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 M2Community 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 2021, 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.
<table>
<thead>
<tr>
<th><strong>EDITOR IN CHIEF</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Park, Taesung</td>
<td>Seoul National University, Korea</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>ADVISORY EDITORIAL BOARD</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Batzer, Mark A.</td>
<td>Louisiana State University, U.S.A.</td>
</tr>
<tr>
<td>Church, George M.</td>
<td>Harvard University, U.S.A.</td>
</tr>
<tr>
<td>Lee, Byungkook</td>
<td>National Institute of Health, U.S.A.</td>
</tr>
<tr>
<td>Matsuda, Fumihiko</td>
<td>Kyoto University, Japan</td>
</tr>
<tr>
<td>Sakaki, Yoshiyuki</td>
<td>RIKEN Genomic Science Center, Japan</td>
</tr>
<tr>
<td>Seo, Jeong-Sun</td>
<td>Seoul National University, Korea</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>ASSOCIATE EDITORS</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cho, Soo Young</td>
<td>National Cancer Center, Korea</td>
</tr>
<tr>
<td>Choi, Murim</td>
<td>Seoul National University, Korea</td>
</tr>
<tr>
<td>Han, Kyudong</td>
<td>Dankook University, Korea</td>
</tr>
<tr>
<td>Huh, Sun</td>
<td>Hallym University, Korea</td>
</tr>
<tr>
<td>Kim, Sangsoo</td>
<td>Soongsil University, Korea</td>
</tr>
<tr>
<td>Oh, Bermseok</td>
<td>Kyung Hee University, Korea</td>
</tr>
<tr>
<td>Oh, S. June</td>
<td>Inje University, Korea</td>
</tr>
<tr>
<td>Park, Hyun Seok</td>
<td>Ewha Womans University, Korea</td>
</tr>
<tr>
<td>Yoon, Kyong-Ah</td>
<td>Konkuk University, Korea</td>
</tr>
<tr>
<td>Won, Sungho</td>
<td>Seoul National University, Korea</td>
</tr>
<tr>
<td>Woo, Hyun Goo</td>
<td>Ajou University, Korea</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>EDITORIAL BOARD</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahn, Chul Woo</td>
<td>University of Texas, U.S.A.</td>
</tr>
<tr>
<td>Chen, Jyh-Yih</td>
<td>Academia Sinica, Taiwan</td>
</tr>
<tr>
<td>Cordaux, Richard</td>
<td>University of Poitiers, France</td>
</tr>
<tr>
<td>Divakar, Darshan Devang</td>
<td>King Saud University, Saudi Arabia</td>
</tr>
<tr>
<td>Hiroki, Yokota</td>
<td>Indiana University, U.S.A.</td>
</tr>
<tr>
<td>Kim, Junhyong</td>
<td>University of Pennsylvania, U.S.A.</td>
</tr>
<tr>
<td>Kohane, Isaac S.</td>
<td>Harvard University, U.S.A.</td>
</tr>
<tr>
<td>Liang, Ping</td>
<td>Brock University, Canada</td>
</tr>
<tr>
<td>Marquardt, Jens</td>
<td>Mainz University, Germany</td>
</tr>
<tr>
<td>Mishra, Siddhartha K.</td>
<td>Harisingh Gour Central University, India</td>
</tr>
<tr>
<td>Ohno-Machado, Lucila</td>
<td>Harvard University, U.S.A.</td>
</tr>
<tr>
<td>Parine, Narasimha Reddy</td>
<td>King Saud University, Saudi Arabia</td>
</tr>
<tr>
<td>Pawan, K. Dhar</td>
<td>RIKEN Genomic Science Center, Japan</td>
</tr>
<tr>
<td>Salem, Abdel Halim</td>
<td>Arabian Gulf University, Bahrain</td>
</tr>
<tr>
<td>Shahik, Shah Md.</td>
<td>University of Chittagong, Bangladesh</td>
</tr>
<tr>
<td>Sree, N. Sreenath</td>
<td>Case Western Reserve University, U.S.A.</td>
</tr>
<tr>
<td>Srikulnath, Kornsorn</td>
<td>Kasetsart University, Thailand</td>
</tr>
<tr>
<td>Terwilliger, Joseph</td>
<td>Columbia University, U.S.A.</td>
</tr>
<tr>
<td>Valdes, Jorge</td>
<td>Centro de Genómica y Bioinformática, Chile</td>
</tr>
<tr>
<td>Van, Steen</td>
<td>Kristel University of Liége, Belgium</td>
</tr>
<tr>
<td>Zhang, Feng</td>
<td>Fudan University, China</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>ETHICS EDITOR</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chung, Yeun-Jun</td>
<td>The Catholic University of Korea, Korea</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>STATISTICS EDITOR</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Han, Buhm</td>
<td>Seoul National University, Korea</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>MANUSCRIPT EDITOR</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chang, Soo Hee</td>
<td>Infolumi, Korea</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>LAYOUT EDITOR</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeong, Eun Mi</td>
<td>M2community, Korea</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>WEBSITE AND JATS XML FILE PRODUCER</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bae, Hyo-Jeong</td>
<td>M2community, Korea</td>
</tr>
</tbody>
</table>
Editorial
Editor’s introduction to this issue (G&I 19:2, 2021)
Taesung Park

Mini review
OMICS approaches in cardiovascular diseases: a mini review
Md. Mehadi Hasan Sohag, Saleh Muhammed Raqib, Syaefudin Ali Akhmad

Original articles
A protein interactions map of multiple organ systems associated with COVID-19 disease
Dhammapal Bharne

Comparative study: nonsynonymous and synonymous substitution of SARS-CoV-2, SARS-CoV, and MERS-CoV genome
Vipan Kumar Sohpal

A network-biology approach for identification of key genes and pathways involved in malignant peritoneal mesothelioma
A. M. U. B. Mahfuz, A. M. Zubair-Bin-Mahfuj, Dibya Joti Podder

Analysis of H3K4me3-ChIP-Seq and RNA-Seq data to understand the putative role of miRNAs and their target genes in breast cancer cell lines
Aditya Jain, Smiline Girija A. S., Paramasivam A., Vijayashree Priyadharsini J.

