = 1, \ldots, n; j = 1, \ldots, 6$) where $y_{ij}$ is a phenotypic value, $u_{0i}$ is a random effect allowing variation around the intercept $\beta_0$, $\beta_h$ denotes genetic and nongenetic effects for eight covariate values $x_{hij}$ (including SNP-age interaction), and $e_{ij}$ is a random error. Based on p-values for SNP-age interaction in the LMM, we found 30 significant SNPs interacting with age for BMI, three SNPs for HIP, three SNPs for WST, and one SNP for WHR in Table 1, meaning that almost 25% of the identified SNPs using Gridbayes had significant p-values for SNP-age interaction. Based on the results of SNP-age interaction, as shown in Table 1, we displayed age-specific changes in BMI for three different genotypes (0, 1, and 2) at four SNPs (Chr5:149489242, Chr6:96377858, Chr10:18571215, and Chr20:40534573), age-specific changes in HIP at one SNP (Chr8:98187086) and age-specific changes in WST at one SNP (Chr2:133806432) in Fig. 5. These figures clearly showed different slopes in phenotypic values over six time points for three genotypes, meaning that there were SNP-age interactions at the identified genetic loci by Gridbayes.

Discussion

In this paper, we performed a meta-analysis based on GWAS results for cross-sectional obesity-related traits such as BMI, HIP,
Table 1. Genome-wide association results for BMI, HIP, WST, and WHR-associated SNPs with a p-value < 0.05 using the LMM method

<table>
<thead>
<tr>
<th>Trait</th>
<th>Chr</th>
<th>Position</th>
<th>Minor</th>
<th>Major</th>
<th>MAF</th>
<th>Beta</th>
<th>SE</th>
<th>p(LMM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI</td>
<td>13</td>
<td>55546160</td>
<td>C</td>
<td>T</td>
<td>0.3864</td>
<td>0.0047</td>
<td>3.53 x 10^-4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>5532644</td>
<td>A</td>
<td>G</td>
<td>0.3059</td>
<td>0.00503</td>
<td>3.53 x 10^-4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>108287752</td>
<td>C</td>
<td>T</td>
<td>0.1474</td>
<td>0.0638</td>
<td>5.71 x 10^-4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>109243327</td>
<td>C</td>
<td>A</td>
<td>0.2209</td>
<td>0.0580</td>
<td>1.24 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>97721824</td>
<td>G</td>
<td>A</td>
<td>0.3989</td>
<td>0.0484</td>
<td>1.71 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>17340809</td>
<td>A</td>
<td>G</td>
<td>0.1842</td>
<td>0.0596</td>
<td>1.72 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>136606085</td>
<td>C</td>
<td>T</td>
<td>0.4770</td>
<td>0.0442</td>
<td>2.06 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>136608677</td>
<td>G</td>
<td>A</td>
<td>0.4774</td>
<td>0.0442</td>
<td>2.32 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>126243505</td>
<td>T</td>
<td>C</td>
<td>0.4071</td>
<td>0.0480</td>
<td>2.89 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>88585081</td>
<td>A</td>
<td>C</td>
<td>0.0772</td>
<td>0.0860</td>
<td>3.94 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>37949055</td>
<td>G</td>
<td>A</td>
<td>0.2599</td>
<td>0.0503</td>
<td>5.30 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>37832172</td>
<td>A</td>
<td>G</td>
<td>0.4460</td>
<td>0.0475</td>
<td>6.85 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>17378993</td>
<td>T</td>
<td>C</td>
<td>0.2553</td>
<td>0.0520</td>
<td>8.00 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>119401114</td>
<td>T</td>
<td>C</td>
<td>0.4033</td>
<td>0.0452</td>
<td>1.02 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>15637524</td>
<td>T</td>
<td>C</td>
<td>0.4354</td>
<td>0.0519</td>
<td>1.18 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>105882996</td>
<td>C</td>
<td>T</td>
<td>0.3231</td>
<td>0.0550</td>
<td>1.45 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>43260694</td>
<td>C</td>
<td>T</td>
<td>0.1624</td>
<td>0.0613</td>
<td>1.82 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>96377858</td>
<td>T</td>
<td>C</td>
<td>0.2070</td>
<td>0.0601</td>
<td>1.96 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>18671098</td>
<td>T</td>
<td>G</td>
<td>0.4910</td>
<td>0.0454</td>
<td>2.13 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>149489242</td>
<td>T</td>
<td>C</td>
<td>0.1006</td>
<td>0.0785</td>
<td>2.25 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>18571215</td>
<td>G</td>
<td>A</td>
<td>0.3777</td>
<td>0.0495</td>
<td>2.37 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>40534573</td>
<td>T</td>
<td>C</td>
<td>0.3207</td>
<td>0.0488</td>
<td>2.56 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5500434</td>
<td>T</td>
<td>C</td>
<td>0.2601</td>
<td>0.0521</td>
<td>2.70 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>136484839</td>
<td>G</td>
<td>T</td>
<td>0.2642</td>
<td>0.0519</td>
<td>2.76 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>16840158</td>
<td>A</td>
<td>G</td>
<td>0.3718</td>
<td>0.0521</td>
<td>3.02 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>104745772</td>
<td>A</td>
<td>G</td>
<td>0.0685</td>
<td>0.0892</td>
<td>3.19 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>119383320</td>
<td>A</td>
<td>C</td>
<td>0.3169</td>
<td>0.0479</td>
<td>3.42 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>237458209</td>
<td>G</td>
<td>A</td>
<td>0.4012</td>
<td>0.0493</td>
<td>3.75 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>78510165</td>
<td>G</td>
<td>T</td>
<td>0.4001</td>
<td>0.0455</td>
<td>4.02 x 10^-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>20359771</td>
<td>C</td>
<td>A</td>
<td>0.3368</td>
<td>0.0523</td>
<td>4.65 x 10^-2</td>
<td></td>
</tr>
<tr>
<td>HIP</td>
<td>5</td>
<td>28989968</td>
<td>G</td>
<td>T</td>
<td>0.1186</td>
<td>0.0700</td>
<td>6.01 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>98187086</td>
<td>A</td>
<td>G</td>
<td>0.1649</td>
<td>0.0612</td>
<td>8.60 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>64914487</td>
<td>G</td>
<td>A</td>
<td>0.0839</td>
<td>0.0816</td>
<td>4.83 x 10^-3</td>
<td></td>
</tr>
<tr>
<td>WST</td>
<td>2</td>
<td>133232051</td>
<td>A</td>
<td>G</td>
<td>0.0532</td>
<td>0.1009</td>
<td>3.05 x 10^-4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>41264019</td>
<td>C</td>
<td>T</td>
<td>0.1530</td>
<td>0.0645</td>
<td>5.06 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>42586542</td>
<td>C</td>
<td>A</td>
<td>0.4095</td>
<td>0.0457</td>
<td>4.61 x 10^-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>133806432</td>
<td>G</td>
<td>T</td>
<td>0.3763</td>
<td>0.0506</td>
<td>4.96 x 10^-3</td>
<td></td>
</tr>
<tr>
<td>WHR</td>
<td>12</td>
<td>114085922</td>
<td>C</td>
<td>T</td>
<td>0.2865</td>
<td>0.0508</td>
<td>3.23 x 10^-2</td>
<td></td>
</tr>
</tbody>
</table>

BMI, body mass index; HIP, hip circumference; WST, waist circumference; WHR, waist-hip ratio; SNP, single nucleotide polymorphism; LMM, linear mixed model.

WST, and WHR from KARE data and then applied our Bayesian method to a subset of SNPs selected by meta-analysis to further detect causal SNPs and SNPs with time-varying effects. To obtain the meta-analysis results, we first conducted GWAS analyses on obesity-related traits separately for each time point and then combined all GWAS results based on the average, trajectory, Meta-Fixed, and Meta-Random methods. For our Bayesian analysis, we first selected a subset of SNPs based on Meta-Random results and the LD clumping analyses and then applied our Bayesian method to those selected SNPs. We found our Bayesian method newly detected various novel loci and there was a reasonable concordance between our Bayesian analysis and the average, trajectory-
Fig. 5. Time-specific changes in body mass index (BMI) for three different genotypes (0, 1, and 2) at four single nucleotide polymorphisms (SNPs) (Chr5:149489242, Chr6:96377858, Chr10:18571215, and Chr20:40534573), age-specific changes in hip circumference (HIP) at one SNP (Chr8:98187086), and age-specific changes in waist circumference (WST) at one SNP (Chr2:133806432). The figures clearly show different slopes in phenotypic values over six time points for three genotypes, meaning that there are SNP-age interactions at identified the genetic loci.
ry, Meta-Fixed, and Meta-Random methods. We also confirmed that almost 25% of the identified SNPs using our Bayesian method had significant p-values for SNP-age interaction, meaning that our Bayesian method can well detect SNPs interacting with age as well as SNPs associated with traits of interest.

With the limited sample size, we restricted our Bayesian analysis to a subset of SNPs based on the BOLT-LMM analysis and LD clumping analysis. With a sufficient sample size, our method can be applied to all available SNPs. We are currently developing a parallel computing algorithm based on a message passing interface to execute multiple groups of SNPs simultaneously. This will make it feasible to apply our method to large-sample GWAS data.

One limitation of the proposed Bayesian model is its interpretability. It cannot readily conclude whether the identified SNPs had main effects and/or interactions with other covariates including time. Novel findings need to be further evaluated to elucidate their mode of action, as we performed the analysis for SNP-age interaction. If the number of identified SNPs is small, we can estimate the effect of each genotypic combination, which we can use to interpret the genetic mode. Our Bayesian model for GWAS data relied on a set of preselected SNPs. How to select a good set of SNPs, especially those with low marginal effects but high interactions with other SNPs or environmental factors, is challenging and deserves further investigation.

**ORCID**

Wonil Chung: https://orcid.org/0000-0002-5766-6247
Hyunji Hwang: https://orcid.org/0000-0002-5884-5394
Taesung Park: https://orcid.org/0000-0002-8294-590X

**Authors’ Contribution**

Conceptualization: WC. Data curation: WC, HH, TP. Formal analysis: WC, HH. Funding acquisition: WC, TP. Methodology: WC. Writing - original draft: WC, HH. Writing - review & editing: WC, HH, TP.

**Conflicts of Interest**

Taesung Park serves as an editor of the Genomics and Informatics, but has no role in the decision to publish this article. All remaining authors have declared no conflicts of interest.

**Acknowledgments**

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (2020R1C1A01012657) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A1A10044154). This work was supported by the Soongsil University Research Fund. GWAS dataset and epidemiological data for KoGES are third-party data and are available under the approval of the data access committee of the National Biobank of Korea (http://www.nih.go.kr/NIH/eng/contents/NihEngContentView.jsp?cid=65714&menuIds=HOME004-MNU2210-MNU2327-MNU2329-MNU2338).

**Supplementary Materials**

Supplementary data can be found with this article online at http://www.genominfo.org.

**References**

7. Gong Y, Zou F. Varying coefficient models for mapping quantitative trait loci using recombinant inbred intercrosses. Genetics
17. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-575.
Identification of the associations between genes and quantitative traits using entropy-based kernel density estimation

Jaeyong Yee¹, Taesung Park², Mira Park³*

¹Department of Physiology and Biophysics, Eulji University, Daejeon 34824, Korea
²Department of Statistics, Seoul National University, Seoul 08826, Korea
³Department of Preventive Medicine, Eulji University, Daejeon 34824, Korea

Genetic associations have been quantified using a number of statistical measures. Entropy-based mutual information may be one of the more direct ways of estimating the association, in the sense that it does not depend on the parametrization. For this purpose, both the entropy and conditional entropy of the phenotype distribution should be obtained. Quantitative traits, however, do not usually allow an exact evaluation of entropy. The estimation of entropy needs a probability density function, which can be approximated by kernel density estimation. We have investigated the proper sequence of procedures for combining the kernel density estimation and entropy estimation with a probability density function in order to calculate mutual information. Genotypes and their interactions were constructed to set the conditions for conditional entropy. Extensive simulation data created using three types of generating functions were analyzed using two different kernels as well as two types of multifactor dimensionality reduction and another probability density approximation method called m-spacing. The statistical power in terms of correct detection rates was compared. Using kernels was found to be most useful when the trait distributions were more complex than simple normal or gamma distributions. A full-scale genomic dataset was explored to identify associations using the 2-h oral glucose tolerance test results and γ-glutamyl transpeptidase levels as phenotypes. Clearly distinguishable single-nucleotide polymorphisms (SNPs) and interacting SNP pairs associated with these phenotypes were found and listed with empirical p-values.

Keywords: genetic association, kernel density estimation, mutual information, quantitative trait

Introduction

Over the past decades, genetic association studies have been conducted to identify genetic variants associated with various traits or diseases [1–3]. Genetic susceptibility for many complex diseases is often analyzed using diagnosis-based categories, although the underlying phenotypes are usually quantitative [4,5]. A genomic association, however, does not necessarily require any classification. Therefore, the intrinsic features of an association may be better reflected by entering the quantitative distributions into the association measurement in their original form. Furthermore, some traits, such as human height, are intrinsically continuous; therefore, meaningful thresholds for categorization may not exist.
Multifactor dimensionality reduction (MDR) has been successfully used as a genomic association measurement method [6]. It can identify interacting genes, and it was originally intended for binary outcomes. This method uses the classification accuracy, measured by constructing a confusion matrix, to quantify an association. Variants of MDR have emerged. For ordered categorical traits with more than two response categories, ordinal MDR uses Kendall’s tau-b as an association measure [7]. For quantitative traits, generalized MDR (GMDR) and quantitative MDR (QMDR) have been proposed. Inheriting the original MDR, the common strategy of these methods is to classify the trait values corresponding to a genotype as a binary state. GMDR utilizes a score statistic after adjusting covariates [8]. QMDR uses T-statistics as the association measure and compares the mean values for each cell with the overall mean to classify the trait distribution [9].

Entropy-based methods of analyzing genomic associations have emerged as another stream of research [10]. According to information theory, mutual information (MI) is defined as the amount of information, or entropy, shared by two random variables [11,12]. In analyses of genomic associations, this concept can be translated into the strength of the association between the genotype and phenotype [10]. MI is regarded as a generalized correlation measure in the sense that it is not limited to linear dependence [13]. MI has been evaluated as a measure for associations and extended to machine learning [14]. The estimation of MI between discrete or categorical random variables is well established. However, when either of two variables is quantitative, estimating MI is not at all straightforward [15]. MI-based test statistics for gene-gene interactions associated with discrete trait values have been proposed [16,17]. Quantitative traits have also been considered with generalized MI, referred to as “k-way interaction information,” but with the assumption of a normal distribution [18]. A more direct estimation of MI with quantitative traits has been suggested using the m-spacing entropy measure [19]. This method estimates MI utilizing the observed spacing of order m between quantitative trait values, without any assumption or classification attempt. The probability density tends to be inversely proportional to the spacing between data points. M-spacing elaborates upon this notion by considering the spacings beyond the immediately adjacent points, resulting in more accurate estimations of probability density. This is the basis, in turn, for a more precise determination of entropy and MI.

Here, we propose another way of analyzing genomic associations for quantitative traits based on the kernel density estimation (KDE). KDE estimates a distribution function by summing kernels over the domain, or the observed data points. Kernels are designed to be normalized and non-negative functions, symmetric around each data point [20]. MI would be obtained with these estimated distribution functions of quantitative traits. We examined the KDE method by varying the kernels and using adaptive bandwidth for them to determine the most proper way of combining KDE and MI estimations for genomic association data. Associations with gene-gene interactions were investigated with quantitative traits of simulation and real datasets. Statistical power was analyzed in terms of the correct detection rates for extensive sets of simulation data obtained by KDE, two types of MDR, and m-spacing. This comparison showed that using kernels may be more useful than other methods when the trait distributions are more complex than simple normal or gamma distributions. A full-scale genomic dataset with the phenotype of the 2-h oral glucose tolerance test was selected from the Korean Association Resource (KARE) project [21], because the distributions were found to be complex. Additionally, γ-glutamyl transpeptidase (γ-GTP) levels were explored as a phenotype. Single-nucleotide polymorphisms (SNPs) and interacting SNP pairs associated with this phenotype were clearly identified and listed with empirical p-values.

**Methods**

MI between the genotype and the quantitative phenotype is investigated to establish a genomic association. Measuring MI requires estimating the entropy and conditional entropy. To estimate them for a quantitative trait, the probability density function (pdf) needs to be estimated first. KDE has been adopted to estimate the pdf for distributions with or without a boundary effect. Fig. 1A shows the use of an additional factor $J$ with the kernel when the variable is transformed. Fig. 1B visualizes when to apply KDE to genomic data to obtain MI.

**Definition of entropy and MI**

When the probability density function, $f(x)$, is known, the entropy, $H$, is defined in the integral form of the pdf as below, which is also called the differential entropy [22].

$$H(f) = - \int f(x) \ln f(x) dx$$  \hspace{1cm} (1)

MI is defined as the difference between the entropy of one set and that conditioned by the other set, where two sets are interchangeable. MI can quantify the association between two sets [11], which, in the scope of this paper, would be paired observations of the phenotype and genotype values. MI is obtained by the difference between the two entropies above.
where $H(P)$ is the amount of information contained in the phenotype distribution \([10]\). The conditional entropy $H(P|G)$ measures the amount of information still necessary to describe the phenotype distribution when the genotype is known. Equivalently, it is the amount of information that the phenotype distribution does not share with the genotype. Therefore, MI in Eq. (2) quantifies the amount of information that the phenotype and genotype distributions share. The more information they share, the more strongly the genetic information contributes to the phenotype.

### Entropy by KDE

To estimate the entropy in (1), we first need to estimate $f(x)$ from the data \([22]\). Let $\{X_i\}$ denote the set of random samples drawn from a distribution with density $f$. Then, the entropy $H$ can be estimated as follows.

$$H = -\frac{1}{n} \sum_{i=1}^{n} \ln \hat{f}(X_i)$$ \hspace{1cm} (3)

The estimation of entropy now becomes equivalent to the estimation of $f$ (i.e., a pdf). For that purpose, KDE can be used to estimate $f$. A simple and known function $K$, called a kernel, may be defined around each data point and summed for the estimation of a pdf, as shown below \([23]\).

\begin{equation}
\hat{f}(x) = \frac{1}{n} \sum_{j=1}^{n} K \left( \frac{x - X_j}{h} \right)
\end{equation}

Here $K(u)$ should be non-negative and symmetric for our purpose, while satisfying the normalization condition. The requirement for a kernel function that it should be normalized in the range of its argument also ensures the normalization of the pdf \([20]\). It should be noted that at an arbitrary point $x$, the pdf is determined by the sum of $n$ individual kernel functions whose centers are at $x = X_j$. The width of a kernel function is controlled by the bandwidth $h$.

### Kernels for a distribution with a boundary

Some phenotype distributions have distinct boundaries. For example, let us examine the phenotype of $\gamma$-GTP levels, as shown in Fig. 2B. Unlike usual Gaussian distributions, which can be found with weight or blood pressure measurements, this histogram is crowded near the boundary value of zero. A skewed distribution like this can be modeled with a gamma distribution, as presented in Fig. 2A. The range supported is $(0, \infty)$. As suggested in Eq. (4), KDE estimates the pdf as the sum of kernels, which is symmetric around each data point. In Fig. 2C, a few kernels are shown along with the estimated pdf. When the density value is significant near the boundary, as in this case, the estimated pdf inevitably has tails outside the supported range. The normality of the pdf is then bro-
Ken, and the estimated shape of the pdf may not reach the real distribution. This eventually results in an inaccurate estimation of MI. One remedy for this is to use a kernel in the following form 

\[
(5)
\]

Here, the kernel is symmetric in ln- \( x \) space, whose range is \((−\infty, \infty)\). The different Jacobian between Eqs. (4) and (5) should be noted; this can be obtained straightforwardly from the fact that the normality of the kernel function is defined as below.

\[
\int_{-\infty}^{\infty} K(u)du = 1
\]

Transforming back to \( x \) space, the estimated pdf fits better, as shown in Fig. 2D. The kernels will not be symmetric in \( x \) space, and the shape will be dependent on the data point \( X_j \), around which the kernels are estimated.

**Choice of the kernel function**

Several types of kernel functions have been proposed that satisfy the symmetric and non-negative conditions imposed for our purpose [23]. Among them, the Epanechnikov kernel has the highest efficiency, which means that it has the smallest asymptotic mean integrated squared error over other kernels when the number of data points is the same [23]. It has a parabolic form as below.
\[ K(u) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{u^2}{2} \right) \]  

(7)

The indicator function, 1(.,.), is used. Meanwhile, the Gaussian kernel has about 5% lower efficiency, which means it requires 5% more data points to achieve the same error level as the Epanechnikov kernel. However, the Gaussian kernel is widely used because of its mathematical convenience. It has the form given below.

\[ K(u) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{u^2}{2} \right) \]  

(8)

The Epanechnikov kernel is also more advantageous for computation due to its relative simplicity, which becomes an important factor with genomic data containing an extensive set of genotypes [25]. We examined these two kernels.

**Determination of bandwidth**

As can be seen in Eqs. (4) and (5), the bandwidth \( h \) should be determined to make arguments for kernel functions. It also plays the role of a weight factor for the sum of kernels at each point. The value of bandwidth can be deduced by setting the derivative of the asymptotic mean integrated squared error with respect to the bandwidth to zero [26]. However, it has a differentiation term of the pdf, which is obviously unknown. An acknowledged replacement is the sample standard deviation, \( \sigma \), and a constant specific to the kernel used [26]. Its expression is as follows, where \( n \) is the number of data points.

\[ h = \hat{\sigma} C_v(K)n^{-1/(2v+1)} \]  

(9)

The bandwidth in Eq. (9) now depends on the shape of data distribution and the kernel shape. We used \( v = 2 \) and \( C_2 = 2.34, 1.06 \) for the kernels in Eqs. (7) and (8), respectively.

**MI by entropy and conditional entropy**

Combining Eq. (3) with Eqs. (4) or (5), the entropy for the whole phenotype, \( P \), can be estimated as follows.

\[ H(P) = \left[ -\frac{1}{n} \sum_{i=1}^{n} \ln \left( \frac{1}{n} \sum_{j=1}^{n} K\left( \frac{X_j - X_i}{h} \right) \right) \right] \]  

(no boundary effect)

\[ \left[ -\frac{1}{n} \sum_{i=1}^{n} \ln \left( \frac{1}{n} \sum_{j=1}^{n} K\left( \ln X_j - \ln X_i \right) \right) \right] \]  

(boundary effect)

For computing the conditional entropy, the phenotype set needs to be divided according to the corresponding genotypes, represented as \( \{ P \mid G = g \} \). Let \( g \) indicate each genotype and \( d \) be used for the order of genomic interaction. Because each SNP has three different forms (AA, Aa, and aa), \( d \)-order interacting SNPs should have \( 3^d \) possible genotypes. The conditional entropy can now be obtained by summing the above KDE calculations on each subset, weighted by the subset size, as below.

\[ H(P \mid G = g) = \sum_{e = 0}^{K-1} \text{probs}(G = g)H(P \mid G = g) \]  

(11)

\[ \left\{ \left[ -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \ln \left( \frac{1}{n} \sum_{g} \frac{1}{n} K\left( \frac{X_j - X_i}{h} \right) \right) \right] \right\} \]  

(no boundary effect)

\[ \left\{ \left[ -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \ln \left( \frac{1}{n} \sum_{g} \frac{1}{n} K\left( \ln X_j - \ln X_i \right) \right) \right] \right\} \]  

(boundary effect)

**Estimation of p-values**

If statistical significance of the obtained MI is required, the p-value is estimated by random permutation of the trait values among samples to make the resultant dataset satisfy the null hypothesis. The maximum MI value of all genotype combinations from this dataset would form a single point of the null distribution of MI constructed by repeated random permutations. Counting the number of points in this null distribution that are larger than or equal to the observed MI would give the desired empirical p-value.

**Results**

**Generation of simulation data**

The application of KDE to a genomic association study and its performance were examined with a simulated dataset. Simulated data were generated based on the Velez models [27], which assume 2-order SNP interactions for binary phenotypes. Penetrance values, \( t_y \), were tabulated for each of the nine possible genotype combinations of two interacting SNPs, along with specified values of the minor allele frequency (MAF) and heritability. To generate quantitative values, we took the penetrance as the mean of the distribution from which the trait value was sampled. Three types of distributions were considered. The first type was a normal distribution, as given below.

\[ y_{ij} \sim N\left(t_y, \sigma^2\right) \quad i, j = 0, 1, 2 \text{ or } AA, Aa, aa \]  

(12)

Another was a gamma distribution, shown below.

\[ y_{ij} \sim \Gamma\left(k, \theta\right) = \Gamma\left(t_y, \theta, \frac{\sigma^2}{t_y}\right) \quad i, j = 0, 1, 2 \text{ or } AA, Aa, aa \]  

(13)

It should be noted how the penetrance, \( t_y \), was used in the distribution functions above, while \( \sigma \) remained a free parameter. When the penetrance, \( t_y \), was larger or smaller than the overall average value,
the class of the samples for the genotype $ij$ was assigned as high or low risk, respectively. To simulate various situations, three distinct values of $\sigma$, (0.8, 1.0, 1.2), were assigned for high- and low-risk subgroups, independently establishing nine different cases. To further investigate the trait distribution, a third type of trait value sampling was done from a mixed form as shown below, with $\sigma$ set to 0.2.

$$y_{ij} \sim \begin{cases} N(t_{ij} - \alpha, \sigma^2) + N(t_{ij} + \alpha, \sigma^2) / 2 & \text{for high risk} \\ N(t_{ij}, \sigma^2) & \text{for low risk} \end{cases}$$

(14)

The high-risk term in Eq. (14) should not be confused with the sum of normally distributed random variables. In that case, it would make just another normal distribution. Here, the high-risk term was intended to be a Gaussian mixture distribution with double peaks. With Eq. (14), the trait value was generated from a bimodal distribution if $t_{ij}$ was found to be larger than the overall average (i.e., a high-risk case). There were also nine combinations of $\sigma$. The number of SNPs was taken as 20 with a single causal pair and 400 samples. The Velez model has seven heritability values, each of which has five different penetrance tables for two different MAFs. For each of those 70 models, along with nine $\sigma$ combinations, 100 simulations were conducted, yielding $70 \times 90 \times 100$ files for the three distribution schemes, respectively. In all, for each of the seven heritability values, simulated datasets generated from 10 models (five penetrance tables and two MAFs), with nine variations in high- and low-risk samplings from the three types of distributions were considered.

**Demonstration of MI**

Fig. 3 shows how MI works for genomic data. The simplest form of simulation data following Eq. (12), with large heritability (0.4), a MAF of 0.2, and a fixed $\sigma$ of 1.0 was used. The leftmost three vertical lines were for the intended causal SNP pair that was simulated as having strong association. The rightmost three lines were for an arbitrarily chosen SNP pair that was supposed to have little association. KDE was performed on these two SNP pairs with Epanechnikov and Gaussian kernels. Analytic calculations for MI were also conducted, taking advantage of the fact that the analytic form of entropy for a normal distribution was given as $\ln \left( \sigma \sqrt{2\pi} \right)$. MI values were represented by the length of the vertical bar connecting $H(P)$ and $H(P|G)$ values, as defined in Eq. (2). Compared to the unassociated MI, the MI for the associated pair was found to be quite large. Their distinction was clear. The Epanechnikov kernel yielded a closer MI to the analytic result, which should be very close to the true value, than the Gaussian kernel.

**Comparison of hit ratios**

In Fig. 4, the empirical power of our KDE method to identify the causal pair was investigated with the simulation data. The hit ratios using the Epanechnikov (KDE-E) and Gaussian (KDE-G) kernels were compared with the results from other methods (m-spacing, QMDR, and GMDR). Each point in the plot with respect to heritability was obtained from calculations of the hit ratio, taking all of the simulation conditions into consideration. Datasets from the normal and mixed generation functions were analyzed using the “no boundary effect” options in Eqs. (10) and (11), while those from the gamma generation function were analyzed as having a boundary effect. The results are plotted separately in (A)–(C). The two kernels showed quite similar performances throughout all the conditions. Considering the simplicity of the mathematical form, therefore, the Epanechnikov kernel should be chosen whenever the amount of calculation is heavy. For the normal and mixed cases in Fig. 4A and 4C, m-spacing results overlapped with the KDE results in high-heritability regions, although small discrepancies might exist for low-heritability regions. However, GMDR and QMDR showed somewhat different performances. In these two cases, shown in Fig. 4A and 4C, only QMDR for high-heritability regions with a normal distribution outperformed KDE, while GMDR showed the lowest performance regardless of the condi-

![Fig. 3. Demonstration of the association strength of a simulated genomic data obtained by kernel density estimation. Length of the vertical line between the paired points of $H(P)$ and $H(P|G)$ represents the association strength measure by mutual information.](https://doi.org/10.5808/gi.22033)
tions. In Fig. 4B, for the phenotypes whose values were drawn from gamma distributions, KDE outperformed all other methods, regardless of the choice of kernels. GMDR performed best only in the two highest-heritability regions. QMDR and GMDR showed an obvious pattern of performance reversal depending on the data generation schemes. GMDR uses a scoring system, and m-spacing does not assume any pdf shape. Therefore, their performance depends little on the shape of the distributions. In contrast, QMDR tries dichotomization, which may take more advantage of symmetric than asymmetric distributions, such as gamma distribution. This may explain the performance reversal between them.

KDE is also a non-parametric method, as is m-spacing. With a symmetric distributions in Fig. 4A and 4C, KDE's performance was found to overlap with that of m-spacing, while showing slightly better performance in the low-heritability region. With a heavily skewed distribution, as in Fig. 4B, KDE showed consistently better performance, although not substantially so, throughout the heritability regions. A gamma distribution simulation was designed such that the shape should be distinct from the normal case, with the choice of $\sigma$ in Eq. (13). Since $t_i$ in that equation is penetrance, which should be smaller than 1, the resultant gamma distribution would have a shape parameter, $k$, smaller than the scale parameter, $\theta$, in most cases because of the used $\sigma$ values. This condition results in a quite skewed gamma distribution, as intended, giving rise to the boundary effect. KDE, as designed, showed consistency and better performance than m-spacing, QMDR, and GMDR, with the exception mentioned above, regardless of the distribution shapes.

Type I error rate

To examine the type I error rate, the same process used to build the simulation dataset was adopted to construct the null dataset, except that no causal pairs were intended. With the null dataset, the empirical p-value was evaluated by permuting the phenotype part 1,000 times. The p-value evaluation was repeated with the entire null dataset. Counting the number of instances in which the p-value obtained turned out to be smaller than the significance threshold
level, which was taken as 0.05, indicates the type I error. Table 1 presents the results. For heritability variation, a total of 9,000 (9 high-low risk deviation combinations × 1,000 repetitions) p-values were produced to estimate the type I error rate for each cell in this table, while for the MAF cells, 31,500 (7 heritability values × 5 penetrance values × 1,000 repetitions) p-values were used. The Epanechnikov kernel was employed. The estimated type I error rate was close to 5% (range, 4.6% to 5.8%), as would be expected if our method preserved this rate. The preservation of type I error by our method was verified over MAF and heritability conditions regardless of the shape of the functions for generating the simulation data.

**Application to real data (2-hour oral glucose tolerance test)**

A genome-wide dataset from the KARE project [21] was investigated for the phenotype of 2-hour oral glucose tolerance test (OGTT-2h) results, as well as γ-GTP levels. The dataset comprised 8,387 valid samples genotyped for 327,872 SNPs over 22 chromosomes. OGTT-2h is often used to diagnose diabetes, with two critical values (140 and 200 mg/dL) [28], as tabulated in Table 2. The OGTT-2h distribution was not too skewed to be regarded as the gamma distribution examined in this paper. Because of the three-stage diagnosis due to the two critical values, a more elaborate categorization than high and low risk might be necessary. Therefore, the OGTT-2h distribution may be explained better with a more complex distribution than a simple normal distribution. Instead, the mixed form examined in Fig. 4C may be appropriate for OGTT-2h. Fig. 5 shows the scree plots for the association strengths estimated with Epanechnikov kernels for the main effect (A) and two-order interactions (B). The top SNPs were identified by rs numbers. The distinction can be observed very clearly, especially in Fig. 5C. In Table 2, the details of the identified SNPs are listed. Among them, rs30500 was found to have a major association with type 1 diabetes by a previous report [29], while rs3780603 was also listed as having an association with type 2 diabetes in another study [30]. It has been suggested that glucose levels can be a prognostic factor in ovarian carcinoma [31]. Interestingly, rs2227311 in Table 3 was also listed as being associated with the risk of ovarian cancer [32]. Diabetic ketoacidosis was recently reported to have an effect on pulmonary disease [33], and rs41417552, found to be associated with the OGTT-2h phenotype, was also reported to be one of the seven associated SNPs associated with pulmonary edema [34]. The top two-order interaction effects on OGTT-2h are listed in Table 3, rs30500, which was selected by the main effect, was also found to participate in the interaction. Its interaction with rs1559347 distinguished itself quite prominently in the two-order association with OGTT-2h.

**Application to real data (γ-GTP)**

The γ-GTP distribution was found to be skewed enough to be regarded as the gamma distribution. Therefore, an analysis was performed by KDE with the boundary effect considered. Fig. 5 shows the scree plots for the association strengths estimated with Epanechnikov kernels for the main effect (B) and two-order interactions (D).