Implications of the simple chemical structure of the odorant molecules interacting with the olfactory receptor 1A1
S. June Oh

Analysis of genome variants in dwarf soybean lines obtained in F6 derived from cross of normal parents (cultivated and wild soybean)
Neha Samir Roy, Yong-Wook Ban, Hana Yoo, Rahul Vasudeo Ramekar, Eun Ju Cheong, Nam-II Park, Jong Kuk Na, Kyong-Cheul Park, Ik-Young Choi
Editor's introduction to this issue (G&I 19:2, 2021)

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

In this issue, there are six original articles and one mini review. The first article by by Sohag et al. (Jagannath University, Bangladesh) provides a short review on omics approaches to cardiovascular diseases (CVDs). The author summarizes the genomics, proteomics, transcriptomics, and metabolomics in CVDs with a well-organized prospect.

The first original article is about a protein interactions map of multiple organ systems as-sociated with coronavirus disease 2019 (COVID-19) disease by Dr. Bharne (University of Hyderabad, India). This study appears to be motivated by reports that reduced antibody levels and disease recurrence in recovered COVID-19 patients require understanding of the epidemic at a key level. Multiple organ failure cases in patients with COVID-19 have highlighted consideration for other organ systems. This study used RNA sequencing data to determine disease-associated differentially regulated genes and related protein interactions in multiple organ systems, which implies the importance of early diagnosis and treatment of the disease. RNA sequencing data were obtained from autopsy specimens of lung, heart, jejunum, liver, kidney, intestine, bone marrow, adipose, placenta, and skin from 24 patients who died of COVID-19 infection. The total number of samples in the sequencing data was 88, including five negative control samples. Using significantly expressed genes in different organ systems, protein interactions of multiple organ systems were then mapped, revealing CAV1 and CTNNB1 as top nodes. A core interactions sub-network was analyzed to identify several functionally important modules such as AR, CTNNB1, CAV1 and PIK3R1 proteins. In addition, this study highlighted some of the druggable targets to analyze in drug re-purposing strategies against the COVID-19 pandemic. I think the protein interaction maps and modular interactions of differentially regulated genes in multi-organ systems would provide the clues to researchers to rapidly investigate novel therapeutics for the COVID-19 pandemic.

The second article by Sohpal (Beant College of Engineering & Technology, India) performed a comparative study of coronaviruses including severe acute respiratory syndrome coronavirus 2, severe acute respiratory syndrome coronavirus, and Middle East respiratory syndrome coronavirus focusing on non-synonymous and synonymous substitutions. Through simulation studies, nucleotide sequence of closely related strains of respiratory syndrome viruses, codon-by-codon with maximum likelihood analysis, z selection and the divergence time were investigated.

The third article by Mahfuz et al. (University of Development Alternative, Bangladesh) presented a network-biology approach for identification of key genes and pathways involved in malignant peritoneal mesothelioma (MPM). To understand the molecular mechanisms responsible for the initiation and progression of MPM, this study aims to identify the key genes and pathways responsible for MPM. Several bioinformatics analyses were performed such as identification of differentially expressed genes, pathway analysis, and protein-protein interaction network analysis, providing an insight into the potential genes and pathways involved in MPM.
The next article by Kotipalli et al. (Centre for Development of Advanced Computing, India) presents the results of epigenetic analysis for breast cancer. Using chromatin immunoprecipitation sequencing, epigenetic regulation of gene expression by in-silico analysis of histone modifications was carried out. Histone modification data of H3K4me3 from one normal-like and four breast cancer cell-lines were used to predict miRNA expression at the promoter level. Predicted miRNA promoters were used as a probe to identify gene targets. This study is expected to provide the insight into predicted role of H3K4me3 mediated gene regulation via the miRNA-mRNA relationship.

The next article by Oh (Inje University College of Medicine, Korea) is about G protein‐coupled receptors, including olfactory receptors. Through simulation study, the interaction between human olfactory receptor 1A1 and the corresponding odorant molecule was investigated. Further, it was also studied how the chemically simple odorant molecule activates the olfactory receptor.

Finally, the last original article by Choi’s group (Kangwon National University, South Korea) presents analysis of genome variants in dwarf soybean lines obtained in F6 derived from cross of normal parents (cultivated and wild soybean). Through whole genome sequencing data analysis, the authors reported DNA variations between the normal and dwarf members of four lines harvested from a single seed parent in an F6 recombinant inbred line population. The list of single nucleotide polymorphisms provides important information for the genetics of soybean plant height and crop breeding and expected to be useful genetic resources for plant breeders.

ORCID

Taesung Park: https://orcid.org/0000-0002-8294-590X
OMICS approaches in cardiovascular diseases: a mini review

Md. Mehadi Hasan Sohag1,2*, Saleh Muhammed Raqib3, Syaefudin Ali Akhmad4

1Department of Genetic Engineering and Biotechnology, Jagannath University, Dhaka 1100, Bangladesh
2Biotechnology Research Initiative for Sustainable Development, Dhaka 1219, Bangladesh
3Baridhara Scholar’s Institution, Dhaka 1212, Bangladesh
4Department of Biochemistry, Faculty of Medicine, Islamic University of Indonesia, Yogyakarta 55584, Indonesia

Ranked in the topmost position among the deadliest diseases in the world, cardiovascular diseases (CVDs) are a global burden with alterations in heart and blood vessels. Early diagnostics and prognostics could be the best possible solution in CVD management. OMICS (genomics, proteomics, transcriptomics, and metabolomics) approaches could be able to tackle the challenges against CVDs. Genome-wide association studies along with next-generation sequencing with various computational biology tools could lead a new sight in early detection and possible therapeutics of CVDs. Human cardiac proteins are also characterized by mass spectrophotometry which could open the scope of proteomics approaches in CVD. Besides this, regulation of gene expression by transcriptomics approaches exhibits a new insight while metabolomics is the endpoint on the downstream of multi-omics approaches to confront CVDs from the early onset. Although a lot of challenges needed to overcome in CVD management, OMICS approaches are certainly a new prospect.

Keywords: cardiovascular diseases, genomics, metabolomics, OMICS, proteomics, transcriptomics

Introduction

In this 21st century, cardiovascular diseases (CVDs) ranked as one of the serious health issues of the developed and developing countries that encompass so many factors such as tissue, organs and multidimensional molecular perturbations make CVD a global burden [1]. World Health Organization declares CVD as the topmost cause of death throughout the world. In 2015, 17.9 million people died from CVDs which represents 31% of all diseases where cardiac arrest is one of the biggest portions of CVDs. Global Heart Atlas affirmed that 200 million prevalent peripheral artery disease cases were estimated in the last decade where three fourth of the population are living in low or middle-income countries, specifically most of the patients are from the southern region of Asia [2]. In the era of the coronavirus disease 2019 (COVID-19) pandemic, CVD becomes comorbid that increased the risk of getting higher morbidity and mortality of COVID-19 [3]. British Heart Foundation defines CVD as a total of all the diseases of the heart and circulation including coronary heart diseases, angina, heart attack, congenital heart diseases, and stroke [4]. As various researchers describe that CVD is enormously influenced by differences in environmental parameters and lifestyle attributes while socioeconomic condi-
tions causing metabolic diseases and age risk factors are also playing critical roles other than because of infection [5,6]. Besides this, the genetic makeup of the individual is one of the pivotal factors of CVD prognostics and diagnostics where OMICS technology could be recommended for further investigations [7]. OMICS is nothing but promising techniques where the combination of genomics, transcriptomics, proteomics, and metabolomics are conglomerated into an umbrella where researchers can predict the probable identification of the disease and make the best solution for the remedy [8].

**Clinical Spectrum of CVD and Its Laboratory Testing Biomarker**

CVD has various symptoms and manifestations as well as its laboratory testing biomarkers. Common signs and symptoms of CVD are associated with a heart problem and vascular problem attributed to which organ affected such as paresis when brain getting stroke disease. Other symptoms are chest pain, pain in the neck, jaw, throat back, and upper abdominal, fatigue, edema, dyspnea/shortness of breath, irregular heartbeat, nausea, loss of appetite, dizziness, numbness, and cyanosis, as well as easily tiring during exercise [9].

CVD may have various clinical spectrum and various laboratory results as a biomarker profile in the diagnosis process, therapeutic monitoring, a predictor for prognosis as well as the etiology of the disease (Fig. 1). For example in the COVID-19 pandemic with a highly clinical spectrum of signs and symptoms of infection by severe acute respiratory syndrome coronavirus 2 (novel coronavirus) presented on a patient includes CVD manifestation affected both heart and vascular on all system. In terms of physiology, regarding the COVID-19 on clinical presentation with the various sign and symptom, there is a specific area of physiology having a pivotal role in exploring the cause-effect relationship and completing the biomarker profiles of the patient that is electrophysiology i.e., electrocardiogram and channel recording to uncover the pathophysiology of CVD. CVD phenotype with complete clinical presentation supported by electrophysiology will be beneficial to

![Fig. 1. OMICS approach from diagnosis process until prognosis as well as the etiology of the disease. CK-MB, creatine kinase-myocardial band; ECG, electrocardiogram; HRV, heart rate variability.](https://doi.org/10.5808/gi.21002)
link CVD biomarker in terms of OMICS approaches supporting the signs as well as the symptoms of CVD patient. Table 1 depicts the biomarker of CVDs particularly in COVID-19 in comparison with other causes [10].

From the basic science perspective, that worthwhile to use OMICS approaches for mitigating that the argument of genetic differences may be at the core of differential disease expression in the sign and symptoms which may be integral to altered responses to standard medical therapies [11].

**OMICS Approaches for CVD Mitigation**

OMICS is an amalgamation of technologies which accumulates initially the universal detection of genes referred to as genomics, processing with proteins and peptides that termed as proteomics, analyzing with RNAs (e.g., mRNA) that defined as transcriptomics, and the study of metabolic molecules as metabolomics in a perspective of biological phenomenon. Additionally, next-generation sequencing (NGS), genome-wide association study (GWAS), RNA sequencing (RNA-Seq), mass spectrophotometry (MS), and other high throughput techniques are the blessing outcomes from the broad ranges of application of OMICS [12]. Data that are found from the OMICS are relevant and referred to in silico manners that are strictly standardized as computational approaches whereas it is said that OMICS records are explored into bioinformatics. Data sets are processed uniquely along with the procedures e.g., acquisition of data, processing and analysis, and storage. Nevertheless, the most influential part of the data processing system is nothing but the management of data which could be treated as big data management as a lot of sequences of nucleic acids from various organisms (genomic, plasmid, and so on) are needed to be retrieved frequently for multi-research purposes. Besides this, stored data are also needed to be managed for the epigenetic study which will play a key role to answer the mystery of life [13]. However, ‘OMICS’ approaches could lead the scientific community in identifying the metabolic pathway and responsible genes associated with CVDs which is the ultimate window opening scenario to de-sign and validate the drugs that will undergo a clinical trial. A figurative outline based on a previous study has been drawn (Fig. 2) to express the general idea of OMICS workflow [14].