---

**Table 1.** Type I error estimation with a significance level (α) of 0.05

<table>
<thead>
<tr>
<th>Type I error rate (%)</th>
<th>Normal</th>
<th>Gamma</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>5.0</td>
<td>5.3</td>
<td>4.9</td>
</tr>
<tr>
<td>0.2</td>
<td>4.7</td>
<td>5.7</td>
<td>5.0</td>
</tr>
<tr>
<td>Heritability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>4.8</td>
<td>4.8</td>
<td>5.3</td>
</tr>
<tr>
<td>0.3</td>
<td>4.9</td>
<td>4.9</td>
<td>4.6</td>
</tr>
<tr>
<td>0.2</td>
<td>4.6</td>
<td>5.6</td>
<td>5.0</td>
</tr>
<tr>
<td>0.1</td>
<td>4.7</td>
<td>5.0</td>
<td>5.1</td>
</tr>
<tr>
<td>0.05</td>
<td>5.0</td>
<td>5.7</td>
<td>4.9</td>
</tr>
<tr>
<td>0.02</td>
<td>4.7</td>
<td>5.8</td>
<td>5.2</td>
</tr>
<tr>
<td>0.01</td>
<td>5.0</td>
<td>5.8</td>
<td>4.6</td>
</tr>
<tr>
<td>Overall</td>
<td>4.8</td>
<td>5.4</td>
<td>5.0</td>
</tr>
</tbody>
</table>

MAF, minor allele frequency.

---

**Table 2.** Main effect found by KDE for OGTT-2h with KARE samples

<table>
<thead>
<tr>
<th>Rs ID</th>
<th>Chromosome</th>
<th>MI</th>
<th>p-value</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1559347</td>
<td>16</td>
<td>0.0069</td>
<td>$2 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>rs2055918</td>
<td>4</td>
<td>0.0066</td>
<td>$3 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>rs30500</td>
<td>5</td>
<td>0.0064</td>
<td>$4 \times 10^{-5}$</td>
<td>[29]</td>
</tr>
<tr>
<td>rs12983584</td>
<td>19</td>
<td>0.0062</td>
<td>$4 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>rs4338946</td>
<td>2</td>
<td>0.0061</td>
<td>$4 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>rs10968001</td>
<td>9</td>
<td>0.0059</td>
<td>$4 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>rs6919172</td>
<td>6</td>
<td>0.0058</td>
<td>$4 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>rs2227311</td>
<td>13</td>
<td>0.0057</td>
<td>$4 \times 10^{-5}$</td>
<td>[32,33]</td>
</tr>
<tr>
<td>rs7468639</td>
<td>9</td>
<td>0.0057</td>
<td>$4 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>rs16898812</td>
<td>5</td>
<td>0.0057</td>
<td>$4 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>rs3780603</td>
<td>9</td>
<td>0.0055</td>
<td>$4 \times 10^{-5}$</td>
<td>[30]</td>
</tr>
<tr>
<td>rs41417552</td>
<td>5</td>
<td>0.0055</td>
<td>$4 \times 10^{-5}$</td>
<td>[34]</td>
</tr>
</tbody>
</table>

KDE, kernel density estimation; OGTT-2h, 2-hour oral glucose tolerance test; KARE, Korean Association Resource; MI, mutual information.

**Table 3.** Interactions found by KDE for OGTT-2h with KARE samples

<table>
<thead>
<tr>
<th>Rs ID pair</th>
<th>Chromosome</th>
<th>MI</th>
<th>p-value</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>(rs30500, rs1559347)</td>
<td>(5,16)</td>
<td>0.0153</td>
<td>$1 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>(rs16898812, rs30500)</td>
<td>(5,5)</td>
<td>0.0140</td>
<td>$1 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>(rs2055918, rs30500)</td>
<td>(4,5)</td>
<td>0.0138</td>
<td>$1 \times 10^{-5}$</td>
<td>-</td>
</tr>
</tbody>
</table>

KDE, kernel density estimation; OGTT-2h, 2-hour oral glucose tolerance test; KARE, Korean Association Resource; MI, mutual information.
A clear distinction can be observed, especially in Fig. 5B. Table 4 lists the details of the identified SNPs. The newly found rs6990123 showed an outstanding association strength compared to others, and it participated in two-order interactions, as shown in Table 5, to make top associated pairs with SNPs absent from the list of the top main effects. Rs2074356 was reported to have a strong association.

**Table 4. Main effect found by KDE for γ-GTP with KARE samples**

<table>
<thead>
<tr>
<th>Rs ID</th>
<th>Chromosome</th>
<th>MI</th>
<th>p-value</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs6990124</td>
<td>8</td>
<td>0.0309</td>
<td>$1 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>rs2074356</td>
<td>12</td>
<td>0.0120</td>
<td>$3 \times 10^{-5}$</td>
<td>[35]</td>
</tr>
<tr>
<td>rs11066280</td>
<td>12</td>
<td>0.0117</td>
<td>$4 \times 10^{-5}$</td>
<td>[36]</td>
</tr>
<tr>
<td>rs4604857</td>
<td>11</td>
<td>0.0105</td>
<td>$1.1 \times 10^{-4}$</td>
<td>-</td>
</tr>
<tr>
<td>rs16872439</td>
<td>8</td>
<td>0.0097</td>
<td>$2.3 \times 10^{-4}$</td>
<td>-</td>
</tr>
<tr>
<td>rs9522473</td>
<td>13</td>
<td>0.0096</td>
<td>$2.4 \times 10^{-4}$</td>
<td>-</td>
</tr>
<tr>
<td>rs2227311</td>
<td>13</td>
<td>0.0091</td>
<td>$3.7 \times 10^{-4}$</td>
<td>-</td>
</tr>
<tr>
<td>rs16875527</td>
<td>4</td>
<td>0.0083</td>
<td>$7.8 \times 10^{-4}$</td>
<td>-</td>
</tr>
<tr>
<td>rs663661</td>
<td>10</td>
<td>0.0081</td>
<td>$8.7 \times 10^{-4}$</td>
<td>-</td>
</tr>
<tr>
<td>rs398182</td>
<td>22</td>
<td>0.0080</td>
<td>$9.5 \times 10^{-4}$</td>
<td>-</td>
</tr>
<tr>
<td>rs12229654</td>
<td>12</td>
<td>0.0075</td>
<td>$1.42 \times 10^{-3}$</td>
<td>[37]</td>
</tr>
</tbody>
</table>

KDE, kernel density estimation; γ-GTP, γ-glutamyl transpeptidase; KARE, Korean Association Resource; MI, mutual information.

**Table 5. Interactions found by KDE for γ-GTP with KARE samples**

<table>
<thead>
<tr>
<th>Rs ID pair</th>
<th>Chromosome</th>
<th>MI</th>
<th>p-value</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>(rs2211730, rs6990124)</td>
<td>(8,8)</td>
<td>0.0392</td>
<td>$1 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>(rs314743, rs6990124)</td>
<td>(5,8)</td>
<td>0.0389</td>
<td>$1 \times 10^{-5}$</td>
<td>-</td>
</tr>
<tr>
<td>(rs6990124, rs1103291)</td>
<td>(8,9)</td>
<td>0.0389</td>
<td>$1 \times 10^{-5}$</td>
<td>-</td>
</tr>
</tbody>
</table>

KDE, kernel density estimation; γ-GTP, γ-glutamyl transpeptidase; KARE, Korean Association Resource; MI, mutual information.

A clear distinction can be observed, especially in Fig. 5B. Table 4 lists the details of the identified SNPs. The newly found rs6990123 showed an outstanding association strength compared to others, and it participated in two-order interactions, as shown in Table 5, to make top associated pairs with SNPs absent from the list of the top main effects. Rs2074356 was reported to have a strong association.
with γ-GTP levels [35], and rs11066280 was reported to have a strong association with type 2 diabetes, which is closely related to γ-GTP [36], rs12229654, which has been reported to be associated with both γ-GTP and high-density lipoprotein cholesterol [37], was also found.

**Discussion**

We investigated genomic associations with quantitative traits, including genomic interactions. Entropy-based MI can measure the association strength if the entropy of the trait could be estimated both by itself and as conditioned on the genotypes. We estimated entropy through KDE.

We explored and compared two types of kernel functions for KDE. The Epanechnikov kernel involves a far lower computational burden than the Gaussian kernel, but it was found to be as powerful as the Gaussian kernel for the genomic association task. There are several other kernels whose efficiencies lie between the Epanechnikov and Gaussian kernels, but under the non-negativity and symmetry constraint, their shapes are quite similar, especially in that their extents are limited by the indicator function, unlike the Gaussian kernel. Therefore, the two kernels investigated may lie at two extremes in terms of efficiency and how they are defined. Other kernels are expected to provide similar results.

When the dataset is made from a skewed distribution with a crowded boundary, using a symmetric kernel inherently leads to an extended tail outside the supported range. A consequence is an incorrect estimation of the association. The real data for γ-GTP, which we reported in the present analysis, may not be correctly analyzed with a usual symmetric kernel. We suggested defining a transformed argument in the kernel to confine the sum of the kernel functions within the supported range. Through these tactics, the hit ratios were found to be stable and superior to those from other methods.

The proposed method can be extended to multivariate phenotype traits, while m-spacing is intrinsically a univariate method. Multivariate traits should be the natural extension of this paper. When the real data are expected to be more complex, beyond a dichotomous classification, our method in this paper would therefore be a legitimate candidate. Phenotypes with more than one threshold can be found, one of which is the OGTT-2h phenotype analyzed here.

Simultaneous associations of SNPs were found with the phenotypes that have been suggested to have OGTT-2h-related traits as a prognostic factor. Therefore, these SNP findings may provide additional evidence for the reported pathways. This might be a benefit of analyzing quantitative traits in their original form.

**ORCID**

Jaeyong Yee: https://orcid.org/0000-0002-4470-8241

Taesung Park: https://orcid.org/0000-0002-8294-590X

Mira Park: https://orcid.org/0000-0003-3827-9089

**Authors’ Contribution**

Conceptualization: MP. Data curation: TP. Formal analysis: JY. Funding acquisition: MP. Methodology: JY, TP, MP. Writing - original draft: JY. Writing - review & editing: JY, TP, MP.

**Conflicts of Interest**

Taesung Park serves as an editor of the Genomics and Informatics, but has no role in the decision to publish this article. All remaining authors have declared no conflicts of interest.

**Acknowledgments**

This research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2021R1A2C1007788).

**References**


Associations between single-nucleotide polymorphisms of the interleukin-18 gene and breast cancer in Iraqi women

Bilal Fadil Zakariya¹, Asmaa M. Salih Almohaidi²*, Seçil Akilli Şimşek¹, Areege Mustafa Kamal³, Wijdan H. Al-Dabbagh⁴, Safaa A. Al-Waysi⁵

¹Department of Biology, Institute of Sciences, Çankırı Karatekin University, Çankırı Merkez 18100, Turkey
²Department of Biology, College of Science for Women, University of Baghdad, Baghdad 10022, Iraq
³Department of Pathology/Oncology, Medical City Teaching Hospital, Baghdad 10011, Iraq
⁴Department of Oncology, Medical City Teaching Hospital, Baghdad 10011, Iraq
⁵Department of Microbiology, Medical City Teaching Hospital, Baghdad 10011, Iraq

According to long-term projections, by 2030, the world’s population is predicted to reach 7.5 billion individuals, and there will be roughly 27 million new cancer cases diagnosed. The global burden of breast cancer (BC) is expected to rise. According to the Ministry of Health-Iraqi Cancer Registry, cancer is the second largest cause of death after cardiovascular disease. This study investigated the interleukin-18 (IL18) single-nucleotide polymorphisms (SNPs) –607C/A rs1946518 and –137G/C rs187238 using the sequence-specific amplification-polymerase chain reaction approach. Regarding the position –607C/A, there was a highly significant difference between the observed and expected frequencies in patients and controls ($\chi^2 = 3.16$ and $\chi^2 = 16.5$), respectively. The AA and CA genotypes were associated with significantly increased BC risk (odds ratio [OR], 3.68; p = 0.004 and OR, 2.83; p = 0.04, respectively). Women with the A allele had a 5.03-fold increased susceptibility to BC. The C allele may be a protective allele against BC (OR, 0.19). Although position –137G/C showed no significant differences in the CC genotype distribution (p = 0.18), the frequency of the CC genotype was significantly higher in patients than in controls. In contrast, patients had a significantly higher frequency of GC genotypes than controls (p = 0.04), which was associated with an increased risk of developing BC (OR, 2.63). The G allele frequency was significantly lower in patients than in controls (55.0% vs. 76.2%, respectively). This SNP may be considered a common genotype in the Iraqi population, with the wild-type G allele having a protective function (OR, 0.19) and the mutant C allele having an environmental effect (OR, 2.63).

Keywords: breast neoplasms, interleukin-18, polymorphisms, rs187238, rs1946518

Introduction

Breast cancer (BC) is the most common cancer in females worldwide, with nearly 2.3 million new cases diagnosed in 2020. It accounts for approximately 11.7% of all cancer cases and 24.5% of all cancer cases in females [1]. Since 2008, the global incidence of BC has increased by more than 20%, and the death rate has risen by 14% [2]. As a result, research on relevant tumor indicators for early diagnosis and monitoring is required, and recent studies have concentrated on the role of the immune system in cancer progres-
sion [3].

BC is a multi-step process that involves numerous genetic changes, such as oncogene activation and cancer suppressor gene inactivation [4]. The gene encoding interleukin (IL)-18 is found on chromosome 11 at positions 11q22.2–q22.3 and has six exons in humans. IL-18 is a multidirectional cytokine that regulates the immune response in various ways; in 1989, IL-18 was first described as an “interferon-inducing factor.” IL-18 plays an essential role in stimulating natural killer cells, and cellular anticancer activities also enhance the expansion of Th1 and cell activation. Furthermore, IL-18 increases the expression of adhesion-related molecules, nitric oxide synthase enzyme synthesis, and chemokine production [5]. IL-18, in combination with IL-2, causes a Th2 cell response and the production of IL-4 and IL-13. Simultaneously, IL-18 can reduce antitumor immunity in a programmed-death-1 (PD-1) dependent manner. PD-1 is a co-inhibitory receptor that constitutes one of the top checkpoints. Many polymorphisms in the IL18 promoter region affect transcript factor binding locations [6], which could be IL-18 expression quantitative trait loci. Several studies have found that cytokine gene polymorphisms impact cytokine production, which may be linked to disease [7].

Single-nucleotide polymorphisms (SNPs) are found in the IL18 gene, particularly in the promoter region bound to −607G/T (rs1946518) and −137G/C (rs187238). These SNPs (−137G/C and −607G/T) in the IL18 gene promoter region are expected to affect IL-18 expression and activity [8]. Arimitsu et al. [9] found that monocytes in individuals with the 137G/G genotype produced considerably more IL-18 than those in individuals with the 137G/C genotype. Furthermore, a link between these two SNPs and cancer susceptibility has been established [10]. Research on the relevance of IL18 polymorphisms in BC risk remains contentious. To our knowledge, only three studies have examined the influence of IL18 genetic variations (−137G/C and −607A/C) on BC susceptibility; Khalili-Azad et al. [11] studied the effect of IL18 polymorphisms (−607A/C and −137G/C) on BC risk in 200 patients and 206 healthy controls. They discovered that CC homozygosity for the −137G/C polymorphism lowered the risk of BC [11]. Other studies found that the IL18 −607A/C polymorphism contributed to an increased risk of BC [12,13].

In light of this information, it is necessary to investigate IL18 SNPs in Iraqi women with BC.

Methods

After the ethical committee of the Medical City directorate approved this study, 100 women who visited the oncology teaching hospital’s breast clinic were recruited from January 28, 2020 to August 11, 2022. Group 1 included 60 women who had recently been histopathologically diagnosed with BC and provided written informed consent for participation in the study. Group 2 included 40 healthy women who served as controls. Patients with other systemic diseases and those taking any hormone-modifying drug were excluded from the study.

DNA extraction and genotyping

Each patient and control had 8–10 mL of blood taken. Blood samples were taken from the cubital vein and placed directly into an anticoagulant tube containing EDTA. Genomic DNA extraction was performed using a Norgen Bioteck kit (Thorold, ON, Canada). The optical density ratio at 260/280 nm was used to assess the quantity and quality of isolated DNA Using a Nano-Drop device (Qubit 4, Invitrogen, Waltham, MA, USA). It was preferable to have a balance of 1.7–1.9. Electrophoresis on a 1% agarose gel was performed to corroborate the findings. Until the genotyping analysis, the DNA samples were stored at −20°C.

Polymerase chain reaction

Polymorphisms were investigated by utilizing sequence-specific amplification–polymerase chain reaction (SSP-PCR) at locations −607C/A rs1946518 and −137G/C rs187238 within the promoter region of IL18. For the targeted location −607C/A, a common reverse primer and two specific forward primers were utilized, with an amplified product size of 196 bp, and a forward control primer was used to amplify a 301-bp fragment covering the polymorphism region as an internal control (Table 1).

In the location −137G/C, a common reverse primer and two specific forward primers were also utilized, with an amplified product size of 261 bp, and a forward control primer was used to amplify a 446-bp fragment covering the polymorphism region as an internal control (Table 2). The polymerase chain reaction (PCR) reactions were carried out in a volume of 25 μL, including 12.5 μL of 2 × Go-Taq Green Master mix (Promega, Madison, WI, USA), 3 μL of genomic DNA, and 6.5 μL of nuclease-free water. All reaction mixtures contained one sequence-specific primer, one Table 1. Primers for the IL18 −607C/A rs1946518 polymorphism

<table>
<thead>
<tr>
<th>Primer sequence</th>
<th>Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward primer 1</td>
<td>5′-GTTGCAGAACAGGTTAAAAATATTAC-3′</td>
</tr>
<tr>
<td>Forward primer 2</td>
<td>5′-GTTGCAGAACAGGTTAAAAATATTAA-3′</td>
</tr>
<tr>
<td>Reverse primer</td>
<td>5′-TACCTCATCGAGACTCCC-3′</td>
</tr>
<tr>
<td>Internal positive control</td>
<td>5′-CTTGTGATATTAATTG-3′</td>
</tr>
</tbody>
</table>

https://doi.org/10.5808/gi.22026
common reverse primer, and one internal position control primer, each at a concentration of 1 μL. As a result, each piece of DNA was subjected to two PCR tests: one for the F1 wild-type allele and one for the F2 mutant allele [8]. The assays were carried out in a thermocycler (Applied Biosystems, Waltham, MA, USA). Denaturation was carried out at 95°C for 5 min, followed by 30 cycles at 94°C for 30 s, 57°C for 30 s, 72°C for 40 s, and 72°C for 10 min. Then, the PCR results were visualized using electrophoresis on 2% agarose gel, and the molecular weight was calculated using a 100-bp DNA ladder.

**Statistical analysis**

The genotype of cytokines was statistically analyzed using the SPSS version 25 (IBM Corp., Armonk, NY, USA). The allele frequencies of the cytokine genes were calculated by the direct gene counting method. A freely available online calculator was used to calculate whether there was a significant departure from Hardy-Weinberg (H-W) equilibrium for two alleles [http://www.had2know.com/academics/Hardy-Weinberg_equilibrium-calculator-3-alleles.html]. H-W equilibrium is the expected frequency of genotypes if mating is non-assortative and there are no mutations from one allele to another. When there are two alleles for a particular gene, A and B, and their respective population frequencies are p and q, the expected frequencies of the genotypes AA, AB, and BB are \( p^2 \), \( 2pq \), and \( q^2 \), respectively. The Pearson chi-square test was used to assess whether there were significant differences between the observed and expected frequencies. The alleles and genotypes of cytokines were presented as percentages and frequencies, and the two-tailed Fisher exact test was calculated to assess the significance of differences between their distributions in BC patients and controls. Odds ratios (ORs) were also estimated to define the association between cytokine alleles and genotypes with BC. OR values can range from between 0 and 1 (for a negative association) to more than 1 (for a positive association).

**Results**

**IL18 –607C/A (rs1946518) SSP-PCR**

This study analyzed the distribution of the genotype and allele frequencies of the rs1946518 polymorphism (at the –607C/A position) in patients and controls (Fig. 1). This polymorphism presented three genotypes (CC, CA, and AA) that corresponded to two alleles (T and A) in BC patients and controls. The genotype frequencies in both groups were not in agreement with H-W equilibrium, and there were highly significant differences between the observed and expected frequencies in both patients and controls (\( \chi^2 = 3.16 \) and \( \chi^2 = 16.5 \), respectively) (Table 3).

The CC genotype was significantly less common in patients if mating is non-assortative and there are no mutations from one allele to another. When there are two alleles for a particular gene, A and B, and their respective population frequencies are p and q, the expected frequencies of the genotypes AA, AB, and BB are \( p^2 \), \( 2pq \), and \( q^2 \), respectively. The Pearson chi-square test was used to assess whether there were significant differences between the observed and expected frequencies. The alleles and genotypes of cytokines were presented as percentages and frequencies, and the two-tailed Fisher exact test was calculated to assess the significance of differences between their distributions in BC patients and controls. Odds ratios (ORs) were also estimated to define the association between cytokine alleles and genotypes with BC. OR values can range from between 0 and 1 (for a negative association) to more than 1 (for a positive association).

**Table 2. Primers for the IL18 –137G/C, rs187238 polymorphism**

<table>
<thead>
<tr>
<th>Primer sequence</th>
<th>Forward primer 1</th>
<th>Forward primer 2</th>
<th>Reverse primer</th>
<th>Internal positive control</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'-CCCCAACCTTTACGGAAGAAAG-3'</td>
<td>5'-CCCCAACCTTTACGGAAGAAAAC-3'</td>
<td>5'-AGGAAGGGCAAAATGCACTGG-3'</td>
<td>5'-CCAATAGGACTGATTATTCGCA-3'</td>
<td></td>
</tr>
</tbody>
</table>

*IL18, interleukin-18.*

**Fig. 1.** Agarose gel electrophoresis image that shows the sequence-specific amplification-polymerase chain reaction product analysis of interleukin-18 (IL18) –607 (rs1946518) (C/A) gene polymorphism. Where M, marker (100–1,200 bp), the presence of C or A allele were observed at 196-bp product size. The (CC) wild type homozygote were showed in C allele only, the (AA) mutant type homozygote were showed in A allele only, whereas the (CA) heterozygote were showed in both C and A allele, internal control at 301-bp product size.
than in controls (15.0% vs. 65.5%, \( p < 0.001 \)), and the magnitude of this negative association was 0.11. However, the AA genotype frequency was higher in patients (51.7%) than in controls.

**Table 3.** Hardy-Weinberg equilibrium–expected genotype frequencies in *IL18* rs1946518

<table>
<thead>
<tr>
<th>Groups</th>
<th>CC</th>
<th>CA</th>
<th>AA</th>
<th>C</th>
<th>A</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient genotypes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td>9</td>
<td>20</td>
<td>31</td>
<td>0.68</td>
<td>0.32</td>
<td>3.16*</td>
</tr>
<tr>
<td>Expected</td>
<td>6.0</td>
<td>26.0</td>
<td>28.0</td>
<td>Not detected</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control genotypes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td>25</td>
<td>6</td>
<td>9</td>
<td>0.30</td>
<td>0.70</td>
<td>16.5*</td>
</tr>
<tr>
<td>Expected</td>
<td>19.6</td>
<td>16.8</td>
<td>3.6</td>
<td>Not detected</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

*IL18*, interleukin-18.

\( \chi^2 > 3.84 \), significant.

**Table 4.** Genotype distribution and allele frequency of *IL18* rs1946518 (–607C/A) in breast cancer patients and controls with risk estimation

<table>
<thead>
<tr>
<th>Study group</th>
<th>Patient [n = 60]</th>
<th>Control [n = 40]</th>
<th>OR (95% CI)</th>
<th>Fisher's exact probability*</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C/C</td>
<td>9 (15.0)</td>
<td>25 (62.5)</td>
<td>0.11 (0.04–0.28)</td>
<td>0.000</td>
<td>0.0001***</td>
</tr>
<tr>
<td>C/A</td>
<td>20 (33.3)</td>
<td>6 (15.0)</td>
<td>2.83 (1.02–7.86)</td>
<td>0.061</td>
<td>0.040*</td>
</tr>
<tr>
<td>A/A</td>
<td>31 (51.7)</td>
<td>9 (22.5)</td>
<td>3.68 (1.49–9.04)</td>
<td>0.003</td>
<td>0.004**</td>
</tr>
<tr>
<td>Allele distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>38 (31.7)</td>
<td>56 (70.0)</td>
<td>0.19 (0.11–0.36)</td>
<td>0.000</td>
<td>0.0001***</td>
</tr>
<tr>
<td>A</td>
<td>82 (68.3)</td>
<td>24 (30.0)</td>
<td>5.03 (2.72–9.30)</td>
<td>0.000</td>
<td></td>
</tr>
</tbody>
</table>

*IL18*, interleukin-18; OR, odds ratio; CI, confidence interval.

Significant, *\( p < 0.05 \), **\( p < 0.01 \), ***\( p < 0.001 \).

**Fig. 2.** Agarose gel electrophoresis image that shows the sequence-specific amplification–polymerase chain reaction product analysis of interleukin-18 (*IL18*) –137 (rs187238) (G/C) gene polymorphism. Where M, marker (100–1,200 bp), the presence of G or C allele were observed at 261-bp product size, the (GG) wild type homozygote were showed in G allele only, the (CC) mutant type homozygote were showed in C allele only, whereas the (G/C) heterozygote were showed in both G and C allele. internal control at 446-bp product size.
there was a highly significant difference between the observed and expected frequencies ($\chi^2 = 5.74$) (Table 5).

At the position –137G/C in IL18, patients had a significantly higher frequency of the GC genotype than controls (43.3% vs. 22.5%; OR, 2.63; $p = 0.04$). In contrast, the frequency of the CC genotype was significantly higher in patients than in controls (23.3% and 12.5%; OR, 2.13). Therefore, the G allele frequency was significantly lower in patients than in controls (55.0% vs. 76.2%, respectively), with a highly significant difference ($p = 0.002$). A negative association was found for the G allele (OR, 0.38) and a positive association for the C allele (OR, 2.36) (Table 6).

Discussion

The present study investigated the two most commonly studied SNPs of the IL18 gene (–607C/A and –137G/C). The patients and controls both showed deviation from H-W equilibrium for the –607C/A genotype, which may have been related to BC or intermarriage in Arab Iraqi society between relatives. This result aligns with that of a previous study on BC [14].

Although the distribution of the –137G/C genotype in the patient group was consistent with H-W equilibrium, this was not the case for the control group, which showed a highly significant difference between the observed and expected frequencies ($\chi^2 = 5.74$). This SNP may be considered a common genotype in the Iraqi population, with the wild-type G allele having a protective function, reducing susceptibility, and the mutant C allele having an environmental effect.

These SNPs’ genotypes and alleles showed significant differences between BC patients and controls. The present study observed that the IL18 –607 CA and AA genotypes were present in about 85% of BC cases. The wild-type CC genotype had a low frequency (about 15%) in BC patients, which may highlight the role of the IL18 –607 polymorphism in the pathogenesis of disease. The A allele was an environmental effect allele, while the C allele had a preventive fraction because the CC genotype showed the highest frequency in the control group (70%). These highly significant findings for the AA and CA genotypes at position IL18 –607 suggest that this polymorphism may play a role in cancer progression.

Several studies on IL18 polymorphisms have been conducted in various populations in multiple countries; one of them has found a link between IL18 polymorphisms and the risk of BC [15]. The IL18 –607C/A polymorphism may be linked to an increased risk of BC in Asian and mixed populations [16]. Furthermore, the present results showed that the IL18 rs1946518 SNP might play a role in BC because the wild-type allele C of the –607 SNP had a protective effect against BC (OR, 0.19). In contrast, the mutant allele A had a positive association (OR, 5.03) suggesting an etiological impact; therefore, women who carry allele A of –607 may be more susceptible to BC than women who have allele C. However, the results of some studies were mixed; in a study involving 72 BC patients and 93 control women, Fathi Maroufi et al. [17] discov-

Table 5. Hardy-Weinberg equilibrium–expected genotype frequencies in IL18 rs187238

<table>
<thead>
<tr>
<th>Groups</th>
<th>CC</th>
<th>CA</th>
<th>AA</th>
<th>C</th>
<th>A</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients’ genotypes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td>20</td>
<td>26</td>
<td>14</td>
<td>0.45</td>
<td>0.55</td>
<td>0.93 NS</td>
</tr>
<tr>
<td>Expected</td>
<td>18.2</td>
<td>29.7</td>
<td>12.2</td>
<td>Not detected</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control genotypes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td>26</td>
<td>9</td>
<td>5</td>
<td>0.24</td>
<td>0.76</td>
<td>5.74$^a$</td>
</tr>
<tr>
<td>Expected</td>
<td>23.3</td>
<td>14.5</td>
<td>2.30</td>
<td>Not detected</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

$^a\chi^2 > 3.84$, significant.

Table 6. Genotype distribution and allele frequency of IL18 rs187238 (–137G/C) in breast cancer patients and controls with risk estimation

<table>
<thead>
<tr>
<th>Genotype groups</th>
<th>Study group</th>
<th>OR (95% CI)</th>
<th>Fisher’s exact probability*</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patient (n = 60)</td>
<td>Control (n = 40)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G/G</td>
<td>20 (33.3)</td>
<td>26 (65.0)</td>
<td>0.27 (0.12–0.62)</td>
<td>0.002</td>
</tr>
<tr>
<td>G/C</td>
<td>26 (43.3)</td>
<td>9 (22.5)</td>
<td>2.63 (1.07–6.48)</td>
<td>0.035</td>
</tr>
<tr>
<td>C/C</td>
<td>14 (23.3)</td>
<td>5 (12.5)</td>
<td>2.13 (0.70–6.47)</td>
<td>0.203</td>
</tr>
<tr>
<td>Alleles distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>66 (65.0)</td>
<td>61 (76.2)</td>
<td>0.38 (0.20–0.71)</td>
<td>0.002</td>
</tr>
<tr>
<td>C</td>
<td>54 (45.0)</td>
<td>19 (23.8)</td>
<td>2.63 (1.40–4.92)</td>
<td>0.03</td>
</tr>
</tbody>
</table>

$^a\chi^2 > 3.84$, significant. Significiant, *p < 0.05, **p < 0.01.
erated that the \textit{IL18} –607A/C polymorphism was not linked to BC in an Iranian population sample. The \textit{IL18} rs187238 polymorphism results showed no significant differences according to the CC genotype frequencies between BC patients and controls.

Nonetheless, it was interesting to note in the present study that the heterozygous GC genotype had an OR of 2.63 for the patient group, implying that the mutant allele C may have had an environmental effect on the Iraqi population, conferring susceptibility to BC, while the wild-type allele G had a preventive effect against BC (OR, 0.38). Our findings show that the \textit{IL18} –137G/C polymorphism is associated with the development of BC. The conversion of G (guanine) to C (cytosine) at position –137G/C of the \textit{IL18} gene removes a nuclear factor binding site for histone-4 transcriptional factor-1 \cite{18}. Genetic variants have been considered the most critical cancer risk factors. Although high-penetrant capability genes (e.g., \textit{BRCA1} and \textit{BRCA2}) have strong links to BC, low-penetrant susceptibility genes that predispose individuals to the disease have yet to be identified; nonetheless, immune responses and surveillance may be affected by genetic variability in a sequence of immune regulatory genes \cite{19}. The \textit{IL18} promoter polymorphism –137G/C has previously been linked to various cancers in different populations, including esophageal squamous cell malignant tumors, prostate cancer in the Chinese population \cite{20}, colorectal cancer in Greek people \cite{21}, and ovarian cancer in native Hawaiians \cite{22}. Additional case-control studies on BC and gastric cancer progression have been published \cite{23,24}. However, there is no link between type 2 diabetes mellitus development and the \textit{IL18} –137G/C gene polymorphism \cite{25}.

### ORCID

- Bilal Fadil Zakariya: https://orcid.org/0000-0002-2354-4065
- Asmaa M. Salih Almohaidi: https://orcid.org/0000-0002-6084-0349
- Seçil Akilli Şimşek: https://orcid.org/0000-0002-5055-1391
- Areege Mustafa Kamal: https://orcid.org/0000-0001-5239-9699
- Wijdan H. Al-Dabbagh: https://orcid.org/0000-0002-8927-327X
- Safaa A. Al-Waysi: https://orcid.org/0000-0001-5940-8084

### Authors’ Contribution

- Conceptualization: AMSA. Data curation: BFZ, WHA. Formal analysis: BFZ. Funding acquisition: BFZ. Methodology: AMK, SAA. Writing – original draft: BFZ. Writing – review & editing: SAŞ, AMSA.

### Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

### Acknowledgments

The authors are grateful to everyone who volunteered for this study. We want to express our heartfelt gratitude to the entire Institute of Science at Cankiri Karatekin University staff for their kind assistance. Also, many thanks go to all the staff at the Oncology Teaching Hospital, Medical City, Baghdad, especially the Cancer Research Laboratory, for their help with the diagnosis and sample collection.

### References

1416.
**In silico** genome wide identification and expression analysis of the WUSCHEL–related homeobox gene family in *Medicago sativa*

Tianhui Yang¹, Ting Gao¹*, Chuang Wang¹, Xiaochun Wang¹, Caijin Chen², Mei Tian³, Weidi Yang¹

¹Institute of Animal Science, Ningxia Academy of Agriculture and Forestry Sciences, Yinchuan 750002, China
²Branch Institute of Guyuan, Ningxia Academy of Agriculture and Forestry Sciences, Guyuan 756000, China
³Institute of Horticultural Science, Ningxia Academy of Agriculture and Forestry Sciences, Yinchuan 750002, China

Alfalfa (*Medicago sativa*) is an important food and feed crop which rich in mineral sources. The WUSCHEL–related homeobox (WOX) gene family plays important roles in plant development and identification of putative gene families, their structure, and potential functions is a primary step for not only understanding the genetic mechanisms behind various biological process but also for genetic improvement. A variety of computational tools, including MAFFT, HMMER, hidden Markov models, Pfam, SMART, MEGA, ProtTest, BLASTn, and BRAD, among others, were used. We identified 34 MsWOX genes based on a systematic analysis of the alfalfa plant genome spread in eight chromosomes. This is an expansion of the gene family which we attribute to observed chromosomal duplications. Sequence alignment analysis revealed 61 conserved proteins containing a homeodomain. Phylogenetic study sung reveal five evolutionary clades with 15 motif distributions. Gene structure analysis reveals various exon, intron, and untranslated structures which are consistent in genes from similar clades. Functional analysis prediction of promoter regions reveals various transcription binding sites containing key growth, development, and stress–responsive transcription factor families such as MYB, ERF, AP2, and NAC which are spread across the genes. Most of the genes are predicted to be in the nucleus. Also, there are duplication events in some genes which explain the expansion of the family. The present research provides a clue on the potential roles of *MsWOX* family genes that will be useful for further understanding their functional roles in alfalfa plants.