**Genomics in CVDs**

Genomics approaches are esteemed as one of the most accurate strategies for the diagnostics and prognostics of CVDs. Single nucleotide polymorphisms (SNPs) are an example of the genomics technology that would be an imperative tool for CVD researches by identifying the candidate gene through the GWAS study. Already, there is stringent evidence that has confirmed in favor of SNPs that could identify biomarkers for various common diseases [15]. Besides this, genomics-related risk profiling and quantitative clinical measurements give a pathway for precision therapies for various cardiovascular abnormalities. In recent days, Genomic variation and phenotypic illness or wellness might be an indicator of clinical implication by Electronic Health Records and Genomics programs [16]. These programs are smoothly run by using clinical data from electronic medical records that is linked to various clinical repositories or biobanks including studies from eMERGE (Electronic Medical Records and Genomics; http://www.gwas.net) network and as well as Clinical Implementation of Personalized Medicine through Electronic Health Records and Genomics (CLIPMERGE PGx) [17]. Additionally, the National Institute of Health has taken an initiative through National Centre for Biomedical Computing (NCBC) entitled “Informatics for Integrating Biology and the Bedside (i2b2; https://www.i2b2.org)” where researchers have the user-friendly software analysis tools for collection and management of clinical data based on genomics research [18]. Besides this, hypertrophic cardiomyopathy was the inception case of the cardiovascular genetic disorder, where a missense mutation occurred in β myosin heavy chain was detected [19]. Mostly, the genome sequencing approaches came with great hope towards the scientific community for identifying DNA sequence variants (DSVs) which is the ultimate consequence of single gene abnormalities. NGS took the definitive place for revealing the gene defect

<table>
<thead>
<tr>
<th>Cardiac biomarker</th>
<th>Definition</th>
<th>Association with COVID-19</th>
<th>Prognostic potential</th>
</tr>
</thead>
<tbody>
<tr>
<td>cTn</td>
<td>Gold-standard necrotic biomarkers for myocardial injury</td>
<td>Acute myocardial injury, ICU admission, in hospital death and severity of inflammation in COVID-19</td>
<td>+++</td>
</tr>
<tr>
<td>BNP</td>
<td>Predictor of adverse outcome following acute myocardial injury</td>
<td>Acute myocardial injury, ICU admission, in hospital death</td>
<td>++</td>
</tr>
<tr>
<td>CK-MB</td>
<td>Biomarker of myocardial damage and reperfusion</td>
<td>Acute myocardial injury, ICU admission, in hospital death</td>
<td>+</td>
</tr>
</tbody>
</table>

CVD, cardiovascular disease; COVID-19, coronavirus disease 2019; cTn, cardiac troponin; ICU, intensive care unit; BNP, brain natriuretic peptide; CK-MB, creatine kinase-myocardial band.
that happened due to cardiovascular problems where all the DSVs are profoundly identified through whole-exome sequencing or whole-genome sequencing methods. Millions of DNA fragments are sequenced here in NGS, where phenotypic expression is also determined that allows investigators to draw out a clear picture of the disorder [20]. Already a report has been shown that intermediate polymorphism of several genes like angiotensin I converting enzyme, Chymase, Coagulation factor III and V have keeping up the adverse preoperative cardiovascular outcomes. Due to the changes in enzymatic activity and down-regulation of the receptors, polymorphism of the above genes can turn into an adverse cardiac event after bypass grafting [21]. Additionally, up-regulation of the Interleukin-6 level also put an unfavorable cardiac status after surgery. Another example of preoperative genomics is associated with platelet glycoprotein IIIa gene (ITGB3) polymorphism. This initiates the aggregation of platelet and increased the level of troponin1 which occur unpleasant results of a patient with surgery like coronary graft occlusion, myocardial infarction, and even death [22]. However, additional functional genomics studies of atrial-natriuretic peptide and brain natriuretic peptide denotes this molecule as biomarkers of heart failure, where the prior knowledge from whole genome picture minimizes the risks of the sudden demise of the patient [23].

**Proteomics in CVDs**

Prior knowledge of protein function and characterization would be the promising solution to identify and mitigate the risk factors towards CVDs. As we know, proteomics approaches could demonstrate the protein patterns and as well as post-translational modifications that help to understand the genetic consequences of the fi-
nal product of responsible gene which is not deeply predictable through the bioinformatics analysis and not also be confined by the genomics studies [24]. By achieving the genomic information, proteomics studies based on accumulating macromolecule databases with the advent of MS or Edman degradation would be done for further progress. The targeted proteome is extremely compared with the entire set of the known proteome to unveil the function and characteristics of disease-specific protein that is named functional proteomics [25]. However, plenty of sophisticated molecular biology research tools like resolving from a gel, electrospray ionization, matrix-assisted laser desorption/ionization are needed to prepare and ionize the sample while peptide sequencing, as well as peptide mass fingerprinting, make the mass spectrometric analysis, is so inevitable techniques in proteomics technology [26]. Though it is not completely understood about cardiac dysfunction in systemic as well as specific heart muscle diseases, considerable changes in myocardial gene and expression of proteins could be the determiner of CVD status. Additionally, two-dimensional electrophoresis was done as a proteomics tool to make a profound dataset of cardiac proteins of humans. Information of hundreds of proteins related to CVDs is deposited here for further analysis and comparison of newly found proteins [27]. Moreover, databases of human cardiac proteins along with the databases for other animals are used for constructing an animal model of heart diseases. Additionally, various reports confirmed that the identification and expression of the protein (e.g., genetic elements, toxic substances, antibodies, and so on) has already been investigated on dilated cardiomyopathy that gives new hope in CVD research [28]. Surprisingly, proteomics also put a positive vibe on cardiac transplantation where identification of cardiac-specific antigen identification might be the global approach. Besides these approaches, novel cardiac biomarkers are also identified through proteomics investigations. Previous studies have been clarified that levels of heat shock protein-27 in tissue samples are determiners of acute myocardial infarction (AMI). As a consequence, proteolytic factors in the blood are also treated as the new insight of biomarker identification strategies to early-stage detection of the acute coronary syndrome [29]. Another biomarker, gamma glutamyl transferase is also associated with the prediction of survival rate in patients with acute ischemic stroke [30]. It has been already well-known that about 177 candidate biomarker protein was reported where proteomics approaches could quantify the measurement of the target proteins [31]. Identification of the protein biomarker from blood serum or plasma is still challenging though a complementary proteomics strategy with multiplex immune analysis could validate the target biomarker. Surprisingly, collected fractionations of plasma/serum of relevant protein or protein complex named interactomes are characterized by immunocapture and liquid chromatography-mass spectrometry. Additionally, glycosylated protein biomarker like erythropoietin is another potential regulator in cardiac patient management [32]. Furthermore, C-reactive protein (CRP) is also regarded as one of the promising biomarkers to identify systemic inflammation in atherogenesis which shows a very close association with chemokines and cytokines for disease development. A molecular entity of the CRP proteins is also evaluated and quantified by the MS spectra [33].

Transcriptomics in CVDs

As with other approaches, transcriptomics also plays an inevitable input in current days CVD research. As we know, that the expression of a gene is entirely written on the RNA as a transcript, so the techniques in transcriptomics made an easy pathway to detect the defective genes regarding CVDs in a very early stage. These techniques are efficient to identify as well as quantify all RNA transcripts that may be regulatory or not, as it is affirmed that occurrences of CVDs are dependent on activation or silencing of responsible cardiac genes. Usually, microarray and RNA-Seq are primary schemes that are used in this technology, where the second one is more comprehensive, specific, and costly than the previous one. Microarray denotes the quantification of RNA is somewhat similar to a genome-wide DNA microarray, while RNA-Seq is a high throughput technology is likely as NGS approaches of complementary DNA. To sum up the transcriptomic analysis, few protocols are used for the annotation of the gene expression, e.g., Heat maps, gene co-expression analysis network, pathway analysis, and so on. Above processing in the experiment would be capable of making a crystal clear visualization with a firm understanding of the entire transcript [34]. However, transcriptomics is the only option to reveal the impact of a variety of RNA e.g., micro RNA (miRNA), small interfering RNA (siRNA), non-coding RNA (ncRNA), and others, as they are thought as eminent players of the cell to cell communication and thought to be as a potential biomarker of CVD prognostics and diagnostics. The availability and stability of ncRNA are profoundly higher than the coding counterpart that could make themselves as intense therapeutic targets as they are not degraded in blood. Previous studies also affirm that the role of ncRNA in CVDs is tremendous because they are associated with different cardiovascular processes including heart development; fibrosis, heart contractility, apoptosis, gene regulation, and so on. This phenomenon will open the horizon of translational research gateway that could introduce personalized medicine in CVD management. Furthermore, quantitative real-time polymerase chain reaction (qPCR) with high throughput arrange-
ment of computational biology gives a rigid strength to the microarray where RNA-Seq approaches give the entire information of complete RNA profiling. These inclusive approaches of transcriptomics disclose the way of identification, progression, and treatment of CVDs [35]. However, a recent study gives a picture of bioinformatics approaches through NGS platforms used in transcriptomics. By sequence mapping and quantifying the expression, RNA-Seq datasets represent the differential detection processes of expressed cardiac genes [36]. Circulating miRNA is reported as the signature molecule to identify the risk level of heart failure patients as well as AMI. Specifically, miRNA-21, miRNA-208, miRNA-423, and miRNA-49 are responsible for heart failure and myocardial infarction, while miRNA-1, miRNA-133, miRNA-328 have also been identified as a biomarker of arrhythmia. Besides this, miRNA-22 could be treated as therapeutics against cardiac autophagy. Overexpression of miRNA-99a in animal models proved that it could improve cardiac function [37]. Samples from the patients with adverse cardiac events could easily be distinguished by the estimation of miRNAs through real-time qPCR due to the cardiac selectivity and plasma stability with rapid release of these biomarkers [38].