**Keywords:** chromosome organization, cis element, *Medicago sativa*, synteny, WOX

---

**Introduction**

The WUSCHEL (WUS)-related homeobox (WOX) gene group is a subfamily that includes plant-specific transcription factors. WOX participates in multiple developmental activities, especially stem cell maintenance and organ development [1]. Typically, WOX amino acids fold into a DNA-binding domain termed the homeodomain, which is encoded by the homeobox (HB) DNA sequence and a homeodomain with 60–66 amino acid residues [2]. In higher plants, many homeodomains-containing transcriptional fac-
tor proteins have been identified in both monocots and dicots [3]. The HB protein superfamily is classified into six families. These classifications include homodomain-leucine zipper (HD-Zip); plant homeodomain (PHD)-finger; BELL; zinc finger-homeodomain (ZF-HD); WOX; and KNOTTED1-like-homeobox (KNOX) [4].

Of these homeodomains, members of the WUS-related gene family have been comprehensively identified or predicted in many plants, such as Arabidopsis, maize, soybean, rice, etc. [5-7]. The model eudicot plant Arabidopsis (Arabidopsis thaliana) contains 15 WOX proteins, which are classified into three clades based on evolutionary relationships i.e., a modern/WUS clade, an intermediate clade, and an ancient clade [8].

Most WUS members are involved in multiple developmental processes, including embryonic development, embryonic polarization, meristematic stem cell maintenance, lateral organ development, seed formation, and regeneration of isolated tissues and organs [9]. Aside from plant growth and development, another important factor to consider is plant response to environmental stresses. In order to survive extreme environmental stress, plants have evolved multiple mechanisms as a defense strategy against external signals by modulating gene expression [10]. Transcription factors such as WRKY, MADS-BOX, NACs, BHLH, and HSF, among many others, have been confirmed to play key roles in regulating plant response to abiotic stresses and are listed in the plant stress transcription factor database (http://caps.ncbs.res.in/stifdb). Recent evidence suggests that WOX genes also play a role in the regulation of abiotic stress resistance. A poplar WOX11/12a gene, for example, has been shown to play an important role in drought tolerance [11]. Overexpression of WOX13 under the rab21 promoter increased drought stress tolerance in rice [12]. HOS9 has been shown in Arabidopsis to regulate cold stress tolerance [13]. Even though several members of the WOX family have been cloned and functionally studied, little is still known about these members and their roles in many plants.

Alfalfa is a popular food and feed crop that is farmed all over the world. It is typically collected as hay, but it can also be processed into silage, grazed, or supplied fresh. The development and quality of alfalfa is limited due to numerous difficult conditions such as a lack of water, cold temperatures, and excessive salt, and productivity is lowered by at least 10%–20% [14]. The entire genome data of the autotetraploid cultivar XinjiangDaYe were released in 2020, resulting in a chromosome-level genome assembly with 32 genes [15]. This gathering will give a wealth of information for identifying important stress-related genes and genetically engineering alfalfa stress tolerance.

In this study therefore, we analyzed the WOX gene family in alfalfa based on available genome sequence. We identified 34 genes of the WOX gene family based on a genome-wide scan approach and predicted their functions by combining the analysis of the phylogenetic tree with that cis-promoting element. We also studied the chromosomal location and gene structures as well as subcellular localization. This study provides further insight into the structure and function of the WOX gene family in alfalfa and is useful for their further genetic studies.

Methods
Identification of WOX members in alfalfa
The latest versions of the genome annotations of alfalfa were retrieved from the genome assembly (https://www.alfalfatoolbox.org/). Previously reported Arabidopsis WOX full-length and homeodomain amino acid sequences were retrieved from The Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org/) aligned with MAFFT v5.3 [16] and then subjected to HMMER v3.0 [17] for building HMM (hidden Markov models) profiles. The HMM profiles were applied to perform HMM search against the annotated alfalfa protein databases with an E-value cutoff of 1e-5. Furthermore, using both the full-length and homeodomain amino acid sequences of Arabidopsis WOXs, a BLASTP search with an E-value cutoff of 0.01 was carried out to identify additional potential WOX proteins. The protein sequences from the two methods outlined above were merged, and redundant entries were manually eliminated. Pfam (https://pfam.xfam.org/) and SMART (http://smart.embl.de/) [18] were used to check the hit sequences for the presence of the homeobox domain.

Sequence alignment and phylogenetic analysis
MAFFT v5.3 was used to align the full-length and homeodomain amino acid sequences of Arabidopsis WOXs and probable WOX members from the alfalfa species using the default parameters with manual editing. Based on the alignment data, two distinct approaches to constructing phylogenetic trees were used. First, using MEGA software 6.06 [19] under the following parameters: Poisson correction, pairwise deletion, and bootstrap values, a neighbor-joining tree was created from the alignment of full-length amino acid sequences of Arabidopsis and alfalfa WOX members (1,000 replicates). Second, the alignment of homeodomain amino acid sequences of AtWOXs and alfalfa WOXs was used to create a Bayesian inference (BI) tree [20]. With the help of ProtTest 2.4, the same JTT + G model was chosen. MrBayes (http://mrbayes.sourceforge.net/) [21] was used to analyze the BI tree for 2,500,000 generations, with trees sampled
every 1,000 generations and a burn-in of 625 while a FigTree v1.4.0 was used to view these tree files.

**Gene chemical structure analyses**
Using both coding and genomic sequences, exon-intron structures were analyzed and depicted using the Gene Structure Display Server (GSDS, http://gsds.cbi.pku.edu.cn/). MEME 4.9.1 (http://meme-suite.org/) was used to identify conserved motifs in *Arabidopsis* WOXs and alfalfa WOX proteins, and WebLogo (http://weblogo.berkeley.edu/logo.cgi) [22] was used to visualize them. The following parameters were set: the distribution of motif occurrences was set to zero or one per sequence; the maximum number of motifs was set to eight; the optimal motif width was set to six and 100; and the optimal number of sites for each motif was set to two and 200.

**Chromosome location and gene duplication analysis**
The distribution of members of the alfalfa WOX family on the alfalfa chromosomes was examined using the WOX gene annotation information in the alfalfa genome database. The duplicate gene pairs were searched from the plant genome duplication database server (http://chibba.agtec.uga.edu/duplication/index/locket). The amino acid sequence of the partly repeated MsWOX gene was determined using the Clustalw program [23].

**Subcellular localization prediction**
A web-based interface for predicting the subcellular localization was retrieved from pre-built ngLOC model database (http://ngloc.unmc.edu/). To generate predictions, WOX protein sequences were supplied in the FASTA format and alfalfa species set as default. the MLCS (Multi-Localization Confidence Score) [24] was searched which reflects if the top two locations are predicted within a close probability margin.

**Synteny and Ka/Ks analysis**
BLASTn was also used to see if Darwinian positive selection impacted the evolution of the MsWOX genes in alfalfa and its diploid progenitors. MCScanX was used to look for synteny blocks containing WOX genes between various alfalfa and *Arabidopsis* genomes and/or subgenomes. Each WOX gene was successfully mapped onto the relevant *Arabidopsis* chromosomes, according to information collected from gff3 files of genome annotation data. To illustrate the blocks and collinearity of homologous gene pairs, the CIRCOS software was utilized. The nonsynonymous to synonymous substitution ratios (Ka/Ks) of all orthologous, paralogous, and homeologous gene pairs were estimated using the program KaKs Calculator version 2.0 using the model average and model averaging methods based on coding sequence alignment.

**Tissue-specific expression analysis**
RNA-sequencing (RNA-seq) data from Phytozome 12 were used to examine the expression patterns of MsWOX genes in distinct tissues. The study covered six different tissues: root, nodule, leaves, flower, pre-elongated stem, and elongated stem. We used the FPKM (fragments per kilobase of transcript per million mapped fragments) data to create a heatmap.

**Results**

**Identification and phylogenetic analysis of MsWOX genes in alfalfa**
Homeodomain sequences from previously identified *Arabidopsis* WOX proteins were used to construct an HMM profile, which was then used as a query to perform HMM searches across related protein databases to identify WOX family genes in alfalfa. BLASTP searches were also carried out to identify other potential WOX proteins, with full-length and homeodomain sequences of known *Arabidopsis* WOX proteins serving as queries. To eliminate duplicate sequences, manual reconstruction was employed. The use of Pfam and SMART analyses to confirm the presence of the homeodomain in each candidate protein increased the dependability of these candidate sequences. Consequently, we discovered 34 candidate WOX genes, which corresponded to previous study (Table 1). These genes encoded proteins ranging in size from 14,609.5 to 153,183.84 kDa, with an instability index of 32 to 66.91, an aliphatic index of 49.78 to 88.07, and isoelectric point (pI) values ranging from 5.33 to 9.79 (Table 1).

**Sequence alignment and phylogenetic analysis**
WOX proteins are plant-specific proteins with a conserved homeodomain. As a result, the sequences of these proteins were aligned to produce sequence logos in order to determine if the domain is conserved in MsWOX members. The findings of the alignment indicated that the homeodomain was highly conserved among the genes (Fig. 1). The homeodomain featured a helix-loop-helix-turn-helix structure and was 61 amino acids long. The homeodomain sequences revealed three highly conserved residues in particular: Q in helix 1, L in helix 2, and W in helix 3. R and E in helix 1, I and P in helix 2, and Q and F in helix 3 are among the other highly conserved amino acid residues in the homeodomain. Surprisingly, all the above-mentioned amino acid residues were found to be highly conserved across all the homeodomain se-
Table 1. Characteristics of alfalfa WOX gene family

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Gene location</th>
<th>A.A. No.</th>
<th>Weight (Da)</th>
<th>Isoelectric points</th>
<th>Instability index</th>
<th>Aliphatic index</th>
<th>GRAVY</th>
</tr>
</thead>
<tbody>
<tr>
<td>MsWOX1</td>
<td>Chr1:3806424-3808395</td>
<td>187</td>
<td>21561.3</td>
<td>9.79</td>
<td>57.8</td>
<td>61.98</td>
<td>-0.858</td>
</tr>
<tr>
<td>MsWOX2</td>
<td>Chr3:85829292-85830284</td>
<td>274</td>
<td>30,925.1</td>
<td>8.52</td>
<td>51.31</td>
<td>70.11</td>
<td>-0.795</td>
</tr>
<tr>
<td>MsWOX3</td>
<td>Chr7:4947090-4947467</td>
<td>274</td>
<td>31,285.2</td>
<td>6.43</td>
<td>49.72</td>
<td>70.77</td>
<td>-0.678</td>
</tr>
<tr>
<td>MsWOX4</td>
<td>Chr7:79672330-79681060</td>
<td>315</td>
<td>34,498.8</td>
<td>6.98</td>
<td>48.96</td>
<td>69.9</td>
<td>-0.43</td>
</tr>
<tr>
<td>MsWOX5</td>
<td>Chr8:241473-243772</td>
<td>832</td>
<td>91,168.9</td>
<td>5.92</td>
<td>51.79</td>
<td>86.07</td>
<td>-0.151</td>
</tr>
<tr>
<td>MsWOX6</td>
<td>Chr8:7362896-7370122</td>
<td>213</td>
<td>24,537.7</td>
<td>6.67</td>
<td>48.71</td>
<td>82.86</td>
<td>-0.888</td>
</tr>
<tr>
<td>MsWOX7</td>
<td>Chr8:22027384-22034137</td>
<td>860</td>
<td>93,547.5</td>
<td>5.99</td>
<td>45.86</td>
<td>86.83</td>
<td>-0.075</td>
</tr>
<tr>
<td>MsWOX8</td>
<td>Chr8:67803988-67805460</td>
<td>1387</td>
<td>153,184</td>
<td>6.11</td>
<td>44.9</td>
<td>88.07</td>
<td>-0.21</td>
</tr>
<tr>
<td>MsWOX9</td>
<td>Chr8:69416480-69417965</td>
<td>832</td>
<td>94,413.6</td>
<td>5.02</td>
<td>52.81</td>
<td>77.34</td>
<td>-0.742</td>
</tr>
<tr>
<td>MsWOX10</td>
<td>Chr1:5066220-5068376</td>
<td>524</td>
<td>59,633.3</td>
<td>5.99</td>
<td>52.97</td>
<td>57.67</td>
<td>-1.055</td>
</tr>
<tr>
<td>MsWOX11</td>
<td>Chr1:100681352-100684467</td>
<td>245</td>
<td>28,008.2</td>
<td>5.33</td>
<td>66.91</td>
<td>52.5</td>
<td>-1.284</td>
</tr>
<tr>
<td>MsWOX12</td>
<td>Chr2:6284516-6286598</td>
<td>245</td>
<td>20,020.2</td>
<td>5.33</td>
<td>65.36</td>
<td>53.71</td>
<td>-1.284</td>
</tr>
<tr>
<td>MsWOX13</td>
<td>Chr2:19355940-19361319</td>
<td>294</td>
<td>33,344</td>
<td>6.57</td>
<td>56.22</td>
<td>63.37</td>
<td>-0.914</td>
</tr>
<tr>
<td>MsWOX14</td>
<td>Chr2:28821465-28822950</td>
<td>870</td>
<td>95,414.2</td>
<td>6.86</td>
<td>44.34</td>
<td>86.25</td>
<td>-0.166</td>
</tr>
<tr>
<td>MsWOX15</td>
<td>Chr2:76808477-76813101</td>
<td>832</td>
<td>91,266.1</td>
<td>5.99</td>
<td>46.75</td>
<td>87.12</td>
<td>-0.11</td>
</tr>
<tr>
<td>MsWOX16</td>
<td>Chr2:81776107-81791035</td>
<td>233</td>
<td>25,965</td>
<td>5.58</td>
<td>48.85</td>
<td>68.58</td>
<td>-0.545</td>
</tr>
<tr>
<td>MsWOX17</td>
<td>Chr2:83689682-83677900</td>
<td>409</td>
<td>47,097.2</td>
<td>8.52</td>
<td>51.12</td>
<td>67.04</td>
<td>-0.589</td>
</tr>
<tr>
<td>MsWOX18</td>
<td>Chr3:38501335-38507307</td>
<td>376</td>
<td>43,056.9</td>
<td>8.42</td>
<td>56.52</td>
<td>62.34</td>
<td>-0.838</td>
</tr>
<tr>
<td>MsWOX19</td>
<td>Chr3:83449292-85346585</td>
<td>306</td>
<td>34,532.9</td>
<td>6.41</td>
<td>60.02</td>
<td>57.42</td>
<td>-0.888</td>
</tr>
<tr>
<td>MsWOX20</td>
<td>Chr3:94036221-9408325</td>
<td>184</td>
<td>21,209.7</td>
<td>6.53</td>
<td>56.58</td>
<td>49.78</td>
<td>-0.79</td>
</tr>
<tr>
<td>MsWOX21</td>
<td>Chr3:99347579-99355806</td>
<td>555</td>
<td>62,699.3</td>
<td>6.31</td>
<td>56.33</td>
<td>56.43</td>
<td>-1.064</td>
</tr>
<tr>
<td>MsWOX22</td>
<td>Chr4:48397488-48403556</td>
<td>262</td>
<td>27,602.4</td>
<td>8.81</td>
<td>50.49</td>
<td>59.24</td>
<td>-0.476</td>
</tr>
<tr>
<td>MsWOX23</td>
<td>Chr4:50875345-50877341</td>
<td>125</td>
<td>14,609.5</td>
<td>6.97</td>
<td>32</td>
<td>71.76</td>
<td>-1.077</td>
</tr>
<tr>
<td>MsWOX24</td>
<td>Chr4:69003312-69005774</td>
<td>418</td>
<td>47,489.3</td>
<td>9.37</td>
<td>48.54</td>
<td>83.06</td>
<td>-0.609</td>
</tr>
<tr>
<td>MsWOX25</td>
<td>Chr4:92442615-9244780</td>
<td>417</td>
<td>46,428.5</td>
<td>5.53</td>
<td>50.26</td>
<td>62.59</td>
<td>-0.621</td>
</tr>
<tr>
<td>MsWOX26</td>
<td>Chr5:13693575-13695090</td>
<td>203</td>
<td>23,472.4</td>
<td>9.23</td>
<td>64.49</td>
<td>60.54</td>
<td>-0.736</td>
</tr>
<tr>
<td>MsWOX27</td>
<td>Chr5:93003502-93004323</td>
<td>778</td>
<td>85,772.5</td>
<td>6.17</td>
<td>50.97</td>
<td>81.22</td>
<td>-0.321</td>
</tr>
<tr>
<td>MsWOX28</td>
<td>Chr5:99882444-99887399</td>
<td>296</td>
<td>32,286.5</td>
<td>7.67</td>
<td>52.38</td>
<td>72.74</td>
<td>-0.317</td>
</tr>
<tr>
<td>MsWOX29</td>
<td>Chr6:2310875-2310420</td>
<td>243</td>
<td>27,513.9</td>
<td>6.75</td>
<td>47.85</td>
<td>65.47</td>
<td>-0.659</td>
</tr>
<tr>
<td>MsWOX30</td>
<td>Chr7:16766076-1677035</td>
<td>218</td>
<td>24,141.7</td>
<td>6.53</td>
<td>50.3</td>
<td>59.95</td>
<td>-0.944</td>
</tr>
<tr>
<td>MsWOX31</td>
<td>Chr7:17564121-17566590</td>
<td>831</td>
<td>91,815.2</td>
<td>6.15</td>
<td>46.72</td>
<td>87.21</td>
<td>-0.152</td>
</tr>
<tr>
<td>MsWOX32</td>
<td>Chr7:51962820-51963935</td>
<td>846</td>
<td>93,017.4</td>
<td>5.84</td>
<td>49.06</td>
<td>82.48</td>
<td>-0.29</td>
</tr>
<tr>
<td>MsWOX33</td>
<td>Chr7:64567276-64574554</td>
<td>270</td>
<td>30,175.7</td>
<td>6.96</td>
<td>53.5</td>
<td>62.52</td>
<td>-0.824</td>
</tr>
<tr>
<td>MsWOX34</td>
<td>Chr7:75163974-75167582</td>
<td>276</td>
<td>31,344.4</td>
<td>7.67</td>
<td>65.18</td>
<td>66.78</td>
<td>-0.752</td>
</tr>
</tbody>
</table>

WOX, WUSCHEL-related homeobox; A.A, amino acid; GRAVY, grand average of hydropathicity index.

In order to examine the evolutionary connection of alfalfa and Arabidopsis WOX members, the MEGA v5.0 platform was used to create a phylogenetic tree using their full-length sequences. The members of the alfalfa clade were given new names based on the original Arabidopsis names discovered in the same group (Fig. 2). The WOX members may be divided into four groups: group 1, group 2, group 3, and group 4. Group 1 has the most Arabidopsis and alfalfa species. While group 4 was the smallest clade in each analyzed species, it had nine WOX members, indicating that WOX members from Arabidopsis and alfalfa species may have an evolutionary link.

Gene motif analysis
The conserved motif analysis validated the differences between the MsWOX genes in the ancient lineage and other clades. A total of 15 motifs were discovered and labeled as motif 1 through motif 15. Motif 3 was found in every MsWOX protein and was shown to encode the WOX homeodomain. Furthermore, genes from the same evolutionary group have nearly identical gene architectures.
For example, motif 4 was shared by all members of group 1. Furthermore, all WOX genes in group 2 had just three motifs, whereas WOX5, WOX6, WOX7, WOX8, and WOX31, which all belong to the first group, had the most motifs. The location of these conserved motifs revealed more about the WOX genes in alfalfa plants. Except for MsWOX9, all MsWOX members in group 1 have a WUS-box; this suggests that MsWOX members from various clades may be engaged in distinct biological processes in alfalfa (Fig. 3).

**MsWOX exon-intron organization**

The exon-intron structure is closely related to the function of the corresponding gene, and it reflects the evolutionary relationship of multigene families when combined with phylogenetic analysis. To gain a better understanding of the MsWOX gene structure in alfalfa, the exon/intron organization was studied. According to the findings, all MsWOX genes have zero to five exons. Surprisingly, both MsWOX13a/13b have three exons in the ancient clade. Generally, the number of exons in MsWOX genes varies between clades. MsWOX8 of group 1, for example, had the most exon number of 29, whereas MsWOX23 of group 2 had only one exon. Small differences in gene characteristic are also observed within members of the same group. For example, MsWOX23 had no intron, whereas MsWOX22, MsWOX26, MsWOX20, MsWOX12, and MsWOX1 all had one despite belonging to the same clade. Untranslated regions were present in 17 members, the vast majority of whom belonged to group 1. But generally, except for MsWOX8 which featured complicated exon-intron structures and significant variation in the number of introns; and MsWOX23 which lacked introns and untranslated regions, exon/intron architectures were found to be similar among WOX members within the same subgroup (Fig. 4). This organization of exons and introns in a gene family could give information about its evolution.

**Chromosome organization**

All the MsWOX genes were mapped into eight chromosomes. The number of genes per chromosome ranged from one in chromosome 6 to seven in chromosome 7. Gene members from similar group were located on same chromosomes. For example, chromosome 1 with three genes, MsWOX1, MsWOX2 of were all located on chromosome 1. MsWOX3 however is from group whose orthologous members such as MsWOX23, MsWOX24, MsWOX25, MsWOX26, MsWOX27, MsWOX28, and MsWOX29 are all located on chromosome 7. There are duplications in MsWOX9 (MsWOX9A and MsWOX9B) and MsWOX5 (MsWOX5A and MsWOX5B) as well as a triplication MsWOX13 (MsWOX13A,
MsWOX13B, and MsWOX13C). Chromosome 6 had only one gene MsWOX22 (Fig. 5).

Collinearity and Ka/Ks analysis
To illustrate the locus linkage of homologous MsWOX genes among genomes, synteny analysis of WOX genes in alfalfa and its diploid progenitors was undertaken. As shown in Fig. 6, syntenic genes were a couple of genes connected by a line, whereas those connected by lines of the same color signified the same type of MsWOX gene, such as MsWOX10 and MsWOX21. As a result, we can see that many chromosomes in all four genomes/subgenomes were connected by the same-colored line, indicating that these genomes/subgenomes were evolutionarily related and that the WOX genes were so crucial that most of them survived polyploidization.

The nonsynonymous (Ka), synonymous (Ks), and Ka/Ks ratios were calculated to assess selection pressure among duplicated WOX gene pairs. Ka/Ks = 1 implies that genes evolved in a neutral fashion; Ka/Ks > 1 or Ka/Ks 0 imply that genes were being positively selected or purified, respectively. All duplicated WOX gene pairs in alfalfa and its diploid progenitors exhibited Ka/Ks values less than

Fig. 2. Phylogenetic tree showing the relatedness of alfalfa WUSCHEL-related homeobox (WOX) proteins to that of Arabidopsis.
one (Supplementary Table 1), with MsWOX15-MsWOX5 having the lowest value of 0.0475257 and MsWOX13-MsWOX15 having the highest value of 0.818763 (Supplementary Table 1).

**Cis-regulatory element and transcription factor binding**

Cis-regulatory elements are unique DNA sequences found upstream of gene coding sequences that govern the expression of stress-responsive genes by interacting with transcription factors. Thus, the cis-elements in the putative promoter regions of the 34 MsWOX genes were investigated to further study the possible characteristics of the MsWOX family genes that are engaged in plant regulatory network management. Consequently, we discovered that MYB included the most cis-regulatory elements, followed by Dof, ERF, Bzip, MIKC-MADS, and AP2 (Fig. 6, Supplementary Table 2). Furthermore, the cis-elements were found in all of the genes. The ERFs were found on MsWOX23 and MsWOX10, whereas the MYBs were found on MsWOX16. TCP is the least represented element, with only one member in MsWOX11 (Fig. 7, Supplementary Table 2).

**Tissue-specific expression analysis**

The heatmap revealed that MsWOX12, MsWOX19, MsWOX24, and MsWOX25 had nearly identical expression patterns in all tissues. Their expressions varied between –0.5 and 0.5. Some gene expressions were found to be tissue-specific. Flowers, for example, had the highest representation of highly expressed genes among the organs, while the leaf, with just MsWOX32 strongly expressed, had the lowest number of genes. In each tissue, just a few genes were expressed. MsWOX1, MsWOX3, MsWOX8, MsWOX7, MsWOX8, MsWOX13, MsWOX14, and MsWOX31 were all found in significant concentrations in the pre-elongated stem. The leaf, on the other hand, has only one highly expressed gene which is MsWOX32. These tissue expression patterns particular to subfamilies may be linked to gene activities. Heat maps highlighted the expression patterns of the paralogous pairings, and we discovered...
that most of the paralogous pairs with high sequence similarity had comparable expression patterns (Fig. 8).

To verify and enrich the expression profiles of MsWOX genes, real-time reverse transcription-PCR (qRT-PCR) analysis of 28 selected genes was conducted in five different tissues. The gene expression pattern detected by qRT-PCR was generally consistent with the RNA-seq data. MsWOX20 and MsWOX22 expression, for example, was both increased in the roots. Furthermore, these genes were intimately involved in a variety of organ-specific developmental processes (Fig. 9).

**Discussion**

As sequencing technology improves, so does the efficiency and accuracy of sequencing. As a result, more and more species’ genomic data are being published. This provides data that may be used to investigate gene structure and function prediction in the context of whole-genome comparison and identification. Here, putatively, 34 MsWOX genes from alfalfa species were identified and analyzed in this study. Phylogenetic analysis using two different methods revealed similar topologies, which were endorsed further by exon-intron organization analysis, motif assessment, and functional prediction. The alfalfa WOX proteins, along with *Arabidopsis* WOX proteins, were categorized into four well-organized groups. Genes from the same clade could have similar roles. For example, WUS-box in WOX11 and WOX12 are both involved in root development under abiotic stress in Brassica [25].

This study’s findings, combined with those from previous ones,
Fig. 5. Chromosomal locations of MsWOX genes in alfalfa. The number of chromosomes was labeled on the top of each chromosome. The location of each WUSCHEL-related homeobox (WOX) genes was marked on the chromosome. The left pane is the chromosome size.
suggest that the WOX genes in the Medicago species are highly conserved in structure and function in plant development and stress resistance. Unique features in some WOX members are also visible, which is due in part to the extensive expansion of some subgroups via gene duplication, as well as gene loss in some other subgroups following species divergence. A conserved homologous domain of 61 amino acid residues was discovered in all 34 WOX transcription factor family members in alfalfa, according to protein structure analysis. When compared to other species’ WOX proteins, we see a “helix-loop-helix-turn-helix” homeodomain structure. This domain is essential for DNA recognition and binding [26]. The homeodomain’s most conserved amino acids are the final amino acid at the second helix, the last amino acid residue of the turn structure, and the amino acid residue V of the middle position of the second helix structure [27]. According to our observation, the three amino acid residues are in the interior of the homeodomain, indicating that they may play an important transcription role as previously suggested [28].

Moreover, the angle of formation of “helix-turn-helix” structure in ancient branches is smaller than that of intermediate clades and modern clades, which may lead to functional changes in the evolution process of WOX transcription factor. The residues may correlate and play a role in the evolutionary process [3]. In addition, the WUX-box with LRP domain might play important role in the adventitious shoot organogenesis [29]. We also observe a conserved motif in the C-terminal region through multiple sequence alignment; other studies, accordingly, have found that the IC-WOX domain may be involved in root evolution [30].

As the binding sites of transcription factors, cis-acting elements in the promoter of the gene determine its expression patterns [31].

**Fig. 6.** Chart depicting the collinear relationships of MsWOX genes in alfalfa. The inner-colored lines show syntenic blocks in homoeologous chromosomes among alfalfa WUSCHEL-related homeobox (WOX) genes.
our study, a series of pant development-and stress-related cis-acting elements were detected in the promoter of MsWOXs. The maximum number of cis-acting elements were MYBs and ERF elements, which play important role in plant growth, development, and response [32, 33].

Whole-genome duplication (WGD) has been proven in several studies to impact the number of gene families in the genome [34]. WGD events in the alfalfa plant genome may have influenced the amount of MsWOX members. The old lineage of WOX genes, for example, is represented by group 1. Through multiplication, we see expansion events in WOX9 (WOX9A and WOX9B) and WOX5 (WOX5A and WOX5B). In group 2, duplicated Arabidopsis WOX13 members were found together with WOX9 duplicated members, whereas the other WOX5 was found on a separate branch of the phylogenetic tree. The fast evolution of WOX genes is suggested by these gene expansions. Gene duplication may have happened after the alfalfa species was separated from other plants, resulting in the isolated members [15]. The protein motifs and exon-intron organization of the isolated group of WOX9 members were also discovered to be distinct. It was revealed that the separated members had bigger gene sizes and longer introns than the original members. Furthermore, all isolated WOX13 subgroup members from alfalfa species lacked motif 3, which was present in all other WOX13 subgroup members. Surprisingly, isolated WOX9 binds more MYBs and ERF transcription factors, according to a cis-element analysis. This is consistent with the discovery
that the MYB and ERF families have developed swiftly and selectively in response to diverse environmental pressures [35-37].

In addition, both MsWOX9 and MsWOX13 members have a highly preserved motif layout. When compared to ancient and contemporary clade subgroups, the gene architectures of these two subgroups have the largest variation, with exon counts ranging from 3–5 in the WOX9 subgroup genes and 2–4 in the MsWOX11 subgroup genes. Previously, it was discovered that genes in the enlarged MsWOX5 play an important role in drought tolerance in the Jatropha plant [38]. Different functions of MsWOX9 proteins from other species have also been discovered, confirming this theory. Lie et al. [39] found that Arabidopsis WOX9/STIP in combination with WOX8/STPL is necessary for embryo patterning and vegetative SAM maintenance, whereas Petunia EVERGREEN/WOX9 is required for inflorescence growth and architecture [40]. All of this demonstrated that the intermediate clade was rapidly evolving, with significant changes in gene architecture, expression patterns, and likely gene functions among its members, which may be influenced by environmental factors like desiccation.

Except for MsWOX23, all WOX members of the current lineage have a homeodomain and a WUS box. The gene lacks the second exon which harbors the WUS box. The WUS box has been identified as the AtWUS functional domain essential for SAM cell identity induction and maintenance [41]. Also, the motif organization and gene architectures, as well as the functional prediction of current clade members within the same subgroup, were found to be identical, except for MsWOX23, which has no homologue in alfalfa species. Therefore, despite being evolutionary like MsWOX11 and MsWOX12, its function may be different due to the absence of the essential WUS box.

In conclusion, we identified 34 WOX genes in alfalfa using bioinformatics methods and a genome-wide database. According to structural features, the MsWOX genes are classified into four groups (group 1, group 2, group 3, and group 4) with 20 MsWOX genes belonging to group 1, seven in group 2, four in group 3, and three in group 4. According to chromosomal mapping, all the MsWOX genes were distributed on eight chromosomes. Collinearity analysis of MsWOX genes indicated considerable collinearity in 81.3% of alfalfa WOX genes. Numerous WOX genes are implicated in multiple gene duplication events, according to the complicated linear connection. Also, some WOX genes in alfalfa had no introns, while most MsWOX genes in the same subgroup exhibited comparable patterns of exon length, intron number, and conserved motifs, according to structural analysis. Cis elements and transcription factor binding analysis of the MsWOX gene revealed high abundance of MYBs, ERFs, and AP2 transcription factors whose number varied widely per motif. TCP was among the least. These findings indicate that MsWOX genes may have a role in development under abiotic stress.

**ORCID**

Tianhui Yang: https://orcid.org/0000-0002-3211-9781
Ting Gao: https://orcid.org/0000-0002-3738-7988
Chuang Wang: https://orcid.org/0000-0001-7492-1863
Xiaochun Wang: https://orcid.org/0000-0001-6024-3678
Caijin Chen: https://orcid.org/0000-0001-5472-570X
Mei Tian: https://orcid.org/0000-0003-0634-7964
Fig. 9. Expression analysis of MsWOX genes in different tissues by qRT-PCR. The expression level of each gene was calculated relatively to the average biological replicate of sample which was expressed at the lowest level and converted to the log base 10 of the value. Different letters indicate statistically significant differences when analyzed by One-way ANOVA and a multiple comparison using Tukey's test at p ≤ 0.05.
Weidi Yang: https://orcid.org/0000-0002-9281-1382

Authors’ Contribution

Conceptualization: TY, TG. Data curation: CW. Formal analysis: TY, XW. Funding acquisition: TG. Methodology: TY, CC, MT, WY. Writing - original draft: TY, TG. Writing - review & editing: TY, TG, CW, XW, CC, MT, WY.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This study was funded by Major breeding project in Ningxia Hui Autonomous Region (2019NYYZ03) and Natural Science Foundation of Ningxia Hui Autonomous Region (2021AAC03286).

Supplementary Materials

Supplementary data can be found with this article online at http://www.genominfo.org.

References

21. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic
Integrated bioinformatics analysis of validated and circulating miRNAs in ovarian cancer

Berkcan Dogan\textsuperscript{1,2}, Ece Gumusoglu\textsuperscript{3}, Ege Ulgen\textsuperscript{4}, Osman Ugur Sezerman\textsuperscript{4}, Tuba Gunel\textsuperscript{3*}

\textsuperscript{1}Department of Medical Genetics, Faculty of Medicine, Bursa Uludag University, Bursa 16059, Turkey
\textsuperscript{2}Department of Translational Medicine, Institute of Health Sciences, Bursa Uludag University, Bursa 16059, Turkey
\textsuperscript{3}Department of Molecular Biology and Genetics, Faculty of Science, Istanbul University, Istanbul 34134, Turkey
\textsuperscript{4}Department of Biostatistics and Medical Informatics, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul 34750, Turkey

Keywords: integrative analysis, KEGG pathway, microRNAs, ovarian cancer, pathfindR

Introduction

Despite the numerous research and clinical studies, ovarian cancer (OC) still has a high mortality rate among gynecological cancers since lack of effective biomarkers for early detection and prognosis \cite{1-3}. Also, the treatment efficiency of OC is low depending on multiple challenges such as late diagnosis, lack of reliable markers, the development of resistance to current therapeutics, and phenotype of heterogeneity \cite{1,4}. For these chal-
Challenges of OC, new studies need to reveal underlying molecular mechanisms, and to discover molecular biomarkers for early diagnosis, prevention, and targeted therapy [4].