**Metabolomics in CVDs**

Disturbances in metabolic signals alter the usual characteristics of the organ system and the cardiovascular system is not an exception. Contemporarily, the notion of CVDs could alter the fates of the intermediate metabolites in different metabolic pathways, which signs a new thought on the pathophysiological status of a CVD patient. In terms of CVD, there is evidence that because lack of endogenous ascorbate leads to the underlying cause of CVD compared with the animal having higher endogenous ascorbate [39]. Alongside other technologies, metabolomics will also offer insight into the diagnosis and prognosis of CVDs through the prediction and analysis of cellular metabolism [14]. Reports illustrate that metabolomics stands on the endpoint of the downstream circle of OMICS approaches where biological fluids or tissue samples are analyzed with nuclear magnetic resonance (NMR) and liquid chromatography-based mass spectrometry (LC/MS) as well. In terms of specificity and throughput attribute, NMR is superrelative to targeted and non-targeted LC/MS while the opposite scenario is observed in the case of sensitivity and metabolome coverage [40]. Furthermore, analysis in ketogenic and glucogenic regulation alongside macromolecular metabolic pathways, metabolome like circulating branched-chain amino acids could interact with insulin resistance results from the prognosis and better treatment of CVDs with comorbidities like obesity and diabetes. Besides this, lipid metabolism alteration is also regarded as one of the strongest incidents for CVD occurrences. Metabolomics study could unveil the early detection of specific cholesterol esters, choline, triacylglycerols, and a density level of lipoproteins considerably associated with chronic heart disease. Additionally, metabolomics strategies also describing inborn metabolic errors which could be associated with thousands of metabolites in cardiometabolic pathways [14,40]. Metabolic screening of choline, trimethylamine N-oxide, and betaine leads towards a conclusion of the association of dietary metabolites with the development of CVDs. Hence, persistent arterial fibrillation is also regarded as the hypermetabolic state where glucose metabolism is impaired and prominently associated with ketone bodies [41]. Recent updates in metabolomics study depict the early vascular aging syndrome motives which are firmly associated with cardiovascular morbidity and mortality. Arterial stiffness is measured by four lysophosphatidylcholines that predict the risks of hypertension as well [42]. Till now, more than eight thousand metabolomes are enlisted into databases of human metabolome projects that actively take part in the understanding of system biology. Already biomarkers of myocardial ischemia are known where approximately six metabolites are confirmed. Another recent study has stated that eighteen dietary metabolites could signal the early onset [40,41]. Reprogrammed cardiac metabolic pathway increased the uptake of glucose due to cardiac hypotrophy where fatty acid oxidation decreased to a minimum level. This metabolic alteration leads the patients towards heart failure. Besides this, a considerable amount of changes are found in fatty acid oxidation in the time of transition to heart failure. Here, LC-MS-based metabolomics showed a decreased level of TCA cycle intermediates [43].

Novel metabolic pathways related to an inflammatory protein like fibrinogen and CRP are not confined yet irrespective of the age, gender, diet, or drug effects of the patients. However, MS-dependent metabolomics approaches initiate the eye-catching processes of detection based on physiological attributes in targeted or untargeted metabolic pathways for assessing the risks of CVDs with possible remedy recommendations [44].

**Conclusion**

Numerous numbers of researchers are engaging in CVD researches to find out the way of an ultimate solution against this global burden. It is certain that, for the improvement of CVD management, early diagnosis is a must. Multi-omics approaches have opened a broad avenue to detect the early onset of CVDs symptom though a long way to go for the inclusive conclusion. However, the Integration of OMICS strategies with clinical researches
will be a strong shield to fight against CVDs.

**ORCID**

Md. Mehadi Hasan Sohag: https://orcid.org/0000-0001-6058-1626
Saleh Muhammed Raqib: https://orcid.org/0000-0001-5074-7143
Syafudin Ali Akhmad: https://orcid.org/0000-0001-9554-8578

**Authors' Contribution**

Conceptualization: MMHS, SMR, SAA. Data curation: MMHS. Formal analysis: MMHS, SAA. Funding acquisition: MMHS, SMR, SAA. Methodology: MMHS, SAA. Writing - original draft: MMHS. Writing - review & editing: MMHS, SAA, SMR.

**Conflicts of Interest**

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

**References**


https://doi.org/10.5808/gi.21002


Coronavirus disease 2019 (COVID-19) is an on-going pandemic disease infecting millions of people across the globe. Recent reports of reduction in antibody levels and the re-emergence of the disease in recovered patients necessitated the understanding of the pandemic at the core level. The cases of multiple organ failures emphasized the consideration of different organ systems while managing the disease. The present study employed RNA sequencing data to determine the disease associated differentially regulated genes and their related protein interactions in several organ systems. It signified the importance of early diagnosis and treatment of the disease. A map of protein interactions of multiple organ systems was built and uncovered CAV1 and CTNNB1 as the top degree nodes. A core interactions sub-network was analyzed to identify different modules of functional significance. AR, CTNNB1, CAV1, and PIK3R1 proteins were unfolded as bridging nodes interconnecting different modules for the information flow across several pathways. The present study also highlighted some of the druggable targets to analyze in drug re-purposing strategies against the COVID-19 pandemic. Therefore, the protein interactions map and the modular interactions of the differentially regulated genes in the multiple organ systems would incline the scientists and researchers to investigate in novel therapeutics for the COVID-19 pandemic expeditiously.

Keywords: cluster analysis, COVID-19, gene ontology, protein interaction maps, RNA-Seq

Introduction

Coronavirus disease 2019 (COVID-19) is a pandemic disease caused by the novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (World Health Organization). As on 11 August 2020, it reported to infect more than 78 million and decease over 1.7 million people around the world (https://www.worldometers.info/coronavirus/). The disease cases are further escalating causing human sufferings. Currently, several vaccines are being evaluated at various clinical stages [1,2] and some available drugs are being investigated to re-purpose [3,4] in the treatment of manageable cases of the COVID-19 disease. The SARS-CoV-2 is a highly transmissible virus containing unusually larger RNA as genome and spike-like glycoprotein envelope [5]. It is different from other corona viruses in having strong binding affinity with human cell surface receptors [6]. The virus begins the process of infection by binding to human cell receptors such as angiotensin-converting enzyme 2 (ACE2), transmembrane serine protease 2, cyclophilins, CD147, and CD26 [7]. The ACE2 are the functional receptors of SARS-CoV-2 and are distributed in the cells of lung, heart, kidney and intestinal tissues [8]. Therefore, the virus can transmit to several organ systems and evade the host immune response leading
to multi-organ failure and death. Hence, infection in the multiple organs should be quickly assessed [9] to manage the individual patients early and reduce the risk of decompensation.