MicroRNAs (miRNAs) are small non-coding RNAs that are about ~22 nucleotides in length. Their function is transcriptional and post-transcriptional regulation of gene expression by targeting mRNAs [5,6]. According to the effect on cancer, there are two types of miRNAs which are tumor suppressor miRNAs and oncomiRs. Depending on the type of cancer, oncomiRs or tumor suppressor miRNAs are inhibited or stimulated, respectively [7]. Especially, dysregulations of specific miRNAs affect cancer cell proliferation, differentiation, metastasis, and recurrence formation [8,9]. Various miRNAs have been shown to play different roles in OC. The mechanisms of miRNAs to impact OC and signaling pathways are still unknown. [2,10].

Many experimental studies have verified the detected interactions with bioinformatics analysis and proved the accuracy and predictivity with in-silico tools or databases. Although significant progress has been made about in-silico analysis in evaluating miRNAs, the need for new tools/databases is increasing day by day. It is also because of the limitations in existing tools/databases, that has increased with the development of high-throughput miRNAs technologies to analyze miRNA [11,12].

In this study, pathfindR was used for enrichment analysis of target genes. PathfindR uses active subnetworks, where an active subnetwork can be defined as a subnetwork of interconnected genes in a protein-protein interaction network (PIN), predominantly consisting of significantly altered genes. The tool initially maps the input genes with significance values onto the PIN and identifies active subnetworks, then it performs enrichment analysis on the identified subnetwork gene sets. In general, enrichment approaches overlook the relational information captured in the PIN and the genes neighboring the significant genes are not considered. By identifying active subnetworks, pathfindR exploits interaction information to enhance enrichment analysis. Active subnetworks allow the inclusion of possibly relevant genes that are not significant but connect significant genes in the PIN, and, in turn, the identification of phenotype-associated connected significant subnetworks [13]. This aids pathfindR uncover relevant mechanisms underlying the studied disease/phenotype.

In this study, we developed an in-silico approach to evaluate target genes of OC-related circulating and previously validated miRNAs which may be targeted for therapeutic approaches and utilized for OC management and diagnosis (Fig. 1).

**Methods**

**Identification of the targets genes of validated miRNAs**

In our functional analyses, we used hsa-miR-885-5p, hsa-miR-1909-5p, and hsa-let7d-3p which were previously defined by our group as the dysregulated miRNAs that can be a candidate biomarker for OC. These candidate miRNAs, which were determined by microarray, were validated by quantitative polymerase chain reaction (qPCR). Both microarray and qPCR results showed that these three miRNAs were downregulated in the OC group compared with healthy individuals [14,15].

Target genes of validated circulating miRNAs were examined by using the miRDB online database (http://www.mirdb.org). As one of the miRNA-target predictions and the functional annotations databases, miRDB provides access to miRNA-target genes and functions of five different species: human, mouse, rat, dog, chicken. All targets in the database were acquired from the MirTarget

![Fig. 1. The workflow for bioinformatics analysis of miRNAs. GEPIA, Gene Expression Profiling Interactive Analysis; qPCR, quantitative polymerase chain reaction.](https://doi.org/10.5808/gi.21067)
database, which was a bioinformatics tool developed with analyzing thousands of miRNA-target interactions obtained from high-throughput techniques. By integrating target prediction and gene ontology enrichment analyses, miRDB presents a streamlined pipeline for quickly identifying miRNA functions [16].

**Functional and pathway enrichment analyses**

Using all (i.e., union of) target genes of the validated miRNAs, active-subnetwork-oriented Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was performed using pathfindR. The tool pathfindR identifies gene sets that form active subnetworks in a PIN using a list of genes. Afterwards, it performs pathway enrichment analysis. An active subnetwork can be defined as a group of interconnected genes in a PIN that predominantly consists of significantly altered genes. Active subnetworks define distinct disease-associated sets of interacting genes. By incorporating interaction information, pathfindR yields more relevant enrichment results.

For assigning a significance value for each target gene (for use with pathfindR), initially, all *Homo sapiens* miRNA-target gene scores were obtained from miRDB (v6.0). The significance for each target gene was defined as the probability of observing a score greater than or equal to the score of this target gene over all *H. sapiens* miRNA-target gene scores [i.e., $P(x \geq \text{observed score})$]. For genes that are targeted by more than one miRNA, the lowest significance was kept. The final list of target genes-significance values was then, filtered keeping genes with significance $\leq 0.5$ (corresponds to a score of 67). The significance value used for pathfindR indicates what proportion of all scores (across all *H. sapiens* miRNA-target gene scores) was as high or higher than the observed score for a given miRNA-target gene pair. The threshold proxy significance value of 0.5 was an ad-hoc choice corresponding to a score of 67. The significance value used for pathfindR yields more relevant enrichment results.

**Overall survival and pathological stage analyses**

The associations between expression signatures of shared genes and overall survival and pathological stage analyses were performed in the TCGA and GTEx dataset by the GEPIA platform. For overall survival plot analysis used log-rank test, also known as the Mantel-Cox test, for the hypothesis test and cohorts’ thresholds adjusted. The Cox proportional hazard ratio and the 95% confidence interval information are included in the survival plot. The method for differentially expressed target gene analysis is one-way ANOVA, using the pathological stage as variable for calculating differential expression. The expression data are first log2(TPM+1) transformed for differential analysis. They included further statistical analysis, Benjamini and Hochberg’s false discovery rate (FDR) adjusted p-value (q-value) $< 0.05$ was identified as statistically significant. Genes with FDR p-value (q-value) $< 0.05$ were identified as significant association with overall survival or pathological stage.

**Results**

**Target genes analysis of miRNAs**

Target genes of validated circulating miRNAs were comprehensively analyzed by miRDB database. Target genes were retrieved separately determined for each miRNA. No specific filters were applied for target gene prediction in miRDB database. Respectively, 422, 230, and 44 target genes were identified with mining target genes of hsa-miR-885-5p, hsa-miR-1909-5p, and hsa-let-7d-3p. Target genes were filtered by a score of 67 for significance values. After filtering, the number of target genes detected for hsa-miR-885-5p, hsa-miR-1909-5p, and hsa-let-7d-3p were respectively 229, 90, and 23 (Table 1). While two common target genes were found in hsa-miR-885-5p and hsa-miR-1909-5p: ERICH3 (glutamate rich 3) and CAPRIN1 (cell cycle associated protein 1), one common target gene were observed in hsa-miR-885-5p and hsa-let-7d-3p: SEC24D (SEC24 homolog D, COPII coat complex component). No common target gene was found for all three miRNAs. The numbers of target and common genes are shown in Table 1.
Table 1: Target genes determined in the miRDB database (score ≥ 67, for representation purposes)

<table>
<thead>
<tr>
<th>miRNAs</th>
<th>Target genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-miR-885-5p</td>
<td>GALNT3, ZNF367, C9orf3, TMEM135, CTNNB1, ZNF354A, OXR1, ZFP91, LRPP4, CPEB2, SYT4L, SLC11A2, ZNF281, PGK2, ASC14, CAMK4, FAM81A, KIF21A, MXD1, LCP1, ATXN11, C9orf78, Ccdc71L, UPF2, ADD1, FAM241A, GPR68, MAN1C1, ELAVL1, ZADH2, PANK3, PFDN4, JAZF1, ORS1E1, ECE1, FURIN, CDC73, FNTA, MAP4K3, UBDT2, LACTB, ATAD2, ENKUR, TMEM41A, FAM98A, KAT2B, EMLIN2, CPSF6, NLRP1, KIAA0825, OSBPL6, ZFY, MAGT1, ZBTB34, NUDT21, CEP170, CPEB3, ING3, SAPC3, ARID2, PDE7A, NUDCD2, TMMR9, ASB5, ZC3H12C, MAN1A2, PCKS5, WWPI, PPARA, DPH19L1, TDRD6, PDCD6IP, VSTM2A, PRK2, ZC3H1A, SH2D1A, HSF2, LAMC2, HSF1, MACF1, WNT5A, VPS13A, GTF2H3, ZNF407, MXE30D, TMEEF2, PHF6, TSC22D2, SLC33A1, SRSF6, DAPLC2, STRIP2, EIF1AX, EIF1AY, SMIM13, TMTT, RX7, CEGBP, ZNF10, PARVA, NSUN5, SEMAGA, UBE2N, CXADR, CKD17, TMEM185B, TCTEX1D1, PDK1, CPSF7, SL-C3501, ZNF516, PARQ, RAD54B, RET, NME5, SLC35F5, HMBOX1, SAMD12, FSBP, Ccdc182, SMARCE1, CLASP2, LRRK2, U2LRA1, B3GLCT, TMEM41B, YWWAE, IMPACT, HSBI1, PPP3R3, HINT3, FOSD1L, KIF1C, MRPL2, SH3BGR2L, TCC3A, ANKFY1, CDK14, ERIC3, ZFHX4, HMGB1, ZFX, TMPRSS11A, UBE2K, NR2C1, TMEM230, PRK3, MRPS10, KLHL29, SRMF2, ETFBKM, GOPC, HRASLS5, AIG1, SLC32A8, PAK5, NUP50, RBPM5, GPR75, DGKH, LARP1, CALN1, ZNF80, ANK54B, STMNR2, NBSA2, BMP3, EYA1, CREB2, SEC24D, TRIM7, LRRC40, TRAF6, U2SURP, EZR, IL6ST, VPS41, JAKMIP3, FREM2, CCR9, SYL1, RPRGIP1, RBBP4, CAPR1N, MYT1, EYA3, POU4F2, RPS9, CTNNA3, FZD10, HGSNAT, ARPP21, TNRC6C, REL, ERLIN, SEMA6D, KHORBS2, RMM14, TGFBR1, AP2M1, MYSM1, WDR43, URI1, TBCC, TVP23B, USF3, GNB4, ABCB8, CENPA, USP14, CLUS2, CD59, TOX, PLEKHA5, C1orox25, OPAL, ABHD17C, SOAT1, UBN2, YWHAH, Ccdc135, I6L8, FAM12A6, DUCN1D5, MIER3, RHOD, DSTE, HIPK3, MDM4</td>
</tr>
<tr>
<td>hsa-miR-1909-5p</td>
<td>F7, PTP4N, RBCTB2, MEFC2F, TAF5, TSAPN16, Ccdc113, COX11, SHOC2, KBTBD12, BCO2, KIF3B, TIGD2, POM121, ERICH3, BLOC1S5, MM20, RAPGEFL1, YY1, B4QALT1, AEBP2, BAZ1A, PLCD2, C2orf48, C2orf19, SERF2, BOI, MOB4, ARID4A, ARID4L, ZNF275, CREB5, PIG2, TBCD10A, FAM83E, ERB3B, EIF4E2, YES1, GLP2R, METTL7B, IL1R2, RPRG, COL8A1, NPK1, CKD15, GGA2, COP7R, KSCAN, SNTB2, MFHAS1, CEMP1, GTPBP1, T1QO, CNGB3, GMPR2, ARPIN-AP352, MEOX2, ABHD2, AP352, APC, ZNF687, YAE1, STMNR1, SCAMP5, Dicer1, FCR1L, RAB36, GSCD, NOL12, FMC1, TOFA, CNKSR2, UBA7, ECHDC2, TCC16, PTPA, SLC35E4, BA1P, CDKN8B, PEX9, DAPL1, EIF4EBP2, CHST9, KCTD16, MPE1A, GDF6, CAPR1N, RIP1L1, RIKM2A, ARL8B, STX5</td>
</tr>
<tr>
<td>hsa-let-7d-3p</td>
<td>MEX3C, SEC24D, BEND2, PI53, PARP11, HMGA2, SH3RF1, PRKACB, GSTK1, PHF14, EXOC7, SIGLEC6, CYP4Z1, KANSL1, NEM1, ABCA8, ZNF486, CASD1, TMEM417, CAMKK2, RLI2, TIMMDC1, KLHL31</td>
</tr>
</tbody>
</table>

Pathway enrichment analyses of target genes of miRNAs

Performing active-subnetwork-oriented enrichment analysis via pathfindR using the union of miRNA-target genes (339 genes in total), 24 enriched KEGG pathways were identified (Table 2). The top three pathways were ubiquitin-mediated proteolysis, spliceosome, and Notch signaling pathway according to p-values. The pathway diagrams for these pathways, in which the target genes of our dysregulated miRNAs are indicated by orange color, are presented in Fig. 3. It is found that a total of ten different target genes were involved in these pathways. A part of the target genes which are TRAF6 (TNF receptor associated factor 6), UBE2N (ubiquitin conjugating enzyme E2 N), UBA7 (ubiquitin like modifier activating enzyme 7), UBE2K (ubiquitin conjugating enzyme E2 K), WWPI (WW domain containing E3 ubiquitin protein ligase 1) are ubiquitin-mediated proteolysis (Fig. 3A) whereas, SRSF2 (serine and arginine rich splicing factor 2), SRSF6 (serine and arginine rich splicing factor 6), and U2SURP (U2 SnRNP associated SURP domain containing) are involved in spliceosome pathway (Fig. 3B). There are only two genes in the Notch pathway: ATXN1L (ataxin 1
like) and KAT2B (lysine acetyltransferase 2B) (Fig. 3C). While nine of these ten genes are (UBE2N, UBE2K, WWP1, TRAF6, U2SURP, SRSF2, SRSF6 KAT2B, and ATXN1L) targeted by hsa-miR-885-5p, only the UBA7 gene is targeted by hsa-miR-1909-5p. There are no target genes of hsa-let-7d-3p associated within three pathways.

### Analysis of differentially expressed target genes

Differentially expressed target gene analysis was conducted to examine the expressions of identified target genes of miRNAs in the TCGA and GTEx datasets. In union of miRNA-target genes (339 genes in total), 93 genes were found to be differentially expressed (p < 0.05) in TCGA and GTEx datasets of when compared to cancer tissues with paired normal tissues (Table 3). None of the common genes (ERICH, CAPRIN1, and SEC24D) among target genes were significant in TCGA and GTEx datasets. While miR-885-5p targets 64 out of 93 genes, miR-1909-5p regulates 20 and let-7d-3p 9 target genes.

### Overall survival and pathological stage analyses

The associations with overall survival outcomes and pathological stage analysis for expression signatures of 93 genes in the TCGA and GTEx datasets were performed by the GEPIA (Table 3). After FDR tests, no significant genes were found to be significant in the analysis of overall survival analyses, but 24 genes were found to be significant with pathological stages analyses (p < 0.05). No significant differentially expressed target genes were found to be with overall survival after statistical analyses. However, 24 differentially expressed target genes were significantly associated pathological stage of OC (Fig. 4). Of 24 genes, ZNF407 (zinc finger protein 407) and UBN2 (ubinuclein 2) genes had the lowest p-value for their relationship with pathological stages for the TCGA and GTEx datasets (p = 0.017856) (Table 3).

### Table 2. Enriched KEGG pathways identified via pathfindR analysis

<table>
<thead>
<tr>
<th>ID</th>
<th>Pathway</th>
<th>p-value</th>
<th>Target genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa04120</td>
<td>Ubiquitin-mediated proteolysis</td>
<td>&lt; 0.001</td>
<td>UBA7, UBE2N, UBE2K, WWP1, TRAF6</td>
</tr>
<tr>
<td>hsa03040</td>
<td>Spliceosome</td>
<td>&lt; 0.001</td>
<td>U2SURP, SRSF2, SRSF6</td>
</tr>
<tr>
<td>hsa04330</td>
<td>Notch signaling pathway</td>
<td>&lt; 0.001</td>
<td>KAT2B, ATXN1L</td>
</tr>
<tr>
<td>hsa05205</td>
<td>Proteoglycans in cancer</td>
<td>0.001160427</td>
<td>ERBB3, CTNNB1, FZD10, EZR, WNT5A, PRKACB</td>
</tr>
<tr>
<td>hsa04390</td>
<td>Hippo signaling pathway</td>
<td>0.001309875</td>
<td>TGFR1, GDF6, WNT5A, FZD10, YWHAH, YWHAH, CTNNB1, APC, CTNNA3</td>
</tr>
<tr>
<td>hsa05160</td>
<td>Hepatitis C</td>
<td>0.001416289</td>
<td>TRAF6, YWHAH, YWHAH, CTNNB1</td>
</tr>
<tr>
<td>hsa04657</td>
<td>IL-17 signaling pathway</td>
<td>0.002692059</td>
<td>TRAF6, ELAVL1</td>
</tr>
<tr>
<td>hsa04144</td>
<td>Endocytosis</td>
<td>0.003025769</td>
<td>AP2M1, TGFBR1, TRAF6, WWP1, PDCD6IP</td>
</tr>
<tr>
<td>hsa03015</td>
<td>mRNA surveillance pathway</td>
<td>0.003942335</td>
<td>NUDT21, CPSF6, CPSF7, UPF2</td>
</tr>
<tr>
<td>hsa05225</td>
<td>Hepatocellular carcinoma</td>
<td>0.008091156</td>
<td>TGFα, TGFBR1, WNT5A, FZD10, AP, CTNNB1, SMARCE1, ARID2</td>
</tr>
<tr>
<td>hsa04670</td>
<td>Leukocyte transendothelial migration</td>
<td>0.008953354</td>
<td>ESR, CTNNB1, CTNNA3</td>
</tr>
<tr>
<td>hsa03013</td>
<td>RNA transport</td>
<td>0.0091471</td>
<td>NUP50, POM121, EIF1AY, EIF1AX, EIF4E2, EIF4EBP2, UPF2</td>
</tr>
<tr>
<td>hsa05203</td>
<td>Viral carcinogenesis</td>
<td>0.011431025</td>
<td>CREBS, YWHAE, YWHAH, REL, PRKACB, GTF2H3, CDK1B1, KAT2B, IL6ST</td>
</tr>
<tr>
<td>hsa04919</td>
<td>Thyroid hormone signaling pathway</td>
<td>0.014664218</td>
<td>PRKACB, KAT2B, CTNNB1</td>
</tr>
<tr>
<td>hsa04520</td>
<td>Adherens junction</td>
<td>0.015121939</td>
<td>CTNNB1, CTNNA3, YES1, TGFBR1</td>
</tr>
<tr>
<td>hsa05226</td>
<td>Gastric cancer</td>
<td>0.016356287</td>
<td>WNT5A, FZD10, AP, CTNNB1, CDK1B, TGFBR1, CTNNA3</td>
</tr>
<tr>
<td>hsa05230</td>
<td>Central carbon metabolism in cancer</td>
<td>0.017277363</td>
<td>PDK1, RET</td>
</tr>
<tr>
<td>hsa04550</td>
<td>Signaling pathways regulating pluripotency of stem cells</td>
<td>0.017619136</td>
<td>IL6ST, WNT5A, FZD10, AP, CTNNB1</td>
</tr>
<tr>
<td>hsa04621</td>
<td>NOD-like receptor signaling pathway</td>
<td>0.018069616</td>
<td>TRAF6, NLRP1, YWHAE</td>
</tr>
<tr>
<td>hsa04934</td>
<td>Cushing syndrome</td>
<td>0.030370611</td>
<td>PRKACB, CREBS, CDK1B, WNT5A, FZD10, CTNNB1, AP</td>
</tr>
<tr>
<td>hsa04310</td>
<td>Wnt signaling pathway</td>
<td>0.033429078</td>
<td>WNT5A, FZD10, CTNNB1, AP, PRKACB, PPP3R1</td>
</tr>
<tr>
<td>hsa05131</td>
<td>Shigellosis</td>
<td>0.034316691</td>
<td>TRAF6, UBE2N</td>
</tr>
<tr>
<td>hsa05216</td>
<td>Thyroid cancer</td>
<td>0.040790494</td>
<td>RET, CTNNB1</td>
</tr>
<tr>
<td>hsa05167</td>
<td>Kaposi sarcoma-associated herpesvirus infection</td>
<td>0.045496474</td>
<td>GNB4, PPP3R1, IL6ST, CTNNB1</td>
</tr>
</tbody>
</table>

ID indicates the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway ID. The pathway column indicates the name of pathways. The p-value column indicates the lowest enrichment p-value obtained through multiple iterations. The target genes column indicates the target genes (that are the targets of at least one miRNA) involved in the given enriched pathway.
Discussion

Recently, a considerable number of research have been made to determine the effect of miRNAs regulate cancer hallmarks and to develop the early diagnosis and prognosis of cancer [18]. Nevertheless, several challenges such as various sampling methods, sample size, detection techniques, gender, and ethnicity or genetic background, affect the reliance on utilizing miRNAs as biomarkers [1,3,10,19]. One of the fundamental research areas for easily detectable, non-invasive, sensitive, and specific miRNA-based biomarker discoveries are of a great value for accurate and effective early diagnosis, risk prediction, prognosis, recurrence, and effective management of OC [8,10,12].

Identification of miRNAs’ target genes has been the focus of computational biology for the last few years. Since detecting all possible miRNA targets with high-throughput technologies is laborious and costly, a wide variety of computational resources has been developed. The methodologies used by databases range from evolutionary conservation evaluations of putative miRNA binding sites to machine learning and classification algorithms. Continuous improvement is needed to develop new tools/databases to accurate predictions of miRNA targets [20,21]. One of these databases is miRDB online database and determines miRNA-target prediction and functional annotations. Estimating that 3.5 million target genes were regulated by 7,000 miRNAs in human, mouse, rat, dog, and chicken, in the latest version of miRDB (v6.0) major updated in 2019. In miRDB, miRNA binding and target down-regulation features used to predict miRNA targets with machine learning methods, generates a prediction scores are in the range of 0–100, and candidate genes with scores ≥ 50 are presented as predicted miRNA targets [16].

The pathfindR tool was used for the enrichment analysis of target genes within the scope of the study. While pathfindR allows better identification of disease-related pathways, it should be noted that the tool requires a significance level per each input gene. To overcome this limitation, we followed an approach where we used...
<table>
<thead>
<tr>
<th>Circulating validated miRNAs</th>
<th>miRDB score</th>
<th>Differential expression analysis</th>
<th>Overall survival analysis</th>
<th>Pathological stage analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hsa-miR-885-5p</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GALNT3</td>
<td>100</td>
<td>2.444</td>
<td>2.18E-17</td>
<td>0.99</td>
</tr>
<tr>
<td>LRP4</td>
<td>95</td>
<td>-1.867</td>
<td>0.86</td>
<td>0.99</td>
</tr>
<tr>
<td>SYTL4</td>
<td>95</td>
<td>-4.934</td>
<td>0.96</td>
<td>0.99</td>
</tr>
<tr>
<td>LCP1</td>
<td>93</td>
<td>2.188</td>
<td>0.83</td>
<td>0.99</td>
</tr>
<tr>
<td>CbofR8B</td>
<td>92</td>
<td>-2.239</td>
<td>0.059</td>
<td>0.6186666666666667</td>
</tr>
<tr>
<td>ADD1</td>
<td>92</td>
<td>-1.492</td>
<td>0.85</td>
<td>0.99</td>
</tr>
<tr>
<td>MAN1C1</td>
<td>91</td>
<td>-2.329</td>
<td>0.053</td>
<td>0.6186666666666667</td>
</tr>
<tr>
<td>ECE1</td>
<td>90</td>
<td>-1.596</td>
<td>0.0067</td>
<td>0.4288</td>
</tr>
<tr>
<td>MAP4K3</td>
<td>89</td>
<td>-1.158</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>ATAD2B</td>
<td>89</td>
<td>-1.679</td>
<td>0.37</td>
<td>0.99</td>
</tr>
<tr>
<td>KAT2B</td>
<td>88</td>
<td>-1.820</td>
<td>0.087</td>
<td>0.6186666666666667</td>
</tr>
<tr>
<td>EMLIN2</td>
<td>88</td>
<td>-1.063</td>
<td>0.88</td>
<td>0.99</td>
</tr>
<tr>
<td>NLRP1</td>
<td>88</td>
<td>-3.098</td>
<td>0.91</td>
<td>0.99</td>
</tr>
<tr>
<td>CEP170</td>
<td>87</td>
<td>-1.235</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>PDE10A</td>
<td>86</td>
<td>-2.166</td>
<td>0.51</td>
<td>0.99</td>
</tr>
<tr>
<td>ZC3H12C</td>
<td>86</td>
<td>-1.207</td>
<td>0.31</td>
<td>0.99</td>
</tr>
<tr>
<td>PCSK5</td>
<td>86</td>
<td>-2.428</td>
<td>0.76</td>
<td>0.99</td>
</tr>
<tr>
<td>LAMC2</td>
<td>84</td>
<td>3.724</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>MACF1</td>
<td>84</td>
<td>-1.498</td>
<td>0.072</td>
<td>0.6186666666666667</td>
</tr>
<tr>
<td>ZNF407</td>
<td>84</td>
<td>-1.085</td>
<td>0.033</td>
<td>0.6186666666666667</td>
</tr>
<tr>
<td>PHF6</td>
<td>84</td>
<td>-1.270</td>
<td>0.69</td>
<td>0.99</td>
</tr>
<tr>
<td>SRSF6</td>
<td>83</td>
<td>-1.279</td>
<td>0.83</td>
<td>0.99</td>
</tr>
<tr>
<td>TPM1</td>
<td>83</td>
<td>1.519</td>
<td>0.33</td>
<td>0.99</td>
</tr>
<tr>
<td>CEBPD</td>
<td>83</td>
<td>-1.953</td>
<td>0.051</td>
<td>0.6186666666666667</td>
</tr>
<tr>
<td>ZNF10</td>
<td>82</td>
<td>2.257</td>
<td>0.8</td>
<td>0.99</td>
</tr>
<tr>
<td>SEMA6A</td>
<td>82</td>
<td>-2.468</td>
<td>0.27</td>
<td>0.99</td>
</tr>
<tr>
<td>UBE2N</td>
<td>82</td>
<td>1.075</td>
<td>0.92</td>
<td>0.99</td>
</tr>
<tr>
<td>CXADR</td>
<td>82</td>
<td>3.495</td>
<td>0.23</td>
<td>0.99</td>
</tr>
<tr>
<td>CDK17</td>
<td>81</td>
<td>-1.058</td>
<td>0.63</td>
<td>0.99</td>
</tr>
<tr>
<td>TMEM185B</td>
<td>81</td>
<td>1.771</td>
<td>0.46</td>
<td>0.99</td>
</tr>
<tr>
<td>ZNF516</td>
<td>81</td>
<td>-2.404</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>HMBOX1</td>
<td>80</td>
<td>-1.798</td>
<td>0.9</td>
<td>0.99</td>
</tr>
<tr>
<td>SAMD12</td>
<td>80</td>
<td>1.632</td>
<td>0.77</td>
<td>0.99</td>
</tr>
<tr>
<td>CLASP2</td>
<td>79</td>
<td>-1.691</td>
<td>0.6</td>
<td>0.99</td>
</tr>
<tr>
<td>LRRK2</td>
<td>79</td>
<td>2.551</td>
<td>0.72</td>
<td>0.99</td>
</tr>
<tr>
<td>LYPLA1</td>
<td>79</td>
<td>1.886</td>
<td>0.89</td>
<td>0.99</td>
</tr>
<tr>
<td>HSPB1</td>
<td>78</td>
<td>1.049</td>
<td>0.031</td>
<td>0.6186666666666667</td>
</tr>
<tr>
<td>TCT30A</td>
<td>77</td>
<td>1.336</td>
<td>0.61</td>
<td>0.99</td>
</tr>
<tr>
<td>ANKFY1</td>
<td>77</td>
<td>-1.215</td>
<td>0.47</td>
<td>0.99</td>
</tr>
<tr>
<td>UBE2K</td>
<td>77</td>
<td>1.396</td>
<td>0.3</td>
<td>0.99</td>
</tr>
<tr>
<td>NR2C1</td>
<td>77</td>
<td>-1.060</td>
<td>0.6</td>
<td>0.99</td>
</tr>
<tr>
<td>G0PC</td>
<td>76</td>
<td>-1.215</td>
<td>0.45</td>
<td>0.99</td>
</tr>
<tr>
<td>HRASLS5</td>
<td>76</td>
<td>-1.137</td>
<td>0.96</td>
<td>0.99</td>
</tr>
<tr>
<td>SLC38A2</td>
<td>76</td>
<td>-1.343</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>RBMS5</td>
<td>75</td>
<td>-2.988</td>
<td>0.62</td>
<td>0.99</td>
</tr>
<tr>
<td>DGKH</td>
<td>75</td>
<td>-1.085</td>
<td>0.78</td>
<td>0.99</td>
</tr>
</tbody>
</table>

(Continued to the next page)
**Table 3. Continued**

<table>
<thead>
<tr>
<th>Circulating validated miRNAs</th>
<th>Differentially expressed target genes</th>
<th>miRDB score (fold change)</th>
<th>overall survival analysis</th>
<th>Pathological stage analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Log2</td>
<td>Log-rank p-value</td>
<td>FDR</td>
</tr>
<tr>
<td>CREBF2</td>
<td>74</td>
<td>-2.524</td>
<td>0.87</td>
<td>0.99</td>
</tr>
<tr>
<td>TRIM7</td>
<td>74</td>
<td>-1.351</td>
<td>0.95</td>
<td>0.99</td>
</tr>
<tr>
<td>IL6ST</td>
<td>73</td>
<td>-1.746</td>
<td>0.61</td>
<td>0.99</td>
</tr>
<tr>
<td>VF541</td>
<td>73</td>
<td>-1.177</td>
<td>0.47</td>
<td>0.99</td>
</tr>
<tr>
<td>JAKMIP2</td>
<td>73</td>
<td>-1.807</td>
<td>0.77</td>
<td>0.99</td>
</tr>
<tr>
<td>FZD10</td>
<td>71</td>
<td>1.693</td>
<td>0.27</td>
<td>0.99</td>
</tr>
<tr>
<td>HGSNAT</td>
<td>71</td>
<td>-1.255</td>
<td>0.66</td>
<td>0.99</td>
</tr>
<tr>
<td>TNRC6C</td>
<td>71</td>
<td>-1.140</td>
<td>0.2</td>
<td>0.99</td>
</tr>
<tr>
<td>SEMA6D</td>
<td>70</td>
<td>-1.887</td>
<td>0.6</td>
<td>0.99</td>
</tr>
<tr>
<td>MYSM1</td>
<td>69</td>
<td>1.920</td>
<td>0.66</td>
<td>0.99</td>
</tr>
<tr>
<td>TBCC</td>
<td>69</td>
<td>1.468</td>
<td>0.08</td>
<td>0.618666</td>
</tr>
<tr>
<td>CENPK</td>
<td>69</td>
<td>1.782</td>
<td>0.7</td>
<td>0.99</td>
</tr>
<tr>
<td>PLEKHA5</td>
<td>68</td>
<td>-1.640</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>ABHD17C</td>
<td>68</td>
<td>2.353</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>UBN2</td>
<td>68</td>
<td>-1.158</td>
<td>0.57</td>
<td>0.99</td>
</tr>
<tr>
<td>FAM126A</td>
<td>67</td>
<td>-1.282</td>
<td>0.55</td>
<td>0.99</td>
</tr>
<tr>
<td>MIER3</td>
<td>67</td>
<td>-1.139</td>
<td>0.79</td>
<td>0.99</td>
</tr>
<tr>
<td>MODM4</td>
<td>67</td>
<td>-1.924</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Hsa-miR-1909-5p</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RCBT2B</td>
<td>95</td>
<td>-1.508</td>
<td>0.33</td>
<td>0.727272</td>
</tr>
<tr>
<td>MEF2C</td>
<td>95</td>
<td>-2.096</td>
<td>0.75</td>
<td>0.833333</td>
</tr>
<tr>
<td>BCO2</td>
<td>88</td>
<td>-1.333</td>
<td>0.4</td>
<td>0.727272</td>
</tr>
<tr>
<td>C4orf19</td>
<td>83</td>
<td>1.547</td>
<td>0.39</td>
<td>0.727272</td>
</tr>
<tr>
<td>ZNF275</td>
<td>81</td>
<td>-1.501</td>
<td>0.87</td>
<td>0.89</td>
</tr>
<tr>
<td>PIGZ</td>
<td>80</td>
<td>1.131</td>
<td>0.63</td>
<td>0.833333</td>
</tr>
<tr>
<td>ERBB3</td>
<td>80</td>
<td>4.024</td>
<td>0.13</td>
<td>0.727272</td>
</tr>
<tr>
<td>METTL7B</td>
<td>78</td>
<td>3.511</td>
<td>0.19</td>
<td>0.727272</td>
</tr>
<tr>
<td>KCNK15</td>
<td>77</td>
<td>3.713</td>
<td>0.27</td>
<td>0.727272</td>
</tr>
<tr>
<td>GGA2</td>
<td>76</td>
<td>-1.710</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>CEMP1</td>
<td>75</td>
<td>-1.950</td>
<td>0.32</td>
<td>0.727272</td>
</tr>
<tr>
<td>GTPBP1</td>
<td>74</td>
<td>1.18</td>
<td>0.44</td>
<td>0.733333</td>
</tr>
<tr>
<td>TRIO</td>
<td>74</td>
<td>-1.925</td>
<td>0.71</td>
<td>0.833333</td>
</tr>
<tr>
<td>MEKO2</td>
<td>73</td>
<td>-1.228</td>
<td>0.025</td>
<td>0.5</td>
</tr>
<tr>
<td>Dicer1</td>
<td>72</td>
<td>-1.709</td>
<td>0.68</td>
<td>0.833333</td>
</tr>
<tr>
<td>SGCD</td>
<td>72</td>
<td>-1.537</td>
<td>0.67</td>
<td>0.833333</td>
</tr>
<tr>
<td>TGFA</td>
<td>71</td>
<td>1.493</td>
<td>0.11</td>
<td>0.727272</td>
</tr>
<tr>
<td>UBA7</td>
<td>70</td>
<td>-1.552</td>
<td>0.17</td>
<td>0.727272</td>
</tr>
<tr>
<td>ECHDC2</td>
<td>70</td>
<td>-3.671</td>
<td>0.75</td>
<td>0.833333</td>
</tr>
<tr>
<td>CDKN1B</td>
<td>69</td>
<td>-1.130</td>
<td>0.3</td>
<td>0.727272</td>
</tr>
<tr>
<td>Hsa-let-7d-3p</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PTGIS</td>
<td>88</td>
<td>-2.804</td>
<td>0.18</td>
<td>0.54</td>
</tr>
<tr>
<td>HMG2A</td>
<td>86</td>
<td>1.875</td>
<td>0.073</td>
<td>0.3285</td>
</tr>
<tr>
<td>GSTK1</td>
<td>81</td>
<td>1.034</td>
<td>0.25</td>
<td>0.5625</td>
</tr>
<tr>
<td>KANS1</td>
<td>74</td>
<td>-1.149</td>
<td>0.47</td>
<td>0.74</td>
</tr>
<tr>
<td>NQO1</td>
<td>72</td>
<td>-1.178</td>
<td>0.73</td>
<td>0.74</td>
</tr>
<tr>
<td>ABG18</td>
<td>72</td>
<td>-5.852</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>CASD1</td>
<td>71</td>
<td>-1.808</td>
<td>0.65</td>
<td>0.74</td>
</tr>
<tr>
<td>RLO2</td>
<td>69</td>
<td>1.1</td>
<td>0.054</td>
<td>0.3285</td>
</tr>
<tr>
<td>TMMDC1</td>
<td>67</td>
<td>1.527</td>
<td>0.65</td>
<td>0.74</td>
</tr>
</tbody>
</table>

TCGA, The Cancer Genome Atlas; GTEx, Genotype-Tissue Expression; FDR, false discovery rate.
Fig. 4. Pathological stage analysis of differently expressed target genes in the ovarian cancer dataset using the GEPIA (Gene Expression Profiling Interactive Analysis) platform (p < 0.05). Pr(>F) values were adjusted to Benjamini and Hochberg’s false discovery rate test.
the scores for miRNA-target gene pairs to calculate a significance level. Additionally, some genes (usually a small proportion) that are not in the PIN (that do not have any curated interactions) have to be discarded, resulting in the loss of possibly relevant genes.