Recent report of re-emergence of the SARS-CoV-2 in a recovered patient [10] necessitated better understanding of the infection, contagion and pathology of the disease. IgG levels and neutralizing antibodies in the recovered patients were decreasing after few months [11]; however, the receptor-binding domain specific antibodies possessed the strong antiviral activity [12]. In addition to these recent findings, perspectives of gene regulations and protein interactions at multi-organ level could play significant role in gaining insights and therapeutic interventions of the disease. The rapid development of sequencing technologies in the past few decades made significant impact on research in molecular biology of viral diagnosis [13] and drug discovery [14]. RNA-sequencing technology provided unprecedented information about the novel and known gene structures and annotations from coding and non-coding transcripts. Analyzing RNA-sequencing data of multiple organ systems associated with COVID-19 could unveil several aspects of the pandemic disease. Therefore, the present study employed RNA-sequencing data of several organ systems from the SARS-CoV-2 infected and deceased individuals to analyze differentially expressed genes and interpret protein interactions that led to identification of several proteins for the therapeutic interventions.

**Methods**

**Identification of differentially expressed genes**

RNA sequencing data was obtained from autopsy specimens of lung, heart, jejunum, liver, kidney, bowel, marrow, fat, placenta and skin of 24 patients deceased due to COVID-19 infection. The total number of samples in the sequencing data was 88 including five negative control samples. The sequencing data was mapped to HG38 Human reference genome and processed using HTSeq-Count [15] to produce raw read counts of mRNA transcripts. Such transcripts read counts for each organ sample were retrieved from the NCBI Gene Expression Omnibus public repository using the accession number GSE150316. Transcripts with total read counts of only one or lesser were filtered out. The resulting transcripts for each of the organ system and the control samples were analyzed using DESeq2 [16] to identify differentially expressed genes (DEGs). The DEGs with the \( p \leq 0.01 \), and with a log2 fold change of \( \geq 1.5 \) and \( \leq -1.5 \) was considered statistically significant up and down-regulated gene products respectively. Further, the genes significantly regulated only in one organ system or in multiple organ systems were predicted.

**Construction and analysis of protein interactions map**

Experimentally verified human protein-protein interactions (PPIs) were retrieved from the Database of Interacting Proteins (DIP) database [17]. From these PPIs, the interactions among the up or down-regulated gene products were extracted and the protein interactions of a specific organ system, multiple organ systems as well as cross-organ protein interactions were recognized. An interaction map of the resulted PPIs was constructed using R igraph [18]. The nodes in the interactions map were colored differently to distinguish the organ-specific, cross-organ, and the multi-organ protein interactions. Topological properties of the interactions map were analyzed to interpret the biological significance of the interactions map.

**Functional annotations and pathways of modules**

The largest component in the protein interactions map was unwrapped as a core interactions sub-network. The core interactions sub-network was processed through edge betweenness clustering algorithm [19] to predict different modules in which the nodes were densely connected among themselves than the nodes of other modules. Each module and the nodes other than that of core interactions sub-network as a whole were employed using the PANTHER’s [20] over-representation analysis with Fisher’s exact test and Bonferroni correction for multiple testing algorithm. Further they were filtered with \( p \leq 0.05 \) and minimum three proteins per function to obtain significant functional annotations and pathways [21].

**Exploration of drug-target interactions**

Total drug protein interactions were retrieved from the MATA-DOR database [22]. The drugs targeting the proteins of the interactions map were extracted from this resource. Further, the type of drug-target protein interactions was interpreted from the results.

**Results and Discussion**

**Differentially regulated genes**

The distribution of log2 fold change values relative to the mean of DESeq2 normalized counts can be visualized in the Supplementary Fig. 1. It reveals that there were several genes which were significantly expressed in different organ systems. The total number of significant up or down-regulated genes in all the organ systems was found to be 8,326. Of these, 3,111 genes were differentially regulated in more than one organ system. A list of differentially regulated genes of all the organ systems with their log2 fold change values and the significant \( p \)-value is shown in the Supplementary Table 1. It was observed from the table that the number of differ-
entially regulated genes was the highest in liver and the least in fat; therefore, the pandemic could be severe in patients associated with liver and fat related diseases. Further, the table revealed several genes that were commonly regulated in multiple organ systems such as IGF2, ITM2C, MAPT, and PPP1R1A genes which were up-regulated in bowel, heart, jejunum, kidney and lung and ABCA3, SFTPA1, SFTPA2, SFTPB, and SLC34A2 genes which were down-regulated in all the organ systems except lung. Some genes were observed to be differently regulated such as ANK2 and CLU both of which were up-regulated in bowel and jejunum but down-regulated in marrow and placenta. A heatmap of organ-wise averaged read counts of genes differentially regulated in more than seven organ systems can be visualized in the Supplementary Fig. 2. It shows that the gene expression counts of lung were the most contradictory to the control samples suggesting that the lungs were severely affected than the other organ systems in the COVID-19 infection. ACE2, the angiotensin 1 converting enzyme 2, was observed to be up-regulated only in heart. CD147 (BSG), a transmembrane protein of the Ig superfamilly, was observed to be up-regulated both in the heart and the marrow while TMPRSS2, a transmembrane protein of serine protease family was significantly down-regulated in bowel, heart, jejunum, marrow, placenta and skin. CD26 (DPP4), a functional receptor on lymphocytes, was observed to be up-regulated in placenta but down-regulated in jejunum. Therefore, differential regulation of the genes and receptors might lead to morbidity and severity of the pandemic disease.