In our previous studies, three downregulated circulating miRNAs (hsa-miR-1909-5p, hsa-miR-885-5p, and hsa-let-7d-3p) in OC patients were consistently validated by comparison with healthy individuals. In these studies, circulating miRNAs, which were determined to be dysregulated by microarray, were then validated with the qPCR [14,15]. In this study, we predicted the validated circulating miRNAs’ target genes with miRDB database and performed functional analysis by pathfindR tool to understand the pathogenesis of OC especially the molecular mechanisms of its development. Also, overall survival and pathological stage analyses were evaluated with differentially miRNAs’ target genes which are commonly found in the TCGA and GTEx datasets.

Two hundred and twenty-nine, 90, and 23 of target genes of hsa-miR-1909-5p, hsa-miR-885-5p, and hsa-let-7d-3p were separately determined via miRDB (score ≥ 67), respectively (Fig. 2). Enriched KEGG pathways of target genes were performed with pathfindR tool. Totally, 339 genes included in active-subnetwork-oriented enrichment analysis and as a result, 24 enriched KEGG pathways were identified (Table 2). Top three pathways with the lowest p-values are ubiquitin-mediated proteolysis (p < 0.001), spliceosome (p < 0.001), and Notch signaling pathway (p < 0.001), respectively (Fig. 3). TRAF6, UBA7, UBE2K, UBE2N, and WWPI genes regulate the ubiquitin-mediated proteolysis pathway. While SRSF2, SRSF6, and U2SURP are involved in the spliceosome, ATXN1L and KAT2B are adjusted to the Notch pathway. Although hsa-miR-885-5p targets ATXN1L, KAT2B, SRSF2, SRSF6, TRAF6, UBE2K, UBE2N, U2SURP, and WWPI genes, hsa-miR-1909-5p just targets UBA7 gene. No target genes of hsa-let-7d-3p were found to be associated with these three pathways.

Our pathway analysis determined different pathways that can be related to the OC development. Of these pathways, only ubiquitin-mediated proteolysis, spliceosome, and Notch signaling pathway showed that their p-value is less than 0.001. Besides the effects of these pathways on OC, they also have roles on other cancer types. Firstly, ubiquitin-mediated proteolysis is an essential mechanism that is responsible for 80%–90% of intracellular protein degradation and is involved in many cellular processes, including tumorigenesis, tumor survival and apoptosis [22,23]. Ubiquitin-mediated pathway modulates BRCA1/2, p53, ERBB2 gene expressions, ERK pathway, cyclin-dependent cell cycle regulation which are related to OC [24]. Bazzaro et al. [25] claimed that upregulation of proteasome subunit levels occurs in OC and proteasome inhibitors may have utility in the treatment of OC. Secondly, splicing mechanism is a crucial process that regulates cellular proliferation, differentiation, and survival [26]. Disregulation of splicing processes and splicing factor genes contribute to cancer, including OC [27,28]. Especially, splicing machinery mutations especially contribute to tumorigenesis. Additionally, it is critical to understand that the molecular mechanism of RNA splicing is causing the development of drug resistance in cancer treatment [26]. Regulation of the relationship between splicing and cancer can be led to splicing-based therapies for cancer treatment [29]. Thirdly, Notch signaling pathway regulates not only cell self renewal and differentiation but also a cell to cell communication [30]. Upregulation of Notch signaling pathway proteins have been identified in OC [31]. Several studies have revealed that it is related to poor overall and disease free survival time, and more advanced stages [30,32]. The Notch signaling pathway plays a specific role in deregulation of signaling cascade has been associated with OC. Targeted therapy against Notch pathway activation can offer clinical benefit to OC [30,31].

According to these studies, our findings also indicate that our dysregulated miRNAs are significant in OC via pathway related-target genes. Briefly, circulating downregulated miRNAs could not suppress their target genes in these pathways, hence, have activated these pathways, resulting in the development of OC. As we can see from the previous studies, these pathways were found to be related with OC. We have discovered target genes of three downregulated circulating miRNAs (hsa-miR-1909-5p, hsa-miR-885-5p, and hsa-let-7d-3p) and also related-pathways of these genes. We propose novel mechanisms between miRNAs, target genes and OC that have not been elucidated previously using pathfindR. Our findings can be used as a diagnostic tool in OC. Our perspective on ATXN1L, KAT2B, SRSF2, SRSF6, TRAF6, UBA7, UBE2N, UBE2K, U2SURP, WWPI and in OC will promote more extensive research on the molecular mechanisms of hsa-miR-1909-5p, hsa-miR-885-5p, and hsa-let-7d-3p and provide a reference for improving the clinical outcome of OC.

The development of prognostic multigene classification protocols can benefit the understanding of tumor biology as well as the prediction of cancer progression and treatment strategies. One critical issue is determining the properly combining the genes [33]. However, studies on the overall survival-related profiles in OC patients have progressed, whereas there have been no large-scale studies based on multicenter validation of gene expression profiles for prediction of disease progression or recurrence in OC patients [34]. Furthermore, pathological staging, which can be determined after surgery and examination of the removed tumor tis-
In conclusion, we offer in-silico evidence that validated circulating miRNAs' target genes and enriched pathways are related to OC and have potential roles in theranostics applications. Especially, enrichment pathways and pathological stage-related genes can be combined with validated miRNAs, their multiple analysis can further enhance the molecular etiology of the OC and also can be employed in future research for biomarker and drug development related to OC. Further experimental investigations are required to validate our results which will ultimately provide a new perspective for translational applications in OC management. Our study will allow a greater understanding of broader clinical application prospects.

**ORCID**

Berkcan Dogan: https://orcid.org/0000-0001-8061-8131

Ece Gumusoglu: https://orcid.org/0000-0003-3807-0330

Ege Ulgen: https://orcid.org/0000-0003-2090-3621

Osman Ugur Sezerman: https://orcid.org/0000-0003-0905-6783

Tuba Gunel: https://orcid.org/0000-0003-3514-5210

**Authors’ Contribution**

Conceptualization: BD, EG, TG. Data curation: BD, EU. Formal analysis: BD, EU. Funding acquisition: BD, TG. Methodology: BD, EG, EU, OUS, TG. Writing - original draft: BD, EU. Writing - review & editing: BD, EG, EU, OUS, TG.

**Conflicts of Interest**

No potential conflict of interest relevant to this article was reported.

**Acknowledgments**

This work was supported by the Istanbul University Scientific Research Projects Department under grant (grant number: 24059). We thank TUBITAK for the scholarship E.U receives (grant number: 118C039).

**References**


7. Svoronos AA, Engelman DM, Slack FJ. OncomiR or tumor sup-


Introduction

Currently, influenza viruses are still a major cause of respiratory disease and can affect all age groups, resulting in a serious public health problem. The estimated infection rate of...
influenza viruses is approximately 5% to 15% of the population [1]. Furthermore, there are more than 500,000 deaths reported worldwide [2]. Seasonal influenza is caused by influenza A (A/H1N1 and A/H3N2 subtypes) and influenza B (B/Victoria and B/Yamagata lineages) viruses. However, the influenza A viruses cause more severity, and lead to more epidemics and pandemics due to the high mutation rates which result from antigenic drifts and antigenic shifts [3]. First, antigenic drift is caused by the accumulation of point mutations that change the properties of the viral hemagglutinin (HA) and neuraminidase (NA) surface proteins to avoid the host immune system. On the other hand, an antigenic shift is a genetic reassortment process when at least two strains of influenza A viruses have infected within the same cell [4]. During viral replication, the high rate of mutation is promoted by error-prone polymerase enzyme [5]. The mutation rates of the influenza A virus have been reported within a range of $2.0 \times 10^{-6}$ to $2.0 \times 10^{-4}$ mutations per site per round of genome replication [6-8]. Therefore, this evidence suggests that each replicated genome of influenza A carries an average of 2–3 mutations per genome [9]. The virus has gradually adapted to its antigenic sites to avoid the host immune response and vaccination [10]. Due to the high mutation rates, the influenza vaccine was less effective (only 29% to 61%) against seasonal outbreaks during 2019–2020 [11].

In Thailand, influenza transmission occurs year-round with two annual peaks: a major peak in the rainy season and a minor peak in winter [12]. Previous studies have reported that influenza was a major cause of morbidity and mortality in Thailand and resulted in crucial economic costs annually. A study conducted during 2005–2008, estimated an annual average of 36,400 influenza-associated hospitalizations and 300 deaths occurred, with significantly higher mortality rates in children and the elderly [13]. Furthermore, several studies examined the genetic variabilities within HA and NA genes of influenza A viruses based on Sanger sequencing [14-17]. Interestingly, whole genome sequencing (WGS) can be applied to characterize viral strains and provide the comprehensive information of the influenza genome for better understanding of the viral evolution and novel viral strains [18].

Nowadays, next-generation sequencing (NGS) has the advantages of massively parallel sequencing thus making it the ideal tool for characterization of the viral whole genome, viral reassortment and viral mutations [18,19]. Consequently, WGS of influenza viruses based on NGS technology can provide the information necessary to understand the characteristics of influenza viruses. This study aimed to investigate the viral genome and mutations of influenza A viruses circulating in Thailand from 2017 to 2018, and this approach can be further applied for preparation against pandemic and epidemic outbreaks in the future.

**Methods**

**Sample collection and influenza diagnosis**

The study protocol was approved by the Institutional Review Board (IRB No. 337/S7) and Institutional Biosafety Committee (MDCU-IBC No. 001/2018) from the Faculty of Medicine, Chulalongkorn University. Briefly, nasal swab samples from patients with influenza-like illness (ILI) were obtained from Bangpakok 9 International Hospital and Chum Phae Hospital from August 2017 to November 2018. The clinical samples were preserved in viral transport media consisting of Hank’s Balanced Salt Solution supplemented with 1% bovine serum albumin, amphotericin B (15 µg/mL), penicillin G (100 U/mL), and streptomycin (50 µg/mL). The nasal swab samples were screened for influenza virus infection using a one-step multiplex reverse transcription-quantitative polymerase chain reaction (RT-qPCR) as described previously [20,21]. Briefly, the assay was performed in a 10 µL final volume, containing 1 µL of RNA sample, 5 µL of 2 × reaction mix, 0.2 µL of SuperScript III RT/Platinum Taq Mix (Invitrogen, Carlsbad, CA, USA), an additional 0.1 mM of MgCl₂, 0.25 µM of each primer, and 0.125 µM of each probe. The one-step multiplex RT-qPCR was performed on the StepOnePlus Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) using the following thermal cycling conditions: at 55°C for 30 min for reverse transcription, followed by 95°C for 10 min, continuing with 40 cycles of 95°C for 15 s and 60°C for 30 s.

**Cell cultures**

Madin-Darby canine kidney (MDCK) cells were obtained from the American Type Culture Collection (ATCC, Manassas, VA, USA) and cultured in Dulbecco’s modified Eagle Medium (DMEM) with high glucose (HyClone, Logan, UT, USA) supplemented with 10% fetal bovine serum (Gibco, Grand Island, NY, USA) and 1% (v/v) penicillin/streptomycin (Gibco) maintained under humidified 5% CO₂ at 37°C [22].

**Influenza virus isolation**

MDCK cells were used for influenza virus isolation and propagation as described in the previous study [23]. Briefly, MDCK cells were seeded in 60 mm tissue culture dishes (SPL Life Science, Pocheon, Korea) at 5 × 10⁵ cells per dish in DMEM medium without antibiotics. When the cells reached around 80% confluence, the media were removed and then washed by phosphate buffer saline (PBS) (Amresco, Solon, OH, USA). Positive influenza samples
The viral cDNAs from the previous step were used as templates for genome amplifications. The primer sets; forward primer (5'-ACGCGTGATCAGCAAAAGCAGG-3') and reverse primer (5'-ACGCGTGATCAGTAGAAACAAGG-3') were used for amplification of influenza A viral genes (8 segments) following the previous study [25]. Briefly, PCR master mix is composed of 1.25 µM of each primer, 0.35 mM of dNTPs, 0.02 U/µL of Phusion High-Fidelity DNA polymerase (Thermo Fisher Scientific), 7.5 µL of cDNA, and nuclease-free water to a final volume of 50 µL. Subsequently, 15 µL of the amplicons were analyzed by 1% agarose gel electrophoresis. The amplicons were purified by the QIAquick PCR Purification kit (Qiagen, Hilden, Germany) following the manufacturer’s protocol. The concentrations of purified PCR products were measured by the Qubit dsDNA High-Sensitivity assay kit (Invitrogen).

DNA library preparation and next-generation sequencing

The purified amplicons (1 µg in 130 µL) from the genome amplification step were sheared to approximately 200 bp fragments by the Covaris M220 Focused-ultrasonicator (Covaris, UK) with 20% duty factor, 50 unit of peak incident power (W), 200 cycles per burst for 150 s. The fragmented DNAs were used for DNA library preparation by using NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) following the manufacturer's instructions. Briefly, 50 µL of DNA fragments were ends repaired and subsequently adapters ligated by using the NEB ligase master mix. Then, DNA libraries (approximately 320 bp) were cleaned up and size selected by AMPure XP beads (Beckman Coulter, Brea, CA, USA). For library enrichment, PCR amplification was carried out by adding the Illumina MiSeq-compatible indexes to the DNA libraries. Afterwards, the enriched DNA libraries were purified by 2% agarose gel electrophoresis with 100 V for 20 min and size selected (approximately 320 bp). Finally, the total DNA libraries were quantified by real-time PCR using the KAPA Library Quantification Kits (Kapa Biosystems, Wilmington, MA, USA). After that, the concentration of each sample was determined and pooled equally at 2 nM of each library. Subsequently, the pooled library was then diluted to 6 µM and paired-end sequenced (2 × 150 bp) on an Illumina MiSeq instrument using MiSeq Reagent Kits v2 (300 cycles) according to the manufacturer protocol (Illumina).

Influenza genome analysis

The MiSeq Reporter Software version 2.4 was used for the primary analysis of FASTQ sequencing data. Low-quality reads (Q-score < 30) and adaptors were trimmed. The passing filtered reads (Q-score ≥ 30) were aligned with the vaccine strains of influenza
A reference genomes (A/California/07/2009 (H1N1) or A/South Australia/55/2014 (H3N2)) for genome characterization and mutation analysis by using CLC Genomics Workbench software (Qiagen). Mutation patterns and frequencies were generated by using GraphPad Prism version 6.01 software (GraphPad Software Inc., San Diego, CA, USA). The FASTQ files and FASTA files of influenza genome sequences were submitted to the Sequence Read Archive (BioProject ID: PRJNA576776) and GenBank as shown in Supplementary Table 1.

**Phylogenetic analysis**

In this study, the HA and NA deduced amino acid sequences of A/H1N1 and A/H3N2 were aligned with reference strains retrieved from the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu database by using the Clustal W program, implemented in the BioEdit sequence alignment editor software v.7.2.5 [27]. Phylogenetic analysis was performed by mean of the maximum likelihood method (1,000 bootstrapping replicates) and LG with Freqs. (+F) model (discrete gamma distribution with 5-rate categories and complete deletion data subset) using the MEGA X software [28].

**Sliding windows analysis of nonsynonymous nucleotide variation**

The nucleotide diversity (π) within each gene of influenza A/H1N1 and A/H3N2 viruses was evaluated by PoPoolation v.1.2.2 to investigate the genetic variations of viruses within the sample [29]. The sliding window analysis of nonsynonymous nucleotide variation (πN) was performed based on Syn-nonsyn-sliding.pl script with the window size of nine codons and a step size of one codon. The average corresponding πN values were calculated and plotted to a middle position of the windows to demonstrate the degree of nonsynonymous substitutions within eight viral gene segments. In addition, the nonsynonymous nucleotide variations (πN) per synonymous nucleotide variation (πS) were analyzed by the Syn-nonsyn-at-position.pl script to investigate the neutrality of selection in each segment. The πN/πS ratios per gene in each influenza subtype were calculated as the average value from individual samples. Lastly, a paired Wilcoxon signed-rank test (p < 0.05) was used to compare pooled average πN and πS values within each subtype of influenza viruses.

**Results**

**Detection and isolation of influenza A viruses**

In this study, 500 nasal swab samples were collected from patients with ILI and detected for influenza A virus by RT-qPCR. Ninety samples were influenza A virus-positive samples including 48.9% (44 samples) of A/H1N1 and 51.1% (46 samples) of A/H3N2 as shown in Table 1. Among these 90 samples, 43 samples (29 of A/H1N1 and 14 of A/H3N2) were successfully isolated in the second passage (P2) of MDCK cells with Ct value ranging from 13 to 28 (Table 2).

**WGS and characterization of influenza A viruses**

From 43 isolated samples, 25 samples were completely amplified as a full genome including 17 samples of A/H1N1 and eight samples of A/H3N2. Finally, 17 samples (12 samples of A/H1N1 and five samples of A/H3N2) passed the quality control of libraries preparation for NGS as shown in Table 1. The result revealed that 17 whole genomes of influenza A viruses were successfully sequenced with an average of 424,151 total reads per sample, 232,578 mapped reads per sample, and 1,720 genome coverage per sample (Table 2). Therefore, these results were highly confident for genome annotation and mutation analysis. The FASTQ data were deposited as BioProject accession no. PRJNA576776 and influenza genome sequences were submitted into GenBank database as summarized in Supplementary Table 1.

**Phylogenetic analysis of influenza A viruses in Thailand**

The HA and NA deduced amino acid sequences were used for phylogenetic analysis. The sequences were aligned with the HA and NA deduced amino acid sequences of the influenza vaccine strains (southern hemisphere influenza seasons during 2012–2019) recommended by the World Health Organization (WHO). The influenza A/H1N1 viruses isolated from this study during 2017–2018 belonged to genetic subclade 6B.1. Interestingly, the results demonstrated that A/H1N1 viruses were closely related to influenza (A/
Table 2. The sample characteristic, virus isolation and NGS data of influenza A virus in this study

<table>
<thead>
<tr>
<th>No.</th>
<th>Sample name</th>
<th>Age (yr)</th>
<th>Sex</th>
<th>Ct from each passage</th>
<th>Total reads</th>
<th>Mapped reads</th>
<th>Average length of mapped read (bp)</th>
<th>Average genome coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A/Thailand/CU-B23883/2017 (H1N1)</td>
<td>2</td>
<td>M</td>
<td>38 30 27</td>
<td>318,034</td>
<td>54,644</td>
<td>67.2</td>
<td>237.7</td>
</tr>
<tr>
<td>2</td>
<td>A/Thailand/CU-B24063/2017 (H1N1)</td>
<td>19</td>
<td>F</td>
<td>46 34 17</td>
<td>171,064</td>
<td>163,257</td>
<td>82.4</td>
<td>905.6</td>
</tr>
<tr>
<td>3</td>
<td>A/Thailand/CU-B24069/2017 (H1N1)</td>
<td>39</td>
<td>M</td>
<td>36 23 18</td>
<td>325,424</td>
<td>309,926</td>
<td>84.4</td>
<td>1,784.8</td>
</tr>
<tr>
<td>4</td>
<td>A/Thailand/CU-B24076/2017 (H1N1)</td>
<td>12</td>
<td>M</td>
<td>36 33 28</td>
<td>143,698</td>
<td>134,494</td>
<td>79.6</td>
<td>726.3</td>
</tr>
<tr>
<td>5</td>
<td>A/Thailand/CU-B24660/2017 (H1N1)</td>
<td>51</td>
<td>F</td>
<td>33 18 16</td>
<td>518,852</td>
<td>442,915</td>
<td>96.0</td>
<td>2,907.9</td>
</tr>
<tr>
<td>6</td>
<td>A/Thailand/CU-B25124/2017 (H1N1)</td>
<td>3</td>
<td>M</td>
<td>31 32 20</td>
<td>134,102</td>
<td>129,280</td>
<td>87.8</td>
<td>771.9</td>
</tr>
<tr>
<td>7</td>
<td>A/Thailand/CU-B25506/2017 (H1N1)</td>
<td>38</td>
<td>F</td>
<td>32 27 18</td>
<td>165,684</td>
<td>159,654</td>
<td>88.7</td>
<td>952.3</td>
</tr>
<tr>
<td>8</td>
<td>A/Thailand/CU-B27534/2017 (H1N1)</td>
<td>31</td>
<td>F</td>
<td>28 13 13</td>
<td>124,720</td>
<td>112,724</td>
<td>61.6</td>
<td>477.5</td>
</tr>
<tr>
<td>9</td>
<td>A/Thailand/CU-B29642/2018 (H1N1)</td>
<td>30</td>
<td>F</td>
<td>32 29 16</td>
<td>1,355,512</td>
<td>520,427</td>
<td>113.6</td>
<td>3,927.2</td>
</tr>
<tr>
<td>10</td>
<td>A/Thailand/CU-B30312/2018 (H1N1)</td>
<td>59</td>
<td>F</td>
<td>31 27 27</td>
<td>199,458</td>
<td>180,341</td>
<td>120.6</td>
<td>1,460.6</td>
</tr>
<tr>
<td>11</td>
<td>A/Thailand/CU-B30648/2018 (H1N1)</td>
<td>29</td>
<td>F</td>
<td>33 22 15</td>
<td>702,634</td>
<td>367,078</td>
<td>113.6</td>
<td>2,771.4</td>
</tr>
<tr>
<td>12</td>
<td>A/Thailand/CU-E1180/2018 (H1N1)</td>
<td>2</td>
<td>M</td>
<td>20 30 13</td>
<td>555,488</td>
<td>357,666</td>
<td>100.3</td>
<td>2,351.7</td>
</tr>
<tr>
<td>13</td>
<td>A/Thailand/CU-B24411/2017 (H3N2)</td>
<td>61</td>
<td>F</td>
<td>34 26 23</td>
<td>217,002</td>
<td>68,933</td>
<td>60.0</td>
<td>284.1</td>
</tr>
<tr>
<td>14</td>
<td>A/Thailand/CU-B24666/2017 (H3N2)</td>
<td>2</td>
<td>F</td>
<td>27 31 15</td>
<td>139,802</td>
<td>90,676</td>
<td>112.1</td>
<td>696.8</td>
</tr>
<tr>
<td>15</td>
<td>A/Thailand/CU-B28277/2017 (H3N2)</td>
<td>24</td>
<td>M</td>
<td>23 19 24</td>
<td>583,222</td>
<td>467,163</td>
<td>129.5</td>
<td>4,229.7</td>
</tr>
<tr>
<td>16</td>
<td>A/Thailand/CU-B29296/2017 (H3N2)</td>
<td>52</td>
<td>F</td>
<td>30 23 22</td>
<td>557,710</td>
<td>22,569</td>
<td>132.3</td>
<td>2,013.4</td>
</tr>
<tr>
<td>17</td>
<td>A/Thailand/CU-B30632/2018 (H3N2)</td>
<td>53</td>
<td>M</td>
<td>31 24 24</td>
<td>998,160</td>
<td>372,082</td>
<td>107.3</td>
<td>2,747.7</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td>424,151</td>
<td>232,578</td>
<td>96.3</td>
<td>1,720.4</td>
</tr>
</tbody>
</table>

NGS, next-generation sequencing.

Exploring deeper detail about the direction of diversity, the ratios of nonsynonymous to synonymous nucleotide diversity (πN/πS) were introduced to examine the changes in nucleotide variations. In brief, the πN/πS ratios > 1 indicate that selective pressure promotes the new variations (positive selection). In contrast, the πN/πS ratios < 1 refer to the new variation being unfavored (negative selection). In addition, the πN/πS ratios = 1 suggest that neutral evolution occurs in these new variations. According to the results shown in Fig. 4, there was no significant positive selection occurring in this study. However, the statistically significant negative selections (p < 0.05) were found in PB1, PA, HA, and NA genes of A/H1N1. Meanwhile, the A/H3N2 exhibited random selections within these 10 genes due to there being no significant difference observed in the πN/πS ratios.

**Nucleotide diversity of influenza A viruses**

The variations of nonsynonymous within influenza A viruses among the samples in this study are summarized in Fig. 3. As shown in Fig. 3A, strong signals appeared in the polymerase (PB2, PB1, and PA) genes as well as in the NP gene in A/H1N1 virus. However, the NA, M, and NS genes showed the low nonsynonymous variations. Interestingly, the HA gene contained the pattern of the variation signals around the middle position of this A/H1N1 gene. As for the results of A/H3N2 (Fig. 3B), the polymerase genes were presented as sharp and multiple peaks of the nonsynonymous nucleotide diversity. In addition, the HA, NP, and NA genes of A/H3N2 displayed sharp signals at the beginning and the end of these genes. Furthermore, the M and NS genes of the A/H3N2 only had peaks around the middle of the genes.

**Discussion**

In this study, 90 out of 500 (18%) nasal swabs obtained from Thai patients with influenza-like-illness during 2017–2018 were positive for influenza A virus detection based on RT-qPCR detection. The percentage of influenza A virus positive in this research was slightly higher than those reported in the previous study (13.2%) during 2016–2017 in Thailand [30]. Previously, several studies...
have demonstrated that the appropriate quality and quantity of DNA are important for the successful NGS platform sequencing [31-33]. In particular, this study has successfully isolated 47.78% (43 of 90 samples) which are positively identified as the influenza A virus, which is higher than the previous study (3.04% of isolation rate) [34]. Also, the positive virus isolations (58.14%, 25 of 43 isolates) can be amplified with universal primers, following the study of Meinel et al. [35], which is appropriate for whole genome characterization and mutation analysis of influenza A virus. For NGS analysis, the result revealed that 17 whole genomes of influenza viruses were successfully sequenced with 232,578 mapped reads (424,151 total reads), average read length of 96.3 bp and average 1,720.4 genome coverage. Furthermore, the complete sequences of the viral genomes provide reliable and highly informative data despite the average genome coverages, depth coverage, which ranged from 237.7 to 4,229.7 (Table 2). The advantages of the NGS-based technique are that it provides the full genome segment and whole genome of influenza virus, as well as effectively reducing both the turnaround time and cost per nucleotide sequence for the whole genome when compared to the Sanger sequencing method [36-38]. However, the sequencing with the Sanger method does not provide the data for quasispecies and nucleotide diversity analysis. Interestingly, the NGS provides more information for minor mutations and selection pressures within the viral genome. Indeed, the nucleotide variations obtained from NGS can be applied for calculation of viral nucleotide diversity within each sample [39,40].

The number of mutations in the HA and NA genes of A/H1N1 might affect the efficiency of a vaccine, and related to deduced amino acid sequences of phylogenetic tree (Fig. 1A and 1B). The vaccine effectiveness of the 2017–2018 flu vaccine against both influenza A viruses is approximately 25% to 52% in Europe and 27% to 44% in the United States [41,42]. Interestingly, the result of the influenza A/H1N1 phylogenetic tree with deduced amino acid se-
quences, which belongs to clade 6B.1, showed a long distance between vaccine strains for 2017–2018 (A/Michigan/42/2015) and our A/H1N1 samples. This result implied that the vaccine might be less effective against A/H1N1 in Thailand. Moreover, the report from the US Centers for Disease Control and Prevention (CDC) also showed the vaccine effectiveness against A/H1N1 was 65% [43]. However, the phylogenetic analysis of both HA and NA deduced amino acid sequences revealed the closer relationship between A/H3N2 isolates (clade 3C.2a1 and 3C.2a2) and A/Hong Kong/4801/2014 strain which was the recommended vaccine virus for A/H3N2 [44]. Therefore, these results implied that the recommended vaccine was more effective against A/H1N1 in Thailand during 2017–2018. Indeed, the phylogenetic trees of influenza A/H1N1 and A/H3N2 obtained in this study were correlated with recent genetic and antigenic characterizations of influenza viruses in Thailand [45].

In this study, the genome of circulating influenza A viruses in Thailand during 2017–2018 was characterized. The result from NGS analysis not only provided the full genome of the virus but also acquired the amino acid substitutions across eight segmented genes. Moreover, there were several known functional mutations of influenza A/H1N1 that had been already characterized. Firstly, the mutations at I354, V344M, and S453T in the PB2 (Fig. 5A) could regulate in the cap-snatching from host RNAs during the viral RNA transcription process [46]. Furthermore, N321K in the PA was reported to increase the polymerase complex activity and the viral replication in the cell culture [46,47]. The amino acid substitution at V100I in the PA-X could trigger down-regulated innate immune response genes (Fig. 5C). Indeed, the amino substitutions at S91R, S181T, I312V, and E391K in the HA might be related to adaptive genetic variations that alter the salt bridge pattern and the membrane fusion stability for major antigenic sites and glycan specificitity [46,47]. Three mutations (K180Q, S202T, and S220T) were located in the HA antigenic sites, which might be involved in the pathogenicity and contributed to the epidemic [48]. Moreover, the mutation (S220T) was observed to affect the infectivity and transmissibility of the virus in humans [49]. The mutation (R240Q) was found in the receptor-binding domain of the

Fig. 2. The phylogenetic analysis of influenza A viruses (H3N2) circulating in Thailand during 2017–2018 (diamond) compared with several World Health Organization recommended influenza vaccine strains (black triangle). The hemagglutinin (HA) (A) and neuraminidase (NA) (B) deduced amino acid sequences were analyzed based on maximum likelihood with 1,000 bootstrapping and LG with Freqs. (+F) model (discrete gamma distribution with 5-rate categories and complete deletion data subset).
HA, which has been reported to increase virus growth [50]. The amino acid variations (D114N, K180Q, S202T, S220T, and K300E) were responsible for loss of antibody neutralization and decreased overall vaccine effectiveness ([51-53]). The amino acid substitutions (N44S, V241I, and N369K) in the NA have been reported to facilitate the stability of the virus [54]. The I188T and N449D mutations in the NA found in this study are similar to those reported in the previous study [55]; however, the function of the mutations has not been well characterized ([Fig. 5F]). Additionally, the nonsynonymous mutation at E55K, L90I, I123V, E125D, K131E, and N205S in the NS1 ([Fig. 5H]) involves the inhibition of host gene expressions related to the interferon response [56,57]. Indeed, the E125D mutation in NS1 interacts with cellular cleavage and polyadenylation specificity factor 30 (CPSF30), which is considered potential in host adaptation to influenza A/H1N1 virus [58,59].

In the influenza A/H3N2, the previous reports found R277Q and D69N at the antigenic epitope C, N137K, and N187K at the antigenic epitope D and E78K/G at the antigenic epitope E of the HA [60,61]. Among these, four amino acid substitutions (N137K, N187K, I422V, and G500E) belong to clade 3C.2a.1 and are represented by A/Singapore/INFIMH-16-0019/2016 (H3N2) virus [62]. The T151N substitution in HA protein was related to the potential N-glycosylation site, affecting antigenic and other viral properties. Moreover, the Q327H substitution in the HA was suggested to bind host proteins ([Fig. 6D]) [62]. Since 2016, the accumulation of mutation at S245N of the NA has contributed to an N-glycosylation site. These mutations (S245N, S247T, and

---

**Fig. 3.** Sliding windows analysis of nonsynonymous nucleotide variation (nN) in eight genes of influenza A virus subtypes H1N1 (A) and H3N2 (B). The nN values were determined by sliding windows with the window size of nine codons and a step size of one codon. The mean corresponding nN values were calculated and plotted to a middle site of the windows.
Fig. 4. The ratio of nonsynonymous nucleotide variation (nN) to synonymous nucleotide variation (nS) analysis in eight genes of influenza A virus subtypes H1N1 (A) and H3N2 (B). Significant at *p < 0.05, **p < 0.01 and ***p < 0.001. The ratio nN/nS > 1: positive selection; nN/nS < 1: negative selection; nN/nS = 1: neutral evolution. The mean nN/nS, standard deviation (S.D.), and p-value (Student’s t-test) in each segment were summarized at the bottom of the figure.

Fig. 5. Mutation patterns with actual mutations frequencies observed in each viral gene segments of influenza A virus (H1N1): (A) PB2, (B) PB1, (C) PA, (D) HA, (E) NP, (F) NA, (G) M, and (H) NS. Amino acid changes were compared to the reference sequence (A/California/07/2009 (H1N1)).
P468H) were introduced to the NA antigenic drift of the circulating A/H3N2 virus [63]. However, N329S mutation could result in a loss of N-glycosylation in the NA [64]. The V303I substitution has been observed in the NA protein (Fig. 6F) with a low resistance to NA inhibitors [65]. Indeed, most mutations of influenza A viruses observed in this study were identified as novel mutations which have not been reported yet (Figs. 5 and 6). However, the function of the novel mutations needs to be further investigated.

Nonsynonymous (πN) and synonymous (πS) mutations of the viruses can be accessed by NGS leading to nucleotide diversity (πN/πS) analysis. According to the previous study, the deep sequencing of A/Wisconsin/67/2005 (H3N2) revealed that the positive selection was observed in the viruses isolated from the chicken kidney, Vero cell culture, and embryonated chicken eggs, whereas the negative selection was found in virus from direct intranasal inoculation in the human challenge [40]. There was no significant nucleotide diversity observed in A/H3N2 viruses in

---

**Fig. 6.** Mutation patterns with actual mutations frequencies observed in each viral gene segments of influenza A virus (H3N2): (A) PB2, (B) PB1, (C) PA, (D) HA, (E) NP, (F) NA, (G) M, and (H) NS. Amino acid changes were compared to the reference sequence (A/South Australia/55/2014 (H3N2)).
our study, and this might be due to the strain of the virus, host cell, or limited numbers of the sample. For πN/πS analysis of influenza A/H1N1 viruses, the mutations existing in the viral genes with statistical significance were PB1, PA, HA, and NA genes in which these mutations were suggested as negative selection. Therefore, to further investigate the πN and πS variations, the sliding window analysis of those genes was performed to ensure that the negative selections were not the outcome of the averaging value across the entire gene. The results of sliding window analysis were consistent with the negative selections from the πN/πS analysis in those genes at which the πN signals were high and sharp at some regions of the genes, while the rest of the genes were relatively low in the πN signals.

In summary, the NGS was successfully applied for whole genome characterizations of influenza A/H1N1 and A/H3N2 viruses that provide the high-throughput data for phylogenetic construction, mutation analysis, and nucleotide diversity. The results revealed that the recommended vaccine A/H1N1 strain might be less effective against the A/H1N1 virus. Moreover, several mutations were demonstrated in both A/H1N1 and A/H3N2, especially in HA and NA genes. Finally, the negative selections were found in the PB1, PA, HA, and NA genes of the A/H1N1. Unfortunately, limited number of samples were successfully propagated, amplified, and sequenced in this study. Nevertheless, the whole genome data obtained from this study might be useful for mutation analysis and can be compared with data obtained from other studies in the future.

ORCID

Somruthai Rattanaburi: https://orcid.org/0000-0002-1821-6673
Vorthon Sawaswong: https://orcid.org/0000-0003-2805-6690
Pattaraporn Nimsamer: https://orcid.org/0000-0003-3741-7558
Orphan Mayuramart: https://orcid.org/0000-0001-5972-6648
Pavaret Sivapornnukul: https://orcid.org/0000-0002-6714-3407
Ariya Khamwut: https://orcid.org/0000-0002-5018-953X
Prangwalai Chanchaem: https://orcid.org/0000-0001-9194-1818
Kritsada Kongnomnan: https://orcid.org/0000-0002-9213-5513
Nungruthai Suntronwong: https://orcid.org/0000-0002-0714-6470
Yong Poovorawan: https://orcid.org/0000-0002-2337-6807
Sunchai Payungporn: https://orcid.org/0000-0003-2668-110X

Authors’ Contribution

Conceptualization: SP. Data curation: SR, VS, SP. Formal analysis: VS, SR, PN. Funding acquisition: SP. Methodology: SR, OM, AK, PC, NS, YP. Writing - original draft: SR. Writing - review & editing: SP, PS, KK.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was partly supported by a grant from the National Science and Technology Development Agency (NSTDA) (P-17-51377), the Royal Golden Jubilee (RGJ) Ph.D. Programme scholarship (PHD/0150/2558), National Research Council of Thailand (NRCT)[2564NRCT321520] and Thailand Research Fund (TRF), the Chulalongkorn University Center of Excellence in Systems Biology and Chulalongkorn Academic Advancement into its 2nd Century Project. We would like to express gratitude to all the members of the Center of Excellence in Clinical Virology (Miss Preeyaporn Vichaiwattana and Mr. Sumeth Korkong), Faculty of Medicine, Chulalongkorn University for facilitating the sample collection. Moreover, we would like to thank the Department of Virology, Faculty of Veterinary Medicine Chulalongkorn University for support in influenza virus isolation (Assistant Prof. Aunyaratana Thontiravong, D.V.M., Ph.D.).

Supplementary Materials

Supplementary data can be found with this article online at http://www.genominfo.org.

References

5. Boivin S, Cusack S, Ruigrok RW, Hart DJ. Influenza A virus polymerase: structural insights into replication and host adaptation


Mathematical modeling of the impact of Omicron variant on the COVID-19 situation in South Korea

Jooha Oh\textsuperscript{1}, Catherine Apio\textsuperscript{2}, Taesung Park\textsuperscript{1,2,*}

\textsuperscript{1}Department of Statistics, Seoul National University, Seoul 08826, Korea
\textsuperscript{2}Interdisciplinary Programs in Bioinformatics, Seoul 08826, Korea

The rise of newer coronavirus disease 2019 (COVID-19) variants has brought a challenge to ending the spread of COVID-19. The variants have a different fatality, morbidity, and transmission rates and affect vaccine efficacy differently. Therefore, the impact of each new variant on the spread of COVID-19 is of interest to governments and scientists. Here, we proposed mathematical SEIQEDVP and SEIQRDV3P models to predict the impact of the Omicron variant on the spread of the COVID-19 situation in South Korea. SEIQEDVP considers one vaccine level at a time while SEIQRDV3P considers three vaccination levels (only one dose received, full doses received, and full doses + booster shots received) simultaneously. The omicron variant’s effect was contemplated as a weighted sum of the delta and omicron variants’ transmission rate and tuned using a hyperparameter $k$. Our models’ performances were compared with common models like SEIR, SEIQR, and SEIQRDV3P using the root mean square error (RMSE). SEIQRDV3P performed better than the SEIQRDV3P model. Without consideration of the variant effect, we don’t see a rapid rise in COVID-19 cases and high RMSE values. But, with consideration of the omicron variant, we predicted a continuous rapid rise in COVID-19 cases until maybe herd immunity is developed in the population. Also, the RMSE value for the SEIQRDV3P model decreased by 27.4%. Therefore, modeling the impact of any new risen variant is crucial in determining the trajectory of the spread of COVID-19 and determining policies to be implemented.

**Keywords:** COVID-19, mathematical models, Omicron, SARS-COV-2, variant

**Introduction**

Coronavirus disease widely known as coronavirus disease 2019 (COVID-19) is a new disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus that emerged from Wuhan, China \cite{1} and spread rapidly across the world becoming a global pandemic in March 2020 \cite{2}. The pandemic caught many governments and people by surprise which led to the implementation of unprecedented intervention policies like school and workplace closures, suspension of public transportation, international travel restrictions, and so forth \cite{3}, with aims of mitigation and suppression. The reason was to not overwhelm the unprepared healthcare systems and lower the number of cases until a pharmaceutical solution was found \cite{4,5}. In addition, many pharmaceutical companies in partnership with government bodies launched one of the fastest vaccine development projects of our decade leading to the development of multiple vaccines \cite{6,7}. The Coalition for Epidemic Preparedness Innovations (CEPI) worked with global health authorities, governments, and vaccine developers to support the development of vaccines against...
COVID-19 [8].

This pandemic also attracted the interest of researchers from different fields since it was a new disease and its transmission pathways and fatality were not yet known. From the outbreak to 4 February, 2022, researchers have contributed a lot in forecasting, and understanding the transmission dynamics, the fatality of SARS-CoV-2, and the evolution of the pandemic, to help in the fight against this new global problem. Among the researchers, statisticians, epidemiologists, and mathematicians contributed to formulating models to capture the transmission dynamics of COVID-19 and forecasting the evolution of the pandemic among different populations amidst government interventions. These mainly included statistical models [9-18], deep-learning models [19-24] and mathematical models [1,14,25,26].

Statistical models offer more precise models and deep-learning techniques are the key to high-quality predictive models [27]. However, both statistical and deep-learning models require real data to make predictions. But with mathematical models, a set of mathematical equations that mimic the current situation is written, and solving them for certain parameters provide information about the disease characteristics [28]. Some of their advantages include mathematical models representing the real situation of the problem being solved and they do not require all data to be available for it to be fitted as deductions from known information about the situation can be used. Also, they can handle sudden changes and complexity with ease. Since the start of the COVID-19 pandemic, mathematical models have been at the forefront of determining and forecasting the spread of COVID-19 and shaping government policies around the world [28].

A seminal paper in 1927 introduced the Susceptible, Infectious, and Recovered (SIR), a mathematical model for infectious diseases [29]. Since then, with advances in information technology and fast computing methods, many variations of the SIR model have been developed. Because mathematical models can easily be understood and definite conclusions about the COVID-19 outbreak can be made from them, Susceptible, Exposed, Infectious and Recovered (SEIR), a modification of SIR and a cascade of other modifications have been constructed and developed for predicting COVID-19 since its declaration as a global pandemic [30-42].

Subsequently, on 8 December, 2020, 272 days after COVID-19 was declared a global pandemic, vaccination started in the United Kingdom [43]. Since then, as of 4 February, over 61.34% of the world population has received at least one dose of a COVID-19 vaccine, and 21.54 million doses of vaccines are administered daily around the world [44]. However, since then new variants of the SARS-CoV-2 virus have appeared. These variants have different transmissibility rates, fatality, and morbidity. Furthermore, existing vaccines have differing efficacy levels against these emerging variants [45,46]. Governments must make decisions and revise policies while considering these new developments like the impact of vaccinations and emerging variants on the spread of COVID-19. However, statistical and deep-learning models would require real data in substantial amounts to perform any forecasting or prediction. On the other hand, these new developments can easily be modeled with little or no data with mathematical models.

For the Korean COVID-19 situation, many models were employed to forecast the future COVID-19 situation in the country amidst government social distancing policies. One paper used the SIR model with time-dependent parameters and deep learning to forecast the spread of COVID-19 in South Korea [47]. Another analysis utilized the SIR model with breakpoint information that allows change in transmission rate at the breakpoints was established [48]. Other uses of the SIR model or its modification used for the Korean population are found elsewhere [49-53]. A modification of the SEIR model that considers transmission rates between age groups and vaccination was also formulated for the Korean population [54]. In this model, five additional groups; quarantined Q, unprotected U, vaccinated V, protected P, and deceased D were added to the standard SEIR model making it the SEIQRDVUP model.

Since the SARS-CoV-2 virus is an RNA virus and lacks the mismatch repair mechanism, the virus replication process is accompanied by a high mutation rate, hence the rise of variants [55]. Common mutant variants include B.1.1.7, B.1.351, B.1.1.28.1, B.1.617.2 (Delta), and B.1.1.529 (omicron), which have all spread rapidly worldwide. The mutations make the virus more contagious (fast-spreading) and difficult to eliminate [56]. However, the SEIQRDVUP model and other previous methods cannot catch the sudden increase in daily cases caused by newer variants with higher transmission rates compared to a previously dominant variant.

To solve this limitation, we formulated a modification of the SEIQRDVUP model to consider a weighted sum of delta and omicron variants’ transmission rates based on variants’ proportions together. In addition, three vaccination levels (only one dose received, full doses received, and full doses+booster shots received) were considered by adding three more compartments of vaccination (V1, V2, and V3) and the removal of the above-mentioned U compartment due to the use of a transmission rate that includes the effect of vaccine efficacy thereby eliminating the ineffectively vaccinated group, U. So, the omicron variant’s effect was contemplated as a weighted sum of the delta and omicron variants’
transmission rate. In this case, the omicron variant’s transmission rate is assumed to be a multiple of the delta’s transmission rate, as explained in detail in the Methods section. This study aims to examine how the omicron variant will affect the COVID-19 situation in Korea with our proposed SEIQRDVP and SEIQRDV3P models. The SEIQRDVP considers only one vaccination level at a time using only one vaccination compartment. However, the SEIQRDV3P considers all three vaccination levels simultaneously.

**Methods**

**Proposed SEIQRDVP and SEIQRDV3P models**

Mathematical methods can be used for the prediction and forecasting of COVID-19 transmission [57-60]. Here, we proposed the SEIQRDVP model, and its flowchart is shown in Fig. 1 below. The susceptible group $S$ is the group of unvaccinated and uninfect ed people that can still be infected by the infectious group. The vaccinated group $V$ is a group of people vaccinated and can still be infected by the infectious group but with a lower transmission rate. In this case, the efficacy of the vaccine is multiplied by the transmission rate. If a host in $S$ or $V$ group gets infected, this host becomes a host of the exposed group, $E$. After the incubation period, a host of $E$ can infect $S$ or $V$ groups, which means that a host of $E$ becomes a host of $I$, the infectious group. When a host of $I$ is determined to be infected, a host will be isolated and becomes a host of an isolated group, $Q$, and loses the ability to infect others. An isolated host will be recovered or be dead and moves to group $R$ or $D$, which are the recovered group and deceased group. Group $P$ is the insusceptible group that has immunity. The following differential equations represent the SEIQRDVP model:

\[
\begin{align*}
\frac{dS}{dt} &= -BS \frac{I}{N} - \nu, \\
\frac{dE}{dt} &= \beta(S+(1-\epsilon)V) \frac{I}{N} - \kappa E, \\
\frac{dI}{dt} &= \kappa E - aI, \\
\frac{dQ}{dt} &= aI - \gamma Q, \\
\frac{dR}{dt} &= (1-f)\gamma Q, \\
\frac{dD}{dt} &= f\gamma Q, \\
\frac{dV}{dt} &= \nu - (1-\epsilon)\beta V \frac{I}{N} - \omega V, \\
\frac{dP}{dt} &= \omega V,
\end{align*}
\]

\[N=S+E+I+Q+R+V+P.\]

where $\beta$ is the transmission rate, $\epsilon$ is vaccine efficacy, $f$ is the mortality rate, $a,\gamma, k$ and $\omega$ are the duration periods from respective previous compartment to the next compartment, $N$ is the total population, and $\gamma$ is the isolation duration. Previously determined model parameters from literature, $\kappa, \alpha, \gamma, f, \epsilon$, and used in our analysis are listed in Table 1. We assumed that the vaccinated host gets immunity 42 days after their first vaccination which means that $1/\omega$ is assumed to be 42 [57]. In Fig. 1, $\nu$ is provided by daily vaccinated cases. Consequently, the remaining parameter $\beta$ is the only unknown parameter estimated by the least-squares method. This process is done using Runge-Kutta fourth-order method and the lsqcurvefit toolbox in MATLAB [61]. In addition, the daily cases are divided into segments with the breakpoints of these segments being determined from the changing levels of the stringency index due to changing government policies. The stringency index was obtained from the Oxford COVID-19 Government Response Tracker (OxCGRT) dataset from the Blavatnik School of Government and the University of Oxford [62,63]. $\beta$ was estimated for each segment independent of other segments, therefore our pro-

![Fig. 1. Flowchart of SEIQRDVP model.](https://doi.org/10.5808/gi.22025)

**Table 1. Previously determined model parameters**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1/\kappa$</td>
<td>The average duration from E to I</td>
<td>4.1 days [64-66]</td>
</tr>
<tr>
<td>$1/\alpha$</td>
<td>The average duration from I to Q</td>
<td>6 days [64,65]</td>
</tr>
<tr>
<td>$1/\gamma$</td>
<td>The average duration from Q to R or D</td>
<td>20.1 days [67]</td>
</tr>
<tr>
<td>$f$</td>
<td>Mortality rate</td>
<td>0.09 [68]</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>Efficacy of vaccination</td>
<td>0.78 [69]</td>
</tr>
</tbody>
</table>
posed model included stringency index as a covariate.

Moreover, the vaccination group can be divided into three: vaccinated (first vaccination), fully vaccinated (second vaccination), and boosted (third vaccination). We call the model that fits the three vaccination levels simultaneously, the SEIQRDV3P model, and its flowchart is shown in Fig. 2. In this model, $v_1, v_2, v_3$ are provided by daily vaccinated, daily fully vaccinated, and daily boosted cases. Also, the efficacy of vaccination for each vaccination group is differently provided with $e_1, e_2, e_3$ and values are 0.75, 0.80, 0.85 [69].

Lastly, the proportion of omicron variants was reflected in the above model as a change in transmission rate, $\beta$. With a transmission rate of delta variant as $\beta_0$ and transmission rate of omicron variant as $\beta_\text{r}$, we assumed that $\beta_\text{r}$ is multiple of $\beta_0$, which means $\beta_\text{r} = k \beta_0$ with hyperparameter $k$. In our cases, we tried 1, 3, 5, and 7 as a value of the hyperparameter $k$, to track the recent rapid increase of the omicron variant. The proportion of the omicron variant in the population is modeled by the parameter $w$. The values of $w$ lie between 0 and 1. The time-series variation of this parameter is known for both train and test data, but its variation for the coming days is unknown. So, a logistic function was fitted to predict the future behavior of $w$. Using past data on the proportion of the omicron variant, the logistic function of the proportion of the omicron variant against time was fitted by the least square method.

Combining these results altogether, the final transmission rate be-cause of the omicron variant was gotten from GISAID, an initiative dedicated to the tracking of virus variants from the influenza viruses and coronavirus [72,73]. This data was divided into train and test data. The training period was chosen from 20 September 2021 to 28 January 2022 since from this date (2021 September 20), the proportion of cases of delta variant had exceeded 90% of the cases. The test data period for prediction was from 29 January 2022 to 4 February 2022.

Results

SEIQRDV3P and SEIQRDV3P models

Modifications of the basic SEIR model to the SEIQR model, to the SEIQRDVUP model, and then to our proposed SEIQRDV3P and SEIQRDV3P models were done and the models’ performances were compared. For each model, using train data, time-dependent $\beta(t)$ using the different models were estimated by the least-squares method (LSE). SEIQR, SEIQRDVUP, and SEIQRDV3P models’ results showed similar fitting with our proposed SEIQRDV3P model. We observed that except for the basic SEIR model and our SEIQRDV3P model, the other models had similar daily cases fitted curves.

Using test data, the prediction error of each model using actual confirmed cases and predicted confirmed cases from models was determined using root mean square error (RMSE). RMSE values for the SEIR, SEIQR, SEIQRDVUP, SEIQRDV3P, and SEIQRDV3P models were calculated as 11,235, 5,079, 5,116, 5,115, and 5,101, respectively, as shown in Table 2. A general decrease in RMSE values with an increase in model complexity is observed. However, the difference in RMSE between SEIQR to SEIQRDV3P models is way smaller than the difference between SEIQR and SEIR models.

Effect of omicron variant

From the above result, SEIQRDV3P and SEIQR models had the

![Fig. 2. Flowchart of SEIQRDV3P model. Parameters for $V_1, V_2, V_3$ to $E$ are $(1-e_1)\beta, (1-e_2)\beta, (1-e_3)\beta$ like Fig. 1.](image-url)

Table 2. RMSE values for each models and at different different k (SEIQRDV3P)

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>$k$ (SEIQRDV3P)</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEIR</td>
<td>11,235.23</td>
<td>1</td>
<td>5,101.342</td>
</tr>
<tr>
<td>SEIQR</td>
<td>5,079.369</td>
<td>3</td>
<td>4,583.178</td>
</tr>
<tr>
<td>SEIQRDVUP</td>
<td>5,116.04</td>
<td>5</td>
<td>4,200.31</td>
</tr>
<tr>
<td>SEIQRDV3P</td>
<td>5,115.755</td>
<td>7</td>
<td>3,705.078</td>
</tr>
<tr>
<td>RMS</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

RMSE, root mean square error.
lowest but almost similar prediction errors, despite large differences in the model structure. This could be because of the recent omicron situation in Korea. As transmissibility between the delta variant, which was originally dominant in Korea, and the omicron variant which is now the dominant variant differ greatly. The SEIQRDV3P model which considers three vaccination levels simultaneously was updated to reflect the different effects of each variant due to their different transmissibility rate. Using train data, time-dependent $\beta(t)$ was estimated by the LSE method. In this case, hyperparameter $k$ was chosen as $3, 5, 7$, which means that the transmission rate of the omicron is $3, 5, 7$ times of transmission rate of the delta. For each selected $k$, the best fitted daily cases curves are shown in Fig. 3. The model with $k = 1$ corresponds to the original SEIQRDV3P model. The x-axis (time) includes both the training and testing period.

Also, RMSE values for each case were calculated as $5,101, 4,583, 4,200, and 3,705$ for each value of hyperparameter $k$ ($1, 3, 5, and 7$). As we include the effect of the omicron variant in the SEIQRDV3P model, we can observe the dramatic decrease in RMSE values. Also, RMSE values decreased as the hyperparameter $k$ increased, as shown in Table 2. This result implies that in a short period, the omicron variant shows way larger transmissibility than the delta variant. Seven days’ prediction after the test data period, which is 5 February 2022 to 11 February 2022, is shown in Fig. 4.

**Discussion**

Since the onset of the global COVID-19 pandemic, mathematical models have been at the forefront of forecasting the future pandemic situation hence policymaking by government bodies. Mathematical models are highly flexible and the impact of different scenarios on the transmission of COVID-19 can be incorporated and predicted, even with the unavailability of data. The mathematical compartmental SEIR model and many of its modifications have been developed.

Governments must revise their testing protocols, social distancing policies, and healthcare protocols with the emergence of each new variant, hence the need of modeling the impact of each emerging variant on the spread of COVID-19. Here, we proposed a modification of the published SEIQRDVUP model, the SEIQRDV3P model which considers one vaccination group at a time, and the SEIQRDV3P model which models the three vaccination levels simultaneously and the impact of the omicron variant. SEIQRDV3P and SEIQRDV3P models’ performance were compared to SEIQRDVUP and other known compartmental mathematical models SEIR and SEIQR models. Firstly, without considering the omicron variant rate, our SEIQRDV3P model doesn’t show much difference from other models contrasted here. This result implies that the SEIQRDV3P model cannot predict a rapid
increase in daily COVID-19 cases without a previous increasing daily case pattern.

However, using a hyperparameter and a weighted sum of transmission rates between two variants, we were able to predict the rapid increase caused by the omicron variant. Omicron rate considering weighted sum lowers the prediction error of the SEIQRD-V3P model from 5,101.342 to 3,705.078 which is 27.4% less than the SEIQRDVUP model. Since the omicron variant has a way larger transmission rate than delta or other previously known variants, it seems that daily incidences will keep increasing until herd immunity for the omicron variant is formed in the population.

However, from January 2022, daily deaths, as well as severity, seem to have decreased considerably. This pattern can imply the low risk and mortality associated with the omicron variant compared to the delta variant, or the impact of vaccination on the population. Therefore, before implementing the ‘Living with COVID-19’ policy in Korea [74], the prediction of deceased and serious patient cases should be preceded. This work can be done by developing the mortality rate in the SEIQRDV3P model to also consider the omicron variant’s mortality with the weighted sum method.

Considering that each variant has its different transmissibility rates, fatality, impact on vaccine efficacy, and morbidity, this generates different model parameter values making it difficult to model all current variants in one model. Therefore, each variant would require its model. Currently, using different model parameters for each variant remained a limitation of this study which we try to solve in the future. Also, SARS-CoV-2 has been known to affect age groups differently. Furthermore, the impact of variant and vaccination policies across different age groups of the population will be considered in our future studies.

With the appearance of new COVID-19 variants appearing after a few months, the fight to end the spread of SARS-CoV-2 even with vaccination has been greatly challenged. These new variants have a different fatality, transmission rate, and efficacy from currently available vaccines. Therefore, their effect on daily cases, deaths, and implemented non-pharmacological policies is of interest to governments and scientists. With the proposed SEIQRDV3P model we found out the new omicron variant will cause a rapid rise in COVID-19 cases in South Korea for some time until herd immunity is developed in the population.

**ORCID**

Jooha Oh: https://orcid.org/0000-0003-0116-0359
Catherine Apio: https://orcid.org/0000-0003-0240-794X
Taesung Park: https://orcid.org/0000-0002-8294-590X
Authors' Contribution

Conceptualization: TP. Data curation: JO. Formal analysis: JO. Funding acquisition: TP. Methodology: JO. Investigation: JO, CA. Visualization: JO, CA. Supervision: TP. Project administration: TP. Writing - original draft preparation: CA, JO. Writing - review and editing: TP, CA, JO.

Conflicts of Interest

Taesung Park serves as an editor of the Genomics and Informatics, but has no role in the decision to publish this article. All remaining authors have declared no conflicts of interest.

Acknowledgments

This research was supported by the Bio and Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. 2021M3ESE3081425).

References

110017.
41. Sahin U, Sahin T. Forecasting the cumulative number of confirmed cases of COVID-19 in Italy, UK and USA using fractional nonlinear grey Bernoulli model. Chaos Solitons Fractals 2020;138:109948.


Comparison of survival prediction models for pancreatic cancer: Cox model versus machine learning models

Hyunsuk Kim¹, Taesung Park², Jinyoung Jang³, Seungyeoun Lee⁴*

¹Department of Statistics, University of California, Berkeley, CA 94720, USA
²Department of Statistics, Seoul National University, Seoul 08826, Korea
³Department of Surgery and Cancer Research Institute, Seoul National University College of Medicine, Seoul 03080, Korea
⁴Department of Mathematics and Statistics, Sejong University, Seoul 05006, Korea

A survival prediction model has recently been developed to evaluate the prognosis of resected nonmetastatic pancreatic ductal adenocarcinoma based on a Cox model using two nationwide databases: Surveillance, Epidemiology and End Results (SEER) and Korea Tumor Registry System-Biliary Pancreas (KOTUS-BP). In this study, we applied two machine learning methods—random survival forests (RSF) and support vector machines (SVM)—for survival analysis and compared their prediction performance using the SEER and KOTUS-BP datasets. Three schemes were used for model development and evaluation. First, we utilized data from SEER for model development and used data from KOTUS-BP for external evaluation. Second, these two datasets were swapped by taking data from KOTUS-BP for model development and data from SEER for external evaluation. Finally, we mixed these two datasets half and half and utilized the mixed datasets for model development and validation. We used 9,624 patients from SEER and 3,281 patients from KOTUS-BP to construct a prediction model with seven covariates: age, sex, histologic differentiation, adjuvant treatment, resection margin status, and the American Joint Committee on Cancer 8th edition T-stage and N-stage. Comparing the three schemes, the performance of the Cox model, RSF, and SVM was better when using the mixed datasets than when using the unmixed datasets. When using the mixed datasets, the C-index, 1-year, 2-year, and 3-year time-dependent areas under the curve for the Cox model were 0.644, 0.698, 0.680, and 0.687, respectively. The Cox model performed slightly better than RSF and SVM.

Keywords: Cox model, random survival forests, support vector machines, survival prediction model

Introduction

Pancreatic cancer is well-known as one of the most lethal cancers worldwide because it has a 5-year overall survival rate of 12.6% as of 2020, while other cancers have 5-year overall survival rates of over 80%. The survival rate strongly depends on the stage of cancer and disease severity. For example, in patients with stage I pancreatic cancer, the 5-year postoperative survival rate is 70.32%, while in patients with stage IV cancer, the 5-year postoperative survival rate is only 3.52%. Therefore, early diagnosis and prediction have been considered promising ways to improve the survival rate of pancreatic cancer.

A survival prediction model for resected pancreatic ductal adenocarcinoma (PDAC) was recently developed with data from the Surveillance, Epidemiology and End Results...
(SEER) database from the United States and external validation using the nationwide Korea Tumor Registry System-Biliary Pancreas (KOTUS-BP) dataset [1]. This prediction model uses a Cox model, which assumes linear associations with many clinicopathologic features. However, it is necessary to investigate nonlinear relationships or complex interactions between clinicopathologic features associated with the survival time. To this end, we applied two machine learning (ML) methods—random survival forests (RSF) and support vector machines (SVM)—to construct predictive models for survival analysis.

In this study, three different schemes were conducted for model development and external evaluation. First, we utilized data from SEER for model development and data from KOTUS-BP for external evaluation. Secondly, these two datasets were used in reverse by taking data from KOTUS-BP for model development and data from SEER for external evaluation. Finally, we mixed these two datasets half and half and utilized the mixed datasets for model development and external validation.

For each of the three different schemes, we developed prediction models using a Cox proportional hazards model, RSF, and SVM. We compared their performance in terms of the C-index and 1-year, 2-year, and 3-year time-dependent areas under the curve (AUCs).

**Methods**

**Data**

This study utilized two nationwide databases: the SEER database from the United States and the KOTUS-BP database from Korea. The datasets were pre-processed as described elsewhere [1]. In the screening process, 9,624 patients from SEER and 3,281 patients from KOTUS-BP were selected. Due to the different sets of covariates in the two datasets, only seven covariates—including age, sex, histologic differentiation, adjuvant treatment, resection margin status, and the American Joint Committee on Cancer (AJCC) 8th edition T-stage and N-stage—were utilized in this study.

The SEER database, which has been maintained by the National Cancer Institute in the United States since 1975, is one of the largest and highest-quality cohort studies, whereas the KOTUS-BP database was launched by the Korean Association of Hepato-Biliary-Pancreatic Surgery in 2014 and has been prospectively registered and regularly managed by pancreatobiliary surgeons at specialized centers in Korea. To unify the study period, patients who underwent upfront curative-intent pancreatectomy between 2004 and 2016 were included.

**Model development scheme**

Three schemes for model development and external validation were conducted. First, we utilized data from SEER for model development and data from KOTUS-BP for validation. Second, we swapped the roles of these two datasets for model development and evaluation. Finally, we mixed these two datasets half and half, and utilized the mixed datasets for model development and external validation.

The Cox proportional hazard (Cox-PH) model and the two ML models had different schemes for the model development process, as shown in Fig. 1, although we used all seven covariates in the Cox-PH model and both ML survival models. While the Cox-PH model was constructed without considering any hyperparameters, both RSF and SVM models require cross-validation (CV) to select the set of hyperparameters that build the best model. First, the model development dataset was divided into 10 subsets. For 10-fold CV, nine of the 10 subsets were used for the training set and the other subset was used for the validation set. The average Harrell C-index of the validation sets in a total of 10 iterations was calculated to compare the performance of models. The final model was then constructed with the entire model development dataset and the set of hyperparameters that resulted in the best average Harrell C-index during 10-fold CV.

**ML methods for survival analysis**

In prospective cohort studies, survival analysis has been useful to investigate the prognostic factors associated with the survival time and to predict disease processes. In traditional survival analysis, a survival prediction model has been constructed on the basis of demographic and clinicopathologic information. In recent years, there has been considerable interest in applying ML methods to predict the survival of cancer patients using a considerable amount of genomic information including traditional clinical covariates. An advantage of ML methods over the classical Cox regression models is their ability to model complicated associations between the survival time and risk factors, leading to better prediction. Unlike regression and classification settings, standard ML methods cannot be directly applied to censored survival data. With consideration of the censoring mechanism, several ML methods have been extended to survival data, such as bagging survival trees [2], RSF [3,4], SVM for survival analysis [5-8], and CoxBoosting [9]. Among these methods, RSF and SVM for survival analysis were used to develop prediction models for PDAC patients in this study.
Random survival forests

The RSF method is an extension of Breiman’s random forest method to right-censored survival data by using a forest of survival trees for prediction. Similar to regression and classification settings, RSF is an ensemble learner formed by averaging a tree base-learner. In survival settings, a binary survival tree is the base-learner, and the ensemble learner is formed by averaging each tree’s Nelson-Aalen’s cumulative hazard function.

There are four main steps in RSF: (1) Draw $B$ bootstrap samples randomly from the given dataset. Since one-third of the training set data is not present in the bootstrapping sample, this leftover data is known as the out-of-bag (OOB) data. (2) For each sample, construct a survival tree using a randomly selected subset of variables among all available variables, and split the nodes using the candidate variables that maximize the survival difference between child nodes. Here, the survival difference is measured by three criteria: the log-rank statistic, gradient-based Breier score, and log-rank score. (3) Grow the tree to the full size with the constraint that a terminal node should contain a certain number of unique uncensored patients. (4) For each terminal node, calculate the cumulative hazard function (CHF) based on the Nelson-Aalen estimator and take the ensemble CHF of the OOB data by averaging the CHF of each tree.

SVM for survival analysis

The SVM method of supervised learning has been very successful, mostly in classification and then extended to the regression problem. The main idea of SVM is to minimize the $\varepsilon$-insensitive loss function, $\max (0, |f(x_i) - y_i| - \varepsilon)$, with a regularization parameter. Here, $f(x_i)$, $y_i$, and $\varepsilon$ are the predicted value, the actual value, and the acceptable margin of error, respectively.

To take into account censored survival data, SVM for regression on the censored data (SVCR) has been proposed by imposing constraints on the SVM formulation for two comparable cases [5,6]. In other words, for censored data, the time to event after censoring is unknown and thus predictions greater than the censoring time are not required to be penalized. However, all survival predictions less than the censoring time are penalized, while uncensored data are treated in the same way as in the ordinary regression approach, since the exact event time is known. A prior study compared three types of SVM [8], including a regression approach, a ranking approach and a hybrid approach combining the regression and ranking approaches. All types of SVM share a common frame, but they differ in their objective function and constraints. In this paper, two types of SVM were considered: the SVCR model proposed by Shivashwamy et al. [5] and ranking support vector machines (RankSVMs) proposed by Van Belle et al.
These two models have been summarized in detail and compared elsewhere [7].