Protein interactions map
The DIP database constituted 6,729 experimentally verified human PPIs. The significant up or down-regulated gene products were detected to engage in 608 PPIs. Of these PPIs, two were specifically observed in bowel, four in heart, five in liver, 36 in marrow, and five in placenta. One hundred and ten PPIs were observed as cross-organ protein interactions where a protein of a specific organ system interacts with a protein of another organ system. Four hundred and forty-six PPIs were observed as multi-organ protein interactions where proteins of multiple organ systems interact with proteins of other organ systems. The involvement of large number of multi-organ protein interactions suggests that the COVID-19 pandemic affects several organ systems to reach its severe pathological state; therefore, early diagnosis and treatment of the pandemic could prevent patient decompensation and thus make easy recovery. A protein interaction map can be visualized in the Fig. 1. In the protein interactions map, the organ-specific, cross-organ and the multi-organ protein interactions were easily distinguishable using color representations of the nodes viz., purple, maroon, burly-wood, orange, yellow, sea-green, tomato, sky-blue, violet, royal-blue and light-green corresponding to bowel, fat, heart, jejunum, kidney, liver, lung, marrow, placenta, skin and multi-organ systems respectively. The map constitutes 608 edges or interactions among 672 nodes or proteins. The number of isolated interactions in the protein interactions map was 77 and the number of connected components was 54. Transitivity or clustering coefficient of the entire interactions map was found to be 0.094 revealing good local connections and sparse sub-graphs. Fitting pow-er-law distribution suggested that the map is a discrete graph. Degree representing the number of interactions for a node was the highest for CAV1 with a value of 15 followed by CTNNB1 with a value of 13. The organ-wise highest degree nodes were HTR2A for bowel, CTNNB1 for fat, ERBB3 for heart, EIF4A1 for jejunum, CDC27 for kidney, ESR1 for liver, S and GINS3 for lung, CDK1 for marrow, DDB1 and TGFBR1 for placenta and KRT5 for skin. Removal of these high degree nodes would disrupt the protein interactions map significantly [23]. The degree distributions of the interactions map indicated that the node degree was decreasing with increase in the number of nodes suggesting a scale free interactions network. It can be viewed in the protein interactions map that ACE2 of heart interact with S protein of lung both of which were found to be up-regulating gene products. The S protein also interact with DPP4 suggesting different downstream regulations. The DPP4 in turn interact with PTPRC which was found to be down-regulating gene product in bowel, heart, jejunum, kidney and placenta. Thus, the map of experimentally validated protein interactions brought about several prospects for the researchers and scientists to investigate in the COVID-19 research.

Functional annotations and pathways
The largest connected component of the protein interactions map was interpreted as the core interactions sub-network and it constituted 306 edges among 265 nodes. The core interactions sub-network can be visualized in the Supplementary Fig. 3. It was observed that the top 5 highest degree nodes contained in the core interactions sub-network implying of high functional significance. The edge betweenness clustering of the core interactions sub-network produced 18 modules or clusters. Modularity of these clusters was 0.84 suggesting good clustering and the significant modular structure [24]. Functional annotation and pathways of each of these modules and of non-core proteins is provided in the Supplementary Table 2. It is perceivable from the table that each of the modules have proteins significantly enriched in similar gene ontology terms such as biological process, molecular functions, cellular components and pathways. The largest modules (cluster 2, 6, and 14) were observed to contain mostly the membrane proteins with various cell binding and signaling activities. Further, the top func-
Fig. 1. A multi-organ protein interactions map of the coronavirus disease 2019. Circles are nodes representing the proteins and the lines between them are edges representing the interactions. Organ-specific nodes are colored uniquely while the multi-organ proteins are colored light-green. Cross-organ protein interactions are interpreted by the interaction between differently colored nodes.
tional annotations and pathways of each of the cluster can be viewed in Supplementary Fig. 4. The modular structure of the core interactions sub-network is represented in the Supplementary Fig. 5. The figure clearly depicts that 36 proteins bridges different modules with 29 interconnections. AR protein bridges six different modules, CTNNB1 5, CAV1 and PIK3R1 4, CCND1 and CTNNA1 3 and CDH1, CDK1 and DDB1 bridges two different modules while 27 other proteins bridge with at least one different module demonstrating their vitality for the flow of information across several pathways.

Drug-target interactions

The number of interacting proteins mapped to MATADOR databases was 222. This indicates that the protein interactions map of COVID-19 is enriched with several significant targets with known drug candidates. Therefore, proteins of the interactions map can be further investigated for drug re-purposing strategies. The Supplementary Table 3 lists all these proteins with the sight of significant regulation, core or non-core interaction, degree, drug, MATADOR score and the type of drug-target interaction. To highlight some of the proteins, a list of drug candidates is shown in the following Table 1. CTNNB1, AR, EGFR, HTR2A, ESR1, INSR, JUN, and PDGFRB are the core and high degree nodes which could be investigating for interventions of the COVID-19 disease. Further, the experimental studies [25] showed that the interacting proteins of this study were targeted by the SARS-CoV-2 spike and other proteins. Therefore, the present study would facilitate and support the scientists and researchers to empathize the complex molecular mechanisms involving multiple organ systems associated with the COVID-19 pandemic.

Table 1. List of high degree nodes and the drug candidates

<table>
<thead>
<tr>
<th>Target</th>
<th>Organ system</th>
<th>Interaction</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTNNB1</td>
<td>Fat</td>
<td>Indirect</td>
<td>Sulindac</td>
</tr>
<tr>
<td>AR</td>
<td>Multi-organ</td>
<td>Direct</td>
<td>Aclarubicin, adapalene, aripiprazole, bezafibrate, carbamazepine, carteolol, eprosartan, isocarboxazid, losartan, norlhydroguaiaretic acid, pargyline, tazarotene, telmisartan, troglitazone, valsartan, warfarin</td>
</tr>
<tr>
<td>EGFR</td>
<td>Multi-organ</td>
<td>Direct</td>
<td>Gefitinib</td>
</tr>
<tr>
<td>HTR2A</td>
<td>Bowel</td>
<td>Direct</td>
<td>Aripiprazole, clozapine, metergoline, mianserin, olanzapine, quetiapine, risperidone, sertindole, zotepine</td>
</tr>
<tr>
<td>ESR1</td>
<td>Liver</td>
<td>Direct</td>
<td>Fulvestrant