In the SVM model, overall survival time, $y$, is explained by the clinical variables $x$ as $y = \varphi(x) + \varepsilon$, where $\varphi(\cdot)$ is called the feature map. Since the feature map usually implies a higher-dimensional space, it is unusual to calculate the feature map itself. Instead, the feature map is directly calculated by kernel $k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$ for variable $x$ between patients $i$ and $j$, which is a consequence of Mercer’s theorem [10]. The entire process of training the model and generating predictions is simply carried out by using the kernel. The kernel plays a significant role in constructing SVM models, and various types of kernels are available, among which a linear kernel and a clinical kernel were considered. The linear kernel is given as $k(x_i, x_j) = x_i^T x_j$, whereas the clinical kernel proposed by Daemen and De Moor [11] is defined as the average of the kernel functions, $k(x_i, x_j)$, of all variables between patients $i$ and $j$. Here

$$k(x_i, x_j) = \frac{\text{max} - \text{min}}{\text{max} - \text{min}} | x_i - x_j |$$

for continuous and ordinal clinical variables and as $k(x_i, x_j) = \begin{cases} 1, & \text{if } x_i = x_j \\ 0, & \text{if } x_i \neq x_j \end{cases}$ for nominal clinical variables. The examples presented by Daemen and De Moor [11] show that this kernel better accounts for clinical data, which often have different scales in covariates, and the differences in values of nominal variables are not necessarily linearly correlated. The final kernel for clinical data is then the sum of the individual kernel matrices divided by the total number of clinical variables. This final kernel describes the similarity of a class of patients based on a set of variables of different types.

Although SVCR and RankSVMs share the same framework to a certain extent, they differ in terms of how they utilize information for their ultimate objective. SVCR is designed to directly predict the survival time and to minimize the absolute error between predicted and observed survival times. In contrast, RankSVMs focuses on predicting the correct ranking of survival times rather than predicting the actual survival time. In this respect, SVCR extends the standard support vector regression to censored data by penalizing incorrect predictions of censored observations [5,6], while RankSVMs takes into account the ranking problem for the censored data by minimizing the empirical risk of incorrectly ranking two observations.

### Survival prediction models

All statistical analyses were done using R version 3.6.2 (The R Foundation for Statistical Computing, Vienna, Austria). The only continuous covariable, age, was reported as the mean ± standard deviation, and the other categorical variables were reported as frequencies with percentages, as shown in Table 1.

Two Kaplan-Meier survival curves were compared using the log-rank test, as shown in Fig. 2. In addition, 5-year survival rates and median survival times were given. Variables with $p$-values less than

### Table 1. Basic statistics and 5-year overall survival rates for seven variables in the SEER and KOTUS-BP databases

<table>
<thead>
<tr>
<th>Variable</th>
<th>SEER database (n = 9,624)</th>
<th>KOTUS database (n = 3,281)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients</td>
<td>5-Year OS (%)</td>
</tr>
<tr>
<td>Age (yr)</td>
<td>65.6 ± 10.4</td>
<td>20.1</td>
</tr>
<tr>
<td>Female</td>
<td>4,755 (49.4)</td>
<td>21.3</td>
</tr>
<tr>
<td>Male</td>
<td>4,869 (50.6)</td>
<td>18.9</td>
</tr>
<tr>
<td>Head</td>
<td>8,079 (83.9)</td>
<td>19.2</td>
</tr>
<tr>
<td>Body/Tail</td>
<td>1,545 (16.1)</td>
<td>25.0</td>
</tr>
<tr>
<td>No adjuvant treatment</td>
<td>2,948 (30.6)</td>
<td>17.3</td>
</tr>
<tr>
<td>Adjuvant treatment</td>
<td>6,676 (69.4)</td>
<td>21.3</td>
</tr>
<tr>
<td>Well differentiated</td>
<td>1,013 (10.5)</td>
<td>37.4</td>
</tr>
<tr>
<td>Moderately differentiated</td>
<td>5,055 (52.5)</td>
<td>20.5</td>
</tr>
<tr>
<td>Poorly differentiated</td>
<td>3,556 (37.0)</td>
<td>14.6</td>
</tr>
<tr>
<td>T1</td>
<td>1,603 (16.7)</td>
<td>32.7</td>
</tr>
<tr>
<td>T2</td>
<td>5,830 (60.6)</td>
<td>18.8</td>
</tr>
<tr>
<td>T3</td>
<td>2,191 (22.7)</td>
<td>14.3</td>
</tr>
<tr>
<td>N0</td>
<td>3,155 (32.8)</td>
<td>32.4</td>
</tr>
<tr>
<td>N1</td>
<td>4,030 (41.9)</td>
<td>20.5</td>
</tr>
<tr>
<td>N2</td>
<td>2,439 (25.3)</td>
<td>14.6</td>
</tr>
</tbody>
</table>

Values are presented as mean±SD or number (%).

SEER, Surveillance, Epidemiology and End Results; KOTUS-BP, Korea Tumor Registry System-Biliary Pancreas; OS, overall survival.

*Log-rank test.
0.05 in the univariate Cox model were entered into a multivariate Cox proportional hazards model to estimate the hazard ratios (HRs) for the corresponding predictors, as shown in Fig. 3.

For the implementation of RSF, the number of binary decision trees, the maximum variables for splitting in each node, and the splitting rules for measuring survival differences are shown in Table 2. The number of trees was 50, 100, 200, 500, and 1,000, and the variables for splitting were given as 10. Although there were seven variables, three variables (histologic differentiation, AJCC 8th edition T-stage, and N-stage) had one more additional variable after one-hot encoding. Three different split rules were applied: log-rank splitting [12,13], gradient-based Brier score splitting [14], and log-rank score splitting [15]. As a result, 150 models for each dataset were constructed, consisting of a combination of the number of trees, the number of variables for splitting, and the split rules.

To implement SVM, 80 models were considered from combinations of various hyperparameters: two SVM models (SVCR and RankSVMs), two types of kernels (linear and clinical kernels), two ways of computing distance between data points (makediff1 and makediff3), and 10 values of the regularization parameter γ as shown in Table 3. The two arguments makediff1 and makediff3 are used in the R package survivalsvm, in which makediff1 computes the distance between two consecutive observations only when the first one is not censored, whereas makediff3 computes the difference between data point i and its neighbor that has the largest survival time that is smaller than the survival time of yi [8].

In total, 80 models were cross-validated and the model with the best validation C-index was chosen.

Advantages of ML methods over the Cox model

Based on three survival predictive models, we investigated personalized treatment policies using the survival rate over time. It is well known that the Cox model assumes a proportional HR over time, which implies that the HRs between different individuals are constant over time. However, the two ML models used in this study reflect more complex interactions between covariates and yield non-constant HRs between different individuals over time. For personalized treatment, it would be more desirable to predict the survival rate over time using the ML models than using the Cox model.

Results

Fig. 2 shows the two Kaplan-Meier survival curves for the SEER and KOTUS-BP datasets. The censoring fractions were 30.7% for SEER and 49.5% for KOTUS-BP. These two survival curves overlapped for up to 20 months and then significantly separated, with a p-value less than 1e-13 from the log-rank test for the equivalence of two survival curves. The median survival times were 21 months for SEER and 24 months for KOTUS-BP. Fig. 3 shows the estimated HRs of each clinical variable from the multivariate Cox model with 95% confidence intervals and p-values for both datasets. The

![Fig. 2. Kaplan-Meier survival curves with 5-year overall survival (OS) rates and median survival times for the Surveillance, Epidemiology and End Results (SEER) and Korea Tumor Registry System-Biliary Pancreas (KOTUS-BP) datasets.](https://doi.org/10.5808/gi.22036)
Fig. 3. Hazard ratios and 95% confidence intervals of seven variables in the Surveillance, Epidemiology and End Results (SEER) and Korea Tumor Registry System–Biliary Pancreas (KOTUS–BP) datasets. AJCC, American Joint Committee on Cancer.
estimated HRs of the seven variables were not meaningfully different between the datasets, and all the variables were significant, except for sex in KOTUS-BP dataset.

Table 4 shows the C-index values and 1-year, 2-year, and 3-year time-dependent AUCs for the final Cox, RSF, and SVM survival analysis models according to the three schemes of model development. The Cox, RSF, and SVM models performed somewhat better when the two datasets were mixed half and half than when the first and second schemes were applied. Although the results are not shown here, the performance of RankSVMs was exceptionally low, with C-index values of 0.548, 0.529, and 0.514 for the three schemes, respectively. This implies that the rank-based approach performed worse than the regression-based approach in this study.

Through 10-fold CV of 150 RSF models, the model with the best validation Harrell C-index was chosen as the final RSF model. The final RSF model consisted of 100 decision trees, used a maximum of two variables in splitting nodes, and used the log-rank test to measure survival differences when two datasets were mixed half and half. The C-index and 1-year, 2-year, and 3-year time-dependent AUCs were 0.6337, 0.6824, 0.6681, and 0.6781, respectively.

Similarly, the model with the best validation Harrell C-index was chosen through 10-fold CV of 80 SVM models. The final SVM model was the SVCR model based on an additive clinical kernel, regularization constant (γ) of 0.1, and the makediff3 method to calculate the distance between data points when the two datasets were mixed half and half. The C-index and 1-year, 2-year and 3-year time-dependent AUCs were 0.6233, 0.6849, 0.6352, and 0.6264, respectively.

The C-index and 1-year, 2-year, and 3-year time-dependent AUCs of the Cox model were 0.6434, 0.6976, 0.6795 and 0.6873, respectively. Comparing these values to those of the two ML survival models, the Cox model consistently performed slightly better than RSF and SVM models. The Cox model also yielded slightly better results when the two datasets were mixed half and half than when the two datasets were not mixed.

In order to consider personalized treatment policies, we compared the predictive survival curves of three different patients using the fitted Cox model and the final RSF model described above.

Table 4. C-index and 1-year, 2-year, 3-year time-dependent AUCs for the Cox, RSF, and SVM models according to three schemes

<table>
<thead>
<tr>
<th>Model</th>
<th>Training (SEER)</th>
<th>Training (KOTUS)</th>
<th>Training (SEER + KOTUS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C-index</td>
<td>Td1 AUC</td>
<td>Td2 AUC</td>
</tr>
<tr>
<td>Training (SEER)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cox</td>
<td>0.65417</td>
<td>0.72545</td>
<td>0.68776</td>
</tr>
<tr>
<td>RSF</td>
<td>0.66520</td>
<td>0.72960</td>
<td>0.70807</td>
</tr>
<tr>
<td>SVM</td>
<td>0.64218</td>
<td>0.72258</td>
<td>0.65812</td>
</tr>
<tr>
<td>Training (KOTUS)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cox</td>
<td>0.65074</td>
<td>0.69346</td>
<td>0.69524</td>
</tr>
<tr>
<td>RSF</td>
<td>0.66293</td>
<td>0.70624</td>
<td>0.71295</td>
</tr>
<tr>
<td>SVM</td>
<td>0.62668</td>
<td>0.66973</td>
<td>0.66769</td>
</tr>
<tr>
<td>Training (SEER + KOTUS)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cox</td>
<td>0.64890</td>
<td>0.70718</td>
<td>0.69108</td>
</tr>
<tr>
<td>RSF</td>
<td>0.66396</td>
<td>0.71328</td>
<td>0.72110</td>
</tr>
<tr>
<td>SVM</td>
<td>0.62538</td>
<td>0.69700</td>
<td>0.64029</td>
</tr>
</tbody>
</table>

AUC, area under receiver operating characteristic curve; RSF, random survival forests; SVM, support vector machines; C-index, Harrell’s concordance index; Td1, 1-year time-dependent; Td2, 2-year time-dependent; Td3, 3-year time-dependent; KoTUS, Korea Tumor Registry System-Biliary Pancreas; SEER, Surveillance, Epidemiology and End Results.
Suppose that the three chosen patients (A, B, and C) are all 50-year-old women, have a tumor in the body or tail of the pancreas, and have not received chemotherapy. Patient A has a well differentiated tumor staged T1 and N0 according to the AJCC 8th edition staging system. Patient B has a moderately differentiated tumor staged T2 and N1, whereas patient C has a poorly differentiated tumor staged T3 and N2. We plotted two predicted survival curves from both the Cox model and RSF model for these three patients over time, as shown in Fig. 4. The predicted survival curves from the Cox model showed relatively consistent differences among these three patients over time, whereas those from the RSF model showed less consistent differences. For example, the slope of the survival curve of patient C suddenly changed at 10 months after the diagnosis and the slope of patient B dramatically changed at 17 months after the diagnosis, whereas the slope of patient A did not change over time. Therefore, it seems that there is a discrepancy between the survival curves generated using the Cox and RSF models. This may imply that different treatment strategies for different patients would maximize treatment efficacy.

Discussion

In light of the development of a predictive survival model for PDAC [1], we considered a comparative study to investigate whether ML methods for survival analysis improve the predictability of the survival rate. In this study, both RSF and SVM methods for survival analysis were considered and compared with the Cox model [1] using the same SEER and KOTUS-BP datasets. In addition to the scheme used in the previous model [1], two other schemes were considered for model development and evaluation. In the second scheme, the roles of these two datasets were reversed, so that KOTUS-BP was used as the training set and SEER was used as the external validation set. In the third scheme, these two datasets were mixed half and half, and one of the mixed datasets was randomly chosen for model development and the remaining dataset was used for external validation. As shown in the Results section, the third scheme yielded slightly better performance for all methods than the other two schemes.

Compared with the Cox model, the performance of the ML survival models was not significantly improved, and RSF performed similarly to the Cox model. However, the performance of SVM differed substantially according to how the survival information was used. The performance of SVCR was comparable to those of the Cox model and RSF, since SVCR utilizes the survival time in the regression model considering the censoring mechanism. In contrast, the performance of RankSVMs was not good because this method only uses the ranking information of the survival times.

The RSF and SVM showed no substantial improvements in performance compared to the Cox model. In this study, only seven clinical variables were shared between the SEER and KOTUS-BP datasets, which might have been too few to maximize the usefulness of ML methods. ML methods are useful to analyze more complex and nonlinear associations among high-dimensional variables such as genetic information. It was also noted that the Harrell C-index of all models, both in the training set and in the test set, was less than 0.70, except for one or two cases.

Although it takes more time to develop ML survival models than a Cox model and there is no substantial performance improvement, these ML survival models have the advantage of allowing nonlinear risk to be predicted over time. As shown in Fig. 4, the trend in the survival curves for the RSF model was different from that for the Cox model. For example, the survival curves of the RSF model had the largest difference between patients B and C when 1 year to 2 years had passed. With this information, clinicians can pay particular attention to patient C in this period. Meanwhile, patients in a period with particularly high risk can be informed in advance, so that they could receive additional health care in that period. The fact that the RSF model outputs the survival curve of each individual might enable more patient-specific
care. Furthermore, ML survival models recommend whether a patient should receive treatment or not [16]. Adjuvant chemotherapy was found to be helpful for almost all of the patients in this study, but other treatments can be rather harmful to some patients.

**ORCID**

Hyunsuk Kim: https://orcid.org/0000-0001-7686-7119  
Taesung Park: https://orcid.org/0000-0002-8294-590X  
Jinyoung Jang: https://orcid.org/0000-0003-3312-0503  
Seungyeoun Lee: https://orcid.org/0000-0001-7941-8933

**Authors’ Contribution**

Conceptualization: SL, HK, JJ, TP. Data curation: JJ, HK. Formal analysis: HK. Funding acquisition: SL, TP. Methodology: SL, HK, TP. Writing - original draft: HK, SL. Writing - review & editing: TP, JJ.

**Conflicts of Interest**

Taesung Park serves as an editor of the Genomics and Informatics, but has no role in the decision to publish this article. All remaining authors have declared no conflicts of interest.

**Acknowledgments**

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2019R1F1A1062005) and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI16C2037).

**References**

A novel mutation in \textit{GJC2} associated with hypomyelinating leukodystrophy type 2 disorder

Sajad Rafiee Komachali$^{1,2}$, Mozghan Sheikholeslami$^1$, Mansoor Salehi$^2$

$^1$Department of Biology, University of Sistan and Baluchestan, Zahedan 98167-45845, Iran
$^2$Medical Genetics Research Center of Genome, Isfahan University of Medical Sciences, Isfahan 81759-54319, Iran

Hypomyelinating leukodystrophy type 2 (HLD2), is an inherited genetic disease of the central nervous system caused by recessive mutations in the gap junction protein gamma 2 (\textit{GJC2}/\textit{GJA12}). HLD2 is characterized by nystagmus, developmental delay, motor impairments, ataxia, severe speech problem, and hypomyelination in the brain. The \textit{GJC2} sequence encodes connexin 47 protein (Cx47). Connexins are a group of membrane proteins that oligomerize to construct gap junctions protein. In the present study, a novel missense mutation gene c.760G>A (p.Val254Met) was identified in a patient with HLD2 by performing whole exome sequencing. Following the discovery of the new mutation in the proband, we used Sanger sequencing to analyze his affected sibling and parents. Sanger sequencing verified homozygosity of the mutation in the proband and his affected sibling. The autosomal recessive inheritance pattern was confirmed since Sanger sequencing revealed both healthy parents were heterozygous for the mutation. PolyPhen2, SIFT, PROVEAN, and CADD were used to evaluate the function prediction scores of detected mutations. Cx47 is essential for oligodendrocyte function, including adequate myelination and myelin maintenance in humans. Novel mutation p.Val254Met is located in the second extracellular domain of Cx47, both extracellular loops are highly conserved and probably induce intramolecular disulfide interactions. This novel mutation in the Cx47 gene causes oligodendrocyte dysfunction and HLD2 disorder.

**Keywords:** connexin 47, \textit{GJC2}, hypomyelinating leukodystrophy type 2, Pelizaeus–Merzbacher-like disease

\section*{Introduction}

Hypomyelinating leukodystrophy type 2 (HLD2), is a genetic disorder of white matter, that nystagmus, progressive spasticity, developmental delay, motor impairments, ataxia, and hypomyelination on the brain are common symptoms in this disorder [1]. The clinical manifestation of HLD2 that also known Pelizaeus–Merzbacher-like disease (MIM #608804) is an insufficient amount of myelin deposition which is observable on brain magnetic resonance imaging (MRI). Primary defects in myelin synthesis and stability are the most common causes of the disease, but myelin damage may also be a factor [1,2]. The majority of individuals with severe hypomyelination begin in infancy or early childhood and develop significant neurological impairments; however, symptoms can also appear in adults [1].

The association between gap junction protein gamma 2 (\textit{GJC2}) gene and HLD2 has been previously reported, homozygous or compound heterozygous mutations in the
The GJC2 gene, earlier known as GJA12 (MIM #608803), is located on the long arm of the chromosome 1 (1q42.13) [4,5]. Connexin 47 (Cx47, GenBank NP 065168.2) is encoded by the GJC2 gene and is a member of a highly conserved protein family of connexins [4,5]. The connexins are a group of membrane proteins that form connexon in the cell membrane [6]. Two connexons, each containing six connexin proteins, link across the extracellular space to form a gap junction channel. Cell growth, regulation, and development can all be aided by gap junctions. Ions, intracellular metabolites, and messenger molecules (with a molecular weight of less than 1–2 kDa) can transfer from one cell's cytoplasm to its opposing neighbors via these channels [6-8].

In the present study, we report a novel missense homozygous GJC2 c.760G > A mutation (p.Val254Met) in two cases of HLD2 in a consanguineous family, broadening the range of HLD2-causing GJC2 mutations. Whole exome sequencing (WES) was used on the DNA sample from the proband who is the affected boy, to identify the mutated gene. Sanger sequencing was used to confirm the discovered mutation in the proband and to investigate this mutation in healthy parents and another affected child.

**Methods**

Genomic DNA was extracted from the peripheral blood of the patients and their parents, using innuPREP Blood DNA Mini Kit (Analytika, Jena, Germany) according to the manufacturer’s protocol.

The sample was under WES using Genome Analyzer HiSeq 4000 (101-bp paired-end reads and 100× depth of coverage; Illumina, San Diego, CA, USA) following the manufacturer’s instructions. The library was generated using SureSelect XT Library Prep Kit (Agilent Technologies, Santa Clara, CA, USA). IllumQC tool used for QC of sequencing data. Sequenced reads were aligned to the GRCH37/UCSC hg19 human reference Genome. Also, the post-alignment processing step includes base quality score recalibration (BQSR) was done before variant calling. Then variant calling and VQSR filtering steps were performed, and in the last part of the primary analysis, we annotated the VCF file through ANNOVAR. Exome sequencing identified 365,494 annotated variants. To find the disease-causing variation, a filtering FASTQ was set up. Variant filtering is a secondary next-generation sequencing analysis step that includes the stages shown in Fig. 1. Due to eliminating benign variants, only variations with a frequency of less than 1% were chosen. Exome Sequencing Project (ESP), 1000 Genomes (1000G), and Exome Aggregation Consortium (ExAC) data were used to cross-verify the frequencies of discovered variations. We applied a neuromuscular-designed panel to filter the remaining variants. This panel is created by using different databases like CeGat, Fulgent, CENTOGENE, and DisGeNET. We were able to reduce the number of candidate variants to around 16 after applying the filtration steps (Table 1). All 16 variants were checked in OMIM for this study. The variants are then classed as benign, likely benign, a variant of uncertain significance, likely pathogenic, or pathogenic using the American College of Medical Genetics/Association for Molecular Pathology criteria.

Clinical information from a physical examination, laboratory tests or imaging, segregation analysis, genotyping and phenotype correlation, previous publication, or a de novo assessment of the variation were all used to assess the variants. Many of these variants may be ruled out simply by checking the phenotype; in most cases, the diagnosis is clinical.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Position/Variant</th>
<th>Zygosity</th>
<th>Inheritance</th>
<th>Disease</th>
<th>Allele frequency (gnomAD)</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSPG2</td>
<td>chr1:22183875:G:A</td>
<td>het</td>
<td>AR</td>
<td>Dyssegmental dysplasia, Silverman-Handmaker type</td>
<td>0.000584</td>
<td>VUS by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td></td>
<td>NM_001291860:exon43:c.</td>
<td></td>
<td></td>
<td>Schwartz-Jampel syndrome, type 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C5300T:p.A1767V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RYR2</td>
<td>chr1:237777434:A:G</td>
<td>het</td>
<td>AD</td>
<td>Arrhythmogenic right ventricular dysplasia 2</td>
<td>0.0000161</td>
<td>VUS by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td></td>
<td>NM_001035:exon37:c.</td>
<td></td>
<td></td>
<td>Ventricular arrhythmias due to cardiac ryanodine receptor calcium release deficiency syndrome</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A500665:p.N1669S</td>
<td></td>
<td></td>
<td>Ventricular tachycardia, catecholaminergic polymorphic, 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCM10</td>
<td>chr10:13230915:C:T</td>
<td>het</td>
<td>AR</td>
<td>Immunodeficiency 80 with or without cardiomypathy</td>
<td>0.013</td>
<td>Benign by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td></td>
<td>NM_018518:exon10:c.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C12507:p.A417V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEST1</td>
<td>chr11:61722577:AGGCTGG----</td>
<td>het</td>
<td>AD</td>
<td>Microcornea, rod-cone dystrophy, cataract, and posterior staphyoma 2</td>
<td>This variant does not have a gnomAD exomes entry.</td>
<td>Pathogenic by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td></td>
<td>NM_001363592:exon3:c.</td>
<td></td>
<td></td>
<td>Bestrophinopathy, AR</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Retinitis pigmentosa, concentric</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Retinitis pigmentosa-50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Vitreoretinocchorioidopathy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPN1</td>
<td>chr11:64953448:G:A</td>
<td>het</td>
<td>AR</td>
<td>Spastic paraplegia 76, AR</td>
<td>0.000446</td>
<td>Pathogenic by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td></td>
<td>NM_001988686:exon5:c.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G517A:p.G173R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WNK1</td>
<td>chr12:974308:-C</td>
<td>het</td>
<td>AR</td>
<td>Neuropathy, hereditary sensory and autonomic, type II</td>
<td>0.248</td>
<td>Benign by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td></td>
<td>NM_213655:exon9:c.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R7266K*45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>THSD1</td>
<td>chr13:52951802:T:C</td>
<td>het</td>
<td>AD</td>
<td>Aneurysm, intracranial berry, 12</td>
<td>0.0429</td>
<td>Benign by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td></td>
<td>NM_199263:exon4:c.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A21446:p.K715R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATM</td>
<td>chr15:45660304:T:G</td>
<td>het</td>
<td>AR</td>
<td>Cerebral creatine deficiency syndrome 3</td>
<td>0.0000239</td>
<td>VUS by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td></td>
<td>NM_001482:exon4:c.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A603C:p.K201N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTU2</td>
<td>chr16:88778615:G:A</td>
<td>het</td>
<td>AR</td>
<td>Microcephaly, facial dysmorphism, renal agenesis, and ambiguous genitalia syndrome</td>
<td>0.562</td>
<td>Benign by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td></td>
<td>NM_001318507:exon6:c.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G490A:p.V164M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERCC2</td>
<td>chr19:45864832:G:A</td>
<td>het</td>
<td>AR</td>
<td>Cerebrooculofacioskeletal syndrome 2</td>
<td>0.0000278</td>
<td>VUS by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td></td>
<td>NM_000130867:exon11:c.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C1115T:p.P372L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NBAS</td>
<td>chr2:15614274:G:A</td>
<td>het</td>
<td>AR</td>
<td>Infantile liver failure syndrome 2</td>
<td>0.0000239</td>
<td>VUS by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td></td>
<td>NM_015909:exon15:c.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C15165:p.R506W</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCCB</td>
<td>chr3:136046527:A:T</td>
<td>het</td>
<td>AR</td>
<td>Propionicacidemia</td>
<td>This variant does not have a gnomAD exomes entry.</td>
<td>Likely Pathogenic by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td></td>
<td>NM_000532:exon13:c.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A1351T:p.T451S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued to the next page)
cases, the reported clinical phenotype of the gene and the patient phenotype are unrelated. It should be emphasized that for autosomal recessive disorders, only homozygous variants were acceptable.

Sanger sequencing was performed by ABI prism 3730 sequencer (Applied Biosystems, Waltham, MA, USA) to validate the pathogenic mutation and segregation the mutation in this family. Mutation Surveyor program version 5.1.2 was used to analyze the sequences (SoftGenetics, State College, PA, USA). Function prediction scores of identified mutation were assessed by PolyPhen2 \[9\], SIFT \[10\], PROVEAN \[11\], and CADD \[12\].

Clinical report
The present study involves a Caucasian consanguineous family with two affected children by HLD2 who have been diagnosed with medical evaluations which are standard procedures in individuals with neurological diseases such as MRI. Local ethics committees obtained informed consent from the subjected family.

Patient 1
Patient 1 is our proband, and the symptoms of the disease appeared in him with developmental delay. He began occupational therapy at the age of four months and was walking and talking by the age of six. After the age of seven, he began to regress. Since eight years old, he has been unable to walk and can only sit and talk in a few phrases. Nystagmus and vision problems are also issues for him. So far, he hasn’t had any seizures.

Patient 2
Recurrent seizures, nystagmus, poor vision, and developmental delay were among the signs of the disease, which first appeared when the daughter was three months old. By the age of five with occupational treatment, she was able to sit and pronounce a few words. However, she has regressed and is presently unable to function as a result of recurrent seizures. She has lost her ability to walk, speak, swallow food, and roll. She now experiences seizures every day that are uncontrollable by medication. Seizures become more severe as time passes.

Results
Magnetic resonance imaging
MRI of the affected boy at the age of 7 years old and the affected girl at the age of 12 years old showed abnormal signal in both cerebral hemispheres white matter high on T2, without mass effect or volume loss mostly due to leukodystrophy. The lateral, 3rd, and 4th ventricles are normal in size and shape midline shift. The brain stem and both cerebellar hemispheres are intact. As well as in the affected girl, a large retention cyst/polyp in the left maxillary sinus is shown in the MRI image (Fig. 2A, B).

Molecular analyses
Performing WES on proband identified a novel homozygous c.760G > A mutation in the GJC2 gene. AD—Allele Depth (Read depth for each allele) is 0, 116 at c.760G > A mutation, which means read for the reference allele = 0 and the alternative allele = 116. Also, DP—Read Depth is 116 (DP = 116). Sanger sequencing confirmed homozygosity of c.760G > A mutation in the proband and his affected sister and heterozygous of this mutation in her parents, suggesting it as the putative disease-causing mutation, and autosomal recessive inheritance pattern (Fig. 2C, D).

Table 1. Continued

<table>
<thead>
<tr>
<th>Gene</th>
<th>Position/Variant</th>
<th>Zygosity</th>
<th>Inheritance</th>
<th>Disease</th>
<th>Allele frequency (gnomAD)</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP4M1</td>
<td>chr7:99704060:G:T NM_001363671:exon14:c. G1081T:p.A361S</td>
<td>het</td>
<td>AR</td>
<td>Spastic paraplegia 50, autosomal recessive</td>
<td>This variant does not have a gnomAD exomes entry.</td>
<td>VUS by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td>CLCN1</td>
<td>chr7:143048771:C:T NM_000083:exon23:c. C2680T:p.R894X</td>
<td>het</td>
<td>AR</td>
<td>Myotonia congenita, dominant Myotonia congenita, recessive</td>
<td>0.00288</td>
<td>Benign by VARSOME Pathogenic by FRANKLIN</td>
</tr>
<tr>
<td>RIMS2</td>
<td>chr8:105105846:C:T NM_001348498:exon19:c. C2992T:p.R998C</td>
<td>hom</td>
<td>AR</td>
<td>Cone-rod synaptic disorder syndrome, congenital nonprogressive</td>
<td>0.201</td>
<td>Benign by VARSOME and FRANKLIN</td>
</tr>
<tr>
<td>GJC2</td>
<td>chr1:228346219:G:A NM_020435:exon2:c. G760A:p.V254M</td>
<td>hom</td>
<td>AR</td>
<td>Spastic paraplegia 44, AR Leukodystrophy, hypomyelinating, 2 Lymphatic malformation 3</td>
<td>This variant does not have a gnomAD genomes entry.</td>
<td>VUS by VARSOME Likely Pathogenic by FRANKLIN</td>
</tr>
</tbody>
</table>

het, heterozygous; AR, autosomal recessive; AD, autosomal dominant; hom, homozygous; VUS, variant of uncertain significance.
There is no report of the c.760G > A mutation at the GJC2 gene in ExAC, 1000G, and other control datasets. We have submitted c.760G > A mutation to ClinVar database, the accession number is SCV001911448. Multiple sequence alignment of the GJC2 gene by Polyphen2 revealed that Val at position 254 is highly conserved among species. The PolyPhen2 [9], SIFT [10], PROVEAN [11], and CADD [12] results all confirmed that the p.Val254Met mutation is destructive and pathogenic (Table 2).

Fig. 2. Clinical and molecular features. (A) Axial T2W patient1 (the affected boy). (B) Axial T2W patient2 (the affected girl). Images show hypomyelination around basal ganglia in both patients. (C) Pedigree of the family. (D) The DNA sequences of the patient 1, patient 2, and their parents. Parents are heterozygote for c.760G>A and both patients are homozygote. The mutated allele is shown with arrows.

Table 2. Prediction of p.Val254 Met substitution effect

<table>
<thead>
<tr>
<th>Prediction tools</th>
<th>Score</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polyphen2 HumDiv</td>
<td>1</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>Polyphen2 HumVar</td>
<td>1</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>Provean</td>
<td>-2.73</td>
<td>Deleterious</td>
</tr>
<tr>
<td>SIFT</td>
<td>0.002</td>
<td>Damaging</td>
</tr>
<tr>
<td>CADD-phred score</td>
<td>26.3</td>
<td>Deleterious</td>
</tr>
</tbody>
</table>
Discussion

Our patient's novel mutation, p.Val254Met affects a highly conserved position in the second extracellular loop of Cx47 that has four transmembrane domains, two extracellular and three cytoplasmic [6]. It is worth noting that the position of this mutation could change the destiny of protein. It could interrupt the oligomerization of a complete gap junction. Based on ClinVar, other reported missense mutations in the extracellular loops of the Cx47 were depicted in figure 3, which have been associated with HLD2 disease (Fig. 3). Cx47, a significant protein for which crystallography has not yet been performed. Identifying any new mutation in Cx47 can thus be a crucial step toward a better understanding of the protein function.

The remarkable point is that the Cx47 is highly expressed in oligodendrocytes. Oligodendrocytes are a type of glial cells that are responsible for myelination in the central nervous system (CNS) [8,13],. Saltatory conduction is enabled by myelin sheaths, which contribute to the speed-up of action potential conduction in neurons. Also providing energy substrates to neurons is another essential function of oligodendrocytes [14,15].

The heterotypic coupling of Cx30–Cx32 and Cx43–Cx47 forms functional channels between astrocytes and oligodendrocytes [16], and this glial gap junction coupling is crucial for oligodendrocytes function, including adequate myelination and maintenance of myelin in humans [7,17]. Cx47 mutations that cause HLD2 are loss-of-function mutations that probably disrupt or alter the ability to generate functional channels with Cx43, indicating that HLD2 is caused by a loss of oligodendrocytes/astrocytes coupling mediated by Cx47/Cx43 channels [17].

Moreover, in previous studies, the discovery of the altered gap junction protein in the endoplasmic reticulum as a result of I33M mutation in the Cx47 gene has been suggested as a possible contributor to the pathogenic process of CNS hypomyelination [18]. Also, a transgenic mouse model of Pelizaeus-Merzbacher-like disorder was carrying the human M283T missense mutation in GJC2. It has been shown that in oligodendrocytes, expression of the homozygous mutant Cx47 gene leads to a complicated and diverse neuropathologic phenotype [19]. Last of all, abnormalities in the expression and distribution of the brain connexin like Cx47

![Fig. 3. Connexin 47 protein (Cx47) sequences. Protein sequences of Cx47 derived from UniProt (Q5T442) and the sequences of extracellular loops are highlighted with yellow. Based on ClinVar, mutations in both extracellular loops of Cx47 that cause hypomyelinating leukodystrophy type 2 were shown. The novel mutation p.Val254Met described in this study is shown within the circle.](https://doi.org/10.5808/gi.22008)
play a particular function in the neuropathologies [15].

In Conclusion, in this study, a novel missense mutation, p.Val254Met in the GJC2 gene, was identified as the cause of HLD2 in two children from a consanguineous family. This gene encodes a gap-junction protein involved in the oligodendrocyte process. This is the first report of p.Val254Met in the GJC2 gene as responsible for the HLD2. It may be beneficial to investigate this mutation at the protein level. In vivo research can also shed light.

**ORCID**

Sajad Rafiee Komachali: https://orcid.org/0000-0002-0326-711X
Mozhgan Sheikholeslami: https://orcid.org/0000-0003-0243-6142
Mansoor Salehi: https://orcid.org/0000-0002-6565-0907

**Authors’ Contribution**

Conceptualization: MS (Mozhgan Sheikholeslami), SRK. Data curation: SRK. Formal analysis: MS (Mozhgan Sheikholeslami), SRK. Funding acquisition: MS (Mansoor Salehi). Methodology: MS (Mozhgan Sheikholeslami), SRK. Writing - original draft: MS (Mozhgan Sheikholeslami). Writing - review & editing: MS (Mansoor Salehi).

**Conflicts of Interest**

No potential conflict of interest relevant to this article was reported.

**Acknowledgments**

We are thankful to the patients and their parents who participated in this study.

**References**


Instructions for authors

Enacted January 2003
Recently revised January 9, 2019

Genomics & Informatics (Genomics Inform) is owned and published by the Korea Genome Organization (KOGO). It is published four times per year (Mar, Jun, Sep, and Dec) in an online version. Genomics & Informatics welcomes high-quality research papers presenting novel data on the topics of gene discovery, comparative genome analyses, molecular and human evolution, informatics, genome structure and function, technological innovations and applications, statistical and mathematical methods, cutting-edge genetic and physical mapping, and DNA sequencing and other reports that present data where sequence information is used to address biological concerns. The journal publishes papers based on original research that are judged after editorial review to make a substantial contribution to the understanding of any area of genomics or informatics. Only manuscripts written in English under the Genomics & Informatics author guidelines are accepted. Genomics & Informatics follows the open access journal policy. All of the content of Genomics & Informatics is freely available online. Digital files can be read, downloaded, and printed without charge.

Manuscripts for submission to Genomics & Informatics should be prepared according to the following instructions. Genomics & Informatics follows the Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals (http://www.icmje.org) from ICMJE and Principles of Transparency and Best Practice in Scholarly Publishing (joint statement by COPE, DOAJ, WAME, and OASPA; (http://doaj.org/bestpractice) if otherwise not described below.

Research and publication ethics

For the policies on research and publication ethics that are not stated in these instructions, the Good Publication Practice Guidelines for Medical Journals (https://www.kamje.or.kr/board/view?b_name=bo_publication&bbo_id=7) and the Guidelines on Good Publication (http://publicationethics.org/resources/guidelines) can be applied. The Editor-in-Chief reserves the right to reject manuscripts that do not comply with the below requirements. The author will be held responsible for false statements or failure to fulfill the below requirements.

Statement of Informed Consent
Copies of written informed consent and Institutional Review Board (IRB) approval for clinical research should be kept. If necessary, the editor or reviewers may request copies of these documents to resolve questions about IRB approval or study conduct.

Statement of Human and Animal Rights
All human investigations must be conducted according to the principles expressed in the Declaration of Helsinki. All studies involving animals must state that the guidelines for the use and care of laboratory animals of the authors’ institution, or of any national law, were followed. Registration of clinical trial research: Any research that deals with a clinical trial should be registered with the primary national clinical trial registry site, such as the Korea Clinical Research Information Service (CRiS, http://cris.nih.go.kr), other primary national registry sites accredited by the World Health Organization (http://www.who.int/ictrp/network/primary/en/), or ClinicalTrials.gov (http://clinicaltrials.gov/), a service of the United States National Institutes of Health.

Authorship
Authorship credit should be based on 1) substantial contributions to conception and design, acquisition of data, and/or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; 3) final approval of the version to be published; and 4) agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Every author should meet all of these four conditions. After the initial submission of a manuscript, any changes whatsoever in authorship (adding author(s), deleting author(s), or re-arranging the order of authors) must be explained by a letter to the editor from the authors concerned. This letter must be signed by all authors of the paper. Copyright assignment must also be completed by every author.

Corresponding author and first author
It does allow multiple corresponding authors for one article. Only one author should correspond with the editorial office. It does accept notice of equal contribution for the first author when the study was clearly performed by co-first authors.

Correction of authorship after publication
It does not correct authorship after publication unless a mistake
has been made by the editorial staff. Authorship may be changed before publication but after submission when an authorship correction is requested by all of the authors involved with the manuscript.

**Conflict of Interest Statement**

The corresponding author must inform the editor of any potential conflicts of interest that could influence the authors’ interpretation of the data. Examples of potential conflicts of interest are financial support from or connections to pharmaceutical companies, political pressure from interest groups, and academically related issues. In particular, all sources of funding applicable to the study should be explicitly stated. As a guideline, any affiliation associated with a payment or financial benefit exceeding $10,000 per annum or 5% ownership of a company or research funding by a company with related interests would constitute a conflict that must be declared. This policy applies to all submitted research manuscripts and review material.

**Originality and Duplicate Publication**

No part of the accepted manuscript should be duplicated in any other scientific journal without the permission of the Editorial Board. If duplicate publication or plagiarism related to the papers of this journal is detected, the authors will be announced in the journal, their institutes will be informed, and the authors will be penalized. All submitted manuscripts are screened by CrossCheck (Similarity Check), a plagiarism detection program provided by iThenticate. The authors assure that no substantial part of the work has been published or is being considered for publication elsewhere. When any of the results is to appear in another journal, details must be submitted to the Editor-in-Chief, together with a copy of the other paper(s) and the expected date(s) of publication.

**Secondary Publication**

It is possible to republish manuscripts if the manuscripts satisfy the condition of secondary publication of the Uniform Requirements for Manuscripts Submitted to Biomedical Journals by the International Committee of Medical Journal Editors (ICMJE), available from http://www.icmje.org/. These are:

- The authors have received approval from the editors of both journals (the editor concerned with the secondary publication must have access to the primary version).
- The priority for the primary publication is respected by a publication interval negotiated by editors of both journals and the authors.
- The paper for secondary publication is intended for a different group of readers; an abbreviated version could be sufficient.
- The secondary version faithfully reflects the data and interpretations of the primary version.
- The secondary version informs readers, peers, and documenting agencies that the paper has been published in whole or in part elsewhere—for example, with a note that might read, “This article is based on a study first reported in the [journal title, with full reference]”—and the secondary version cites the primary reference.
- The title of the secondary publication should indicate that it is a secondary publication (complete or abridged republication or translation) of a primary publication. Of note, the United States National Library of Medicine (NLM) does not consider translations to be “republications” and does not cite or index them when the original article was published in a journal that is indexed in MEDLINE.

**Process to manage research and publication misconduct:** When the Journal faces suspected cases of research and publication misconduct, such as a redundant (duplicate) publication, plagiarism, fabricated data, changes in authorship, undisclosed conflicts of interest, an ethical problem discovered with the submitted manuscript, a reviewer who has appropriated an author’s idea or data, complaints against editors, and other issues, the resolving process will follow a flowchart provided by the Committee on Publication Ethics (http://publicationethics.org/resources/flowcharts). The discussion and decision on suspected cases are done by the Editorial Board of *Genomics & Informatics*.

**Preparation of manuscripts**

**General requirement**

Authors are recommended to keep the length of papers below 10 printed pages (30 typed pages of manuscript, including figures and tables) for original articles, four printed pages for research communications, and two printed pages (approximately 1,400 words or 1,000 words plus one figure) for application notes. All sections of the typescript should be double-spaced on one side of A4 paper (210 × 297 mm), and all pages must be numbered in order.

**Manuscript type**

**Original articles**

Original research articles are full scientific reports of original research. The manuscript should be organized as follows: Title Page, Abstract & Keywords, Introduction, Methods, Results, Discussion, Acknowledgments, References, Tables, and Figure Legends. The Results and Discussion can be combined.
Application notes
Application notes are short communications about novel software, new algorithm implementations, databases, and network services (web servers and interfaces). The manuscripts include the following: Title Page, Abstract & Keywords, Availability, Introduction, Main Text, References, and Supplementary Information.

Clinical genomics
Clinical genomics is for a short report of all kinds of genome analysis data from clinical fields, such as cancer, diverse complex diseases, and genetic diseases. Especially, Genomics & Informatics would encourage submitting cancer panel analysis data for a single cancer patient or a group of patients. Genomics & Informatics also would encourage depositing genome data into the Genomics & Informatics database. The manuscript should be organized as follows: Title Page, Abstract & Keywords, Introduction, Methods, Results, Discussion, Acknowledgments, References, Tables, and Figure Legends. The Introduction, Methods, Results, and Discussion can be combined.

Genome archives
Genome Archives is for a short manuscript announcing the genetic information of recently sequenced prokaryotic and eukaryotic genomes. Genomics & Informatics would encourage depositing the genome data into the Genomics & Informatics database. These genome archive data can make the rationale for sequencing a specific organism. The manuscripts include the following: Title Page, Abstract & Keywords, Introduction, Main Text, References, Tables, and Figure Legends.

Letters to the editor
Critical comments are welcomed for correcting errors of published facts and for providing alternative interpretations of published data. The sequence for a Letter to the Editor is the following: Title Page, Text, References, and Names and Affiliations of Authors. If needed, tables and figures can be included. A Letter to the Editor should not be longer than a printed page.

Review articles
Review Articles are usually solicited by the Editor-in-Chief. Authors wishing to prepare a review article should contact the Editor-in-Chief to discuss the suitability of the subject for the journal. There is no specific requirement for subsections of the body text of the paper.

Opinions / Commentaries
An opinion or commentary piece is a short article that conveys the author’s viewpoint on a research publication, including interpretation of data, value of methods used, and strengths/weaknesses, regarding any topic relevant to the field of research. Opinion (or commentary) articles provide insight, interpretation, and evaluation of specific issues, within the scope of the journal. Opinions should explain the implications of the article and describe the most important conclusions of the paper they are commenting on, highlight controversial issues, mention the strengths and weaknesses of the paper, highlight the presenter’s omission of key facts, and mention supporting arguments that would create a stronger presentation. Opinions are relatively short articles, around 1000 words, allowing maximum freedom of authors’ viewpoints, and are peer-reviewed. The articles are copyedited, citable, published in both PDF and HTML formats, and submitted for indexing in digital archives (e.g., PubMed Central). Authors are not required to pay a fee to publish an opinion (or commentary) article. Commentaries have no set format beyond the basic building blocks of a regular article, i.e., title, manuscript text, subheadings as needed, references, and author information.

Minireviews
Minireview articles are similar to review articles, except for their word limit and references. Minireviews focus on clearly defined topics of current interest, and recent developments in specific fields. Therefore, they offer a fast and easy means to keep abreast of exciting new developments and/or concepts. The word limit for minireview articles is 1000 words (or 2 double-spaced pages), with no more than 30 references. Minireview articles are peer-reviewed, copyedited, citable, published in both PDF and HTML formats, and submitted for indexing in digital archives, such as PubMed Central. Authors are required to pay a fee to publish a minireview.

Research communications
Research communication (RC) intends to deliver significant scientific discovery with broad interest in a short format. RCs may contain unstructured main text that includes introduction, results and discussion. RCs typically have no more than 2 display items (figures and tables) and the main text (not including abstract, references, tables and figure legends) is limited to 1,500 words. RCs may have online supplementary section.

Manuscript Format
Title
The title page should include (1) the full names of all authors with their Open Researchers and Contributors ID (ORCID), and the name(s) and address(es) of the institution(s) at which the work was carried out; (2) the telephone and fax numbers, and the
E-mail address of the corresponding author; and (3) a running title of no more than 50 characters, including spaces. Place an asterisk (*) after the corresponding author.

**Abstract**
The abstract should be unstructured and a single paragraph of fewer than 250 words. References should not be cited in the abstract. Six or fewer keywords should be appended to the abstract in alphabetical order. When possible, the keywords should be those found in the Medical Subject Headings of Index Medicus.

**Main text:**
All papers should be divided into the following sections and appear in this order:

1. **Introduction:** The paper begins with an introduction without subheadings that reviews the literature and states and justifies the purpose of the research.

2. **Methods:** This section should contain sufficient detail so that all procedures can be repeated, in conjunction with the cited references. The manufacturer and model number should be stated in this section—for example, as Sigma Chemical Co. (St. Louis, MO, USA).

3. **Results:** This section should describe the results of the experiments. Extensive interpretation should be reserved for the Discussion section. The results should be presented as concisely as possible. Footnotes should not be used and will be transferred to the text. Gene symbols should be italicized; protein products are not italicized.

4. **Discussion:** This section should provide an interpretation of the results in relation to previously published work and to the experimental system at hand. The Results and Discussion may be combined.

5. **Acknowledgments:** Information concerning the sources of financial support should be included in the acknowledgments.

**Authors’ contribution**
If the number of authors is equal to or greater than two, the authors’ roles should be described according to their specific role. *Genomics & Informatics* participates in the CRediT standard for author contributions. The contributions of all authors must be described using the CRediT Taxonomy of author roles. For each of the categories below, please enter the initials of the authors who contributed in that category. If listing more than one author in a category, separate each set of initials with a space. If no one contributed in a category, you may leave that box blank. The corresponding author is responsible for completing this information at submission, and it is expected that all authors will have reviewed, discussed, and agreed to their individual contributions ahead of this time.

- Conceptualization: AB
- Data curation: EFG
- Formal analysis: AB
- Funding acquisition: CD
- Methodology: AB, CD, EFG
- Writing – original draft: AB, EFG
- Writing – review & editing: AB, CD, EFG

**Reference**
The references should include only articles that are published or in press. Unpublished data, submitted manuscripts, abstracts, and personal communications should be cited within the text only. References are to be numbered in the order of citation within the article in brackets. References with up to six authors must list all names; for more than six authors, the first six names should be listed, followed by “et al.” Journal name titles should be abbreviated in accordance with the NLM Catalog, available from: https://www.ncbi.nlm.nih.gov/nlmcatalog/journals, or the ISO 4 standard, available from: http://www.issn.org/services/online-services/access-to-the-ltwa/?letter=a.

Examples of references are given below:

**Journal article**

**Books**

**Book sections**

**Online document**
- Puniyani AR, Lukose RM. Growing random networks under

Conference paper

Dissertation/Thesis

Tables and figures
Figure legends and tables should be included in the submitted manuscript as separate sections and should be formatted following the style of the journal. Each figure legend should have a brief, separate title that describes the entire figure without citing specific panels. The manuscript should be submitted with a set of figures of sufficient quality for reviewers to judge the data. All figures may be provided in color for the electronic version of the journal, even if the print version is in black and white. Figures will be printed in color only when in the reviewers' opinions the color is essential.

Photographs and illustrations should be of professional quality. Images should be provided as TIFF files. JPEG is also acceptable when the original format is JPEG. Each figure must be of 300 dpi or higher resolution with good contrast and sharpness. If a figure is to be reduced, all elements, including labels, should be able to withstand reduction and remain legible. Electron and light microscopic figures must be original or scanned copies from the original. The magnification should be indicated on each micrograph with a scale bar.

Tables are to be organized in portrait view and may run, if necessary, to subsequent pages in the vertical direction only. Tables should be designed for printing within two (17.5 cm) columns of width in no less than 10-point font and should not exceed more than the width of a journal page. If a table does not fit into this format, consider shortening row or column labels, using more than one table to display the data, eliminating unnecessary data, or converting table data into a figure or transferring part of the table data to the supplement.

Scientific names
The full formal Latin name for a taxon (e.g., Homo sapiens) should be provided the first time that the taxon is mentioned and should be italicized. In subsequent sentences, the scientific name of all taxa in the same genus should be abbreviated to the first initial of the generic name and the species name (e.g., H. sapiens), except where this usage creates confusion or ambiguity. When common names are used, the scientific name should be provided the first time the taxon is mentioned in the abstract and again the first time that taxon is mentioned in the main manuscript [e.g., “red pine (Pinus densiflora)...”]. Other taxonomic designations (e.g., family names) should not be italicized, and common names should not be capitalized.

Units and equations
Standard metric units should be used for describing length, height, weight, and volume. The unit of temperature is given in degrees Celsius (°C). All others are in terms of the International System of Units (SI). All unit symbols must be preceded by one space except percentage (%) and temperature (°C). All equations should be numbered in Arabic numerals.

Abbreviations
Abbreviations must be used as an aid to the reader, rather than as a convenience of the author, and therefore, their use should be limited. Generally, avoid abbreviations that are used less than 3 times in the text, including the tables and figure legends. In addition to abbreviations for SI units, common molecular, chemical, immunological, and hematological terms can be used without definition in the title, abstract, text, tables, and figure legends—e.g., bp, kb, kDa, DNA, cDNA, RNA, mRNA, and PCR. Other common abbreviations are as follows (the same abbreviations are used for plural forms): h (hour; use 0-24:00 h for time), s (second), min (minute), day (not abbreviated), week (not abbreviated), month (not abbreviated), year (not abbreviated), L (liter), mL (milliliter), μL (microliter), g (gram), kg (kilogram), mg (milligram), μg (microgram), ng (nanogram), pg (picogram), g (gravity; not ×g), n (sample size), SD (standard deviation of the mean), and SE (standard error of the mean).

Supplementary materials
Supplementary materials can be provided to support and enhance scientific information. Supplementary files offer additional possibilities for publishing supporting applications, sequence alignment, background datasets, microarray hybridization experiments, high-resolution images, movies, sound clips, and more. Supplementary files will be published alongside the online version of the article on the Genomics & Informatics web site. This material will not be edited or formatted; thus, the authors are responsible for the accuracy and presentation of all such material. Accepted file formats for supplementary materials:
- Quick Time files (.mov)
• Graphical image files (.gif)
• HTML files (.html)
• MPEG movie files (.mpg)
• JPEG image files (.jpg)
• Sound files (.wav)
• Plain ASCII text (.txt)
• Acrobat files (.pdf)
• MS Word documents (.doc)
• Postscript files (.ps)
• MS Excel spreadsheet documents (.xls)
• PowerPoint (.ppt)
• TeX and LaTeX

File sizes must be as small as possible, for quick downloading. 
Recommended specifics are:
• Videos
  - File size: <150 MB
  - Frame rate: 30 frames per second
  - Field order: none (progressive, not interlaced)
  - Aspect ratio: widescreen 16:9
  - Video codec: H.264
  - Video bitrate: 2 Mbps
  - Audio codec: AAC
  - Audio bitrate: 128 kbps
• Images
  - Frame size: 300 dpi in resolution
  - Frame rate: 300 dpi in resolution and 10-15cm in width

Please seek advice from the editorial office before sending files larger than our recommended size to avoid delays in publication.

Accession numbers
Please provide accession numbers for any new data (SNPs, gene sequences, protein sequences, CNVs, microarray data, or structures), which must be deposited in the appropriate genome- or locusspecific database, in a separate section entitled "Accession Numbers," following the Web Resources section (or the Acknowledgments section if no online resources or appendices have been used), directly above the reference list. Please use the following format to list accession numbers: “The accession number(s) for the _______ sequence(s) reported in this paper is/are [database]: [accession number].”

Gender equity (Described according to ICMJE recommendation available from http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html)

Selection and Description of Participants
Clearly describe the selection of observational or experimental participants (healthy individuals or patients, including controls), including eligibility and exclusion criteria and a description of the source population. Because the relevance of such variables as age, sex, or ethnicity is not always known at the time of study design, researchers should aim for inclusion of representative populations into all study types and at a minimum provide descriptive data for these and other relevant demographic variables. Ensure correct use of the terms sex (when reporting biological factors) and gender (identity, psychosocial or cultural factors), and, unless inappropriate, report the sex and/or gender of study participants, the sex of animals or cells, and describe the methods used to determine sex and gender. If the study was done involving an exclusive population, for example in only one sex, authors should justify why, except in obvious cases, (e.g., prostate cancer).” Authors should define how they determined race or ethnicity and justify their relevance.

Submission of Manuscript
The manuscript should be submitted in MS Word file format. The recommended font is Times New Roman with a 11-point font size. All manuscripts must be submitted online through the Genomics & Informatics e-submission system at http://submit.genominfo.org. Any questions concerning manuscript submission should be directed to: Editor, Genomics & Informatics, Korea Genome Organization, The Korean Federation of Science and Technology Societies, Room No. 806, 193 Mallijae-ro, Jung-gu, Seoul 04501, Korea (http://www.kogo.or.kr, Tel: +82-2-558-9394, Fax: +82-2-558-9434, E-mail: kogo@kogo.or.kr).

Peer review and revision of manuscripts
Peer review
A manuscript is generally reviewed by at least two peer reviewers qualified to evaluate the manuscript. It is a single blind peer review. An initial decision will normally be made within one month of receipt of a manuscript. A manuscript that has been published or of which a substantial portion has been published elsewhere will not be accepted. The Editor-in-Chief is responsible for final decisions regarding the acceptance of a peer-reviewed paper.

Manuscript revision
When a manuscript is returned to the corresponding author for revision, the reviewed manuscript must be re-submitted within one month, unless the authors request an extension. A galley proof
and reprint order form will be sent to the corresponding author. The corresponding author is responsible for communicating with the other authors about revisions and final approval of the proofs. The first proofreading is the author's responsibility, and the proof should be returned within three days from the date of receipt.

Copyrights, open access policy and open data policy

**Copyright**
The regulations for acceptance of a manuscript for publication automatically include the consent of the author(s) to transfer the copyright or license to KOGO. Authors should complete a Copyright Agreement Form (CAF) at the time of proofreading. The corresponding author can sign on behalf of any co-authors. The CAF can be obtained from the editorial office. Acceptance of the agreement will ensure full copyright protection and help to disseminate the article to the widest possible readership in print and electronic formats. The authors are responsible for obtaining permission to reproduce copyrighted material from other sources.

**Open access policy**
*Genomics & Informatics* is an open access journal. Articles are distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited for non-commercial purposes. To use the tables or figures of *Genomics & Informatics* in other periodicals, books, or media for scholarly, educational, or even commercial purposes, the process of permission request to the Publisher is not necessary. This is in accordance with the Budapest Open Access Initiative definition of open access. It also follows the open access policy of PubMed Central at the United States National Library of Medicine (http://www.ncbi.nlm.nih.gov/pmc/). All of the content of the journal is available immediately upon publication without an embargo period.

**Archiving policy**
It is accessible without barrier from Korea Citation Index (https://kcigok.kr), National Library of Korea (http://nl.go.kr), or PubMed Central (https://www.ncbi.nlm.nih.gov/pmc/journals/1928/) in the event a journal is no longer published.

**Deposit policy (Self-archiving policy) according to Sherpa/Romeo**
(http://www.sherpa.ac.uk/): Author can not archive pre-print (i.e., pre-refereeing). Author can archive post-print (i.e., final draft post-refereeing).

**Open data policy**
Data sharing is recommended. If the data are already public, the URL site or sources should be disclosed. If data can not be publicized, it can be negotiated with the editor. If there are any inquiries on depositing data, authors should contact the editorial office.

**Clinical data sharing policy**
This journal follows the data sharing policy described in “Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors” (https://doi.org/10.3346/jkms.2017.32.7.1051). As of July 1, 2018, manuscripts submitted to ICMJE journals that report the results of clinical trials must contain a data sharing statement as described below. Clinical trials that begin enrolling participants on or after January 1, 2019 must include a data sharing plan in the trial’s registration. The ICMJE’s policy regarding trial registration is explained at www.icmje.org/recommendations/browse/publishingand-editorial-issues/clinical-trial-registration.html. If the data sharing plan changes after registration, this should be reflected in the statement submitted and published with the manuscript and updated in the registry record. Data sharing statements must indicate the following: whether individual deidentified participant data (including data dictionaries) will be shared; what data in particular will be shared; whether additional, related documents will be available (e.g., study protocol, statistical analysis plan, etc.); and when the data will become available and for how long; by what access criteria data will be shared (including with whom, for what types of analyses, and by what mechanism). Illustrative examples of data sharing statements that would meet these requirements are in Table 1.

**Detailed Description of Use of Articles of Genomics & Informatics Reader benefit**
Publisher applies the Creative Commons Attribution Non-Commercial license to works it publishes and allows free immediate access to, and unrestricted reuse of, original works of all types.

**Reuse benefit**
Publisher applies the Creative Commons Attribution Non-Commercial license to works it publishes and allows free immediate access to, and unrestricted reuse of, original works of all types.

**Copyrights**
Publisher applies the Creative Commons Attribution Non-
Table 1. Examples of data sharing statements that fulfill ICMJE requirements

<table>
<thead>
<tr>
<th>Element</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
<th>Example 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will individual participant data be available (including data dictionaries)?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>What data in particular will be shared?</td>
<td>All of the individual participant data collected during the trial, after deidentification.</td>
<td>Individual participant data that underlie the results reported in this article, after deidentification (text, tables, figures, and appendices).</td>
<td>Individual participant data that underlie the results reported in this article, after deidentification (text, tables, figures, and appendices).</td>
<td>Not available</td>
</tr>
<tr>
<td>What other documents will be available?</td>
<td>Study protocol, statistical analysis plan, clinical study report, analytic code</td>
<td>Study protocol, statistical analysis plan, analytic code</td>
<td>Study protocol</td>
<td>Not available</td>
</tr>
<tr>
<td>When will data be available (start and end dates)?</td>
<td>Immediately following publication. No end date.</td>
<td>Beginning 3 months and ending 5 years following article publication.</td>
<td>Beginning 9 months and ending 36 months following article publication.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>With whom?</td>
<td>Anyone who wishes to access the data.</td>
<td>Researchers who provide a methodologically sound proposal.</td>
<td>Investigators whose proposed use of the data has been approved by an independent review committee (&quot;learned intermediary&quot;) identified for this purpose.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>For what types of analyses?</td>
<td>Any purpose</td>
<td>To achieve aims in the approved proposal.</td>
<td>For individual participant data meta-analysis.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>By what mechanism will data be made available?</td>
<td>Data are available indefinitely at (link to be included).</td>
<td>Proposals should be directed to xxx@yyy. To gain access, data requestors will need to sign a data access agreement.</td>
<td>Proposals may be submitted up to 36 months following article publication. After 36 months, the data will be available in our University's data warehouse but without investigator support other than deposited metadata.</td>
<td>Not applicable</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Information regarding submitting proposals and accessing data may be found at (link to be included).</td>
<td></td>
</tr>
</tbody>
</table>

Data are available for 5 years at a third-party website (link to be included).

ICMJE, International Committee of Medical Journal Editors.

*These examples are meant to illustrate a range of, but not all, data sharing options.

Commercial license to works it publishes. Under this license, although publisher retains ownership of the copyright for content, it allows anyone to download, reuse, reprint, modify, distribute, and/or copy the content.

Author posting benefit:
Publisher applies the Creative Commons Non-Commercial Attribution license to works it publishes. Under this license, although publisher retains ownership of the copyright for content, it allows anyone, including author, to download, reuse, reprint, modify, distribute, and/or copy the content.

Automatic Posting:
Publisher immediately deposits the accepted articles in PubMed Central (http://pubmedcentral.org/) and journal homepage (https://genominfo.org/) upon publication.

Machine readability:
Genomics & Informatics articles can be accessed programatically through PubMed Central or Europe PMC's RESTful Web Service (https://europepmc.org/RestfulWebService). For inquiries, please contact editorial office, as below:

Article processing charge
Neither page charge, article processing fee nor submission fee will be applied since 2019. It is the platinum open access journal

Contact address
Editorial office
Room No. 806, 193 Mallijae-ro, Jung-gu, Seoul 04501, Korea
Tel: +82-2-558-9394
Fax: +82-2-558-9434
E-mail: kogo3@kogo.or.kr
Copyright transfer agreement

The copyright to this article is transferred to Genomics & Informatics effective if and when the article is accepted for publication. The author warrants that his/her contribution is original and that he/she has full power to make this grant. The author signs for and accepts responsibility for releasing this material on behalf of any and all co-authors. The copyright transfer covers the exclusive right to reproduce and distribute the article, including reprints, translations, photographic reproductions, microform, electronic form (offline, online) or any other reproductions of similar nature.

According to the deposit policy (self-archiving policy) of Sherpa/Romeo (http://www.sherpa.ac.uk), authors cannot archive pre-print (i.e. pre-refereeing), but they can archive post-print (i.e. final draft post-refereeing). Authors can archive publisher’s version/PDF.

<table>
<thead>
<tr>
<th>Title of article</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author's signature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>
Publication ethics

For the policies on research and publication ethics that are not stated in these instructions, the Good Publication Practice Guidelines for Medical Journals (http://kamje.or.kr/intro.php?body=publishing_ethics) and the Guidelines on Good Publication(http://publicationethics.org/resources/guidelines) can be applied. The Editor-in-Chief reserves the right to reject manuscripts that do not comply with the below requirements. The author will be held responsible for false statements or failure to fulfill the below requirements.

Statement of Informed Consent

Copies of written informed consent and Institutional Review Board (IRB) approval for clinical research should be kept. If necessary, the editor or reviewers may request copies of these documents to resolve questions about IRB approval or study conduct.

Statement of Human and Animal Rights

All human investigations must be conducted according to the principles expressed in the Declaration of Helsinki. All studies involving animals must state that the guidelines for the use and care of laboratory animals of the authors’ institution, or of any national law, were followed. Registration of clinical trial research: Any research that deals with a clinical trial should be registered with the primary national clinical trial registry site, such as the Korea Clinical Research Information Service (CRiS, http://cris.nih.go.kr), other primary national registry sites accredited by the World Health Organization (http://www.who.int/ictrp/network/primary/en/), or ClinicalTrials.gov (http://clinicaltrials.gov/), a service of the United States National Institutes of Health.

Authorship

Authorship credit should be based on 1) substantial contributions to conception and design, acquisition of data, and/or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; 3) final approval of the version to be published; and 4) agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Every author should meet all of these four conditions.

After the initial submission of a manuscript, any changes whatsoever in authorship (adding author(s), deleting author(s), or re-arranging the order of authors) must be explained by a letter to the editor from the authors concerned. This letter must be signed by all authors of the paper. Copyright assignment must also be completed by every author.

Corresponding author and first author

It does allow multiple corresponding authors for one article. Only one author should correspond with the editorial office. It does accept notice of equal contribution for the first author when the study was clearly performed by co-first authors.

Correction of authorship after publication

It does not correct authorship after publication unless a mistake has been made by the editorial staff. Authorship may be changed before publication but after submission when an authorship correction is requested by all of the authors involved with the manuscript.

Conflict of Interest Statement

The corresponding author must inform the editor of any potential conflicts of interest that could influence the authors’ interpretation of the data. Examples of potential conflicts of interest are financial support from or connections to pharmaceutical companies, political pressure from interest groups, and academically related issues. In particular, all sources of funding applicable to the study should be explicitly stated. As a guideline, any affiliation associated with a payment or financial benefit exceeding $10,000 per annum or 5% ownership of a company or research funding by a company with related interests would constitute a conflict that must be declared. This policy applies to all submitted research manuscripts and review material.

Originality and Duplicate Publication

No part of the accepted manuscript should be duplicated in any other scientific journal without the permission of the Editorial Board. If duplicate publication or plagiarism related to the papers of this journal is detected, the authors will be announced in the journal, their institutes will be informed, and the authors will be penalized. All submitted manuscripts are screened by CrossCheck.
(Similarity Check), a plagiarism detection program provided by iThenticate. The authors assure that no substantial part of the work has been published or is being considered for publication elsewhere. When any of the results is to appear in another journal, details must be submitted to the Editor-in-Chief, together with a copy of the other paper(s) and the expected date(s) of publication.

Secondary Publication

It is possible to republish manuscripts if the manuscripts satisfy the condition of secondary publication of the Uniform Requirements for Manuscripts Submitted to Biomedical Journals by the International Committee of Medical Journal Editors (ICMJE), available from http://www.icmje.org/. These are:

- The authors have received approval from the editors of both journals (the editor concerned with the secondary publication must have access to the primary version).
- The priority for the primary publication is respected by a publication interval negotiated by editors of both journals and the authors.
- The paper for secondary publication is intended for a different group of readers; an abbreviated version could be sufficient.
- The secondary version faithfully reflects the data and interpretations of the primary version.
- The secondary version informs readers, peers, and documenting agencies that the paper has been published in whole or in part elsewhere—for example, with a note that might read, “This article is based on a study first reported in the [journal title, with full reference]”—and the secondary version cites the primary reference.
- The title of the secondary publication should indicate that it is a secondary publication (complete or abridged republication or translation) of a primary publication. Of note, the United States National Library of Medicine (NLM) does not consider translations to be “republications” and does not cite or index them when the original article was published in a journal that is indexed in MEDLINE.

Process to manage research and publication misconduct: When the Journal faces suspected cases of research and publication misconduct, such as a redundant (duplicate) publication, plagiarism, fabricated data, changes in authorship, undisclosed conflicts of interest, an ethical problem discovered with the submitted manuscript, a reviewer who has appropriated an author’s idea or data, complaints against editors, and other issues, the resolving process will follow a flowchart provided by the Committee on Publication Ethics (http://publicationethics.org/resources/flowcharts). The discussion and decision on suspected cases are done by the Editorial Board of Genomics & Informatics.
Author's checklist


☐ 2. Title page: (1) complete title, (2) manuscript type, (3) authors’ name, (4) affiliations, (5) telephone, facsimile and E-mail address of corresponding author, (6) running title (no more than 50 characters).

☐ 3. Abstract in unstructured format within 250 words.

☐ 4. Six or fewer keywords, preferably MeSH terms.

☐ 5. Manuscript is structured as follows:
   - Original Article: Abstract, Keywords, Introduction, Materials and Methods, Results, Discussion, References, Table and Figure.
   - Research Communication: Abstract, Keywords, Main Text, and Conclusion (if applicable), References, Table and Figure.
   - Application Note: Abstract, Keywords, Availability, Introduction, Main Text, and Supplementary Information, References, Table and Figure.

☐ 6. Reference in proper format. Check that all references listed in the references section are cited in the text and vice versa.

☐ 7. All figures and tables referenced in the text and numbered in order of its appearance in the text.

☐ 8. Figures as a separate files, in TIFF or JPG format, minimum 300 dpi.

☐ 9. Each necessary permission statement signed by the appropriate source.

☐ 10. Elucidation of research or project support/funding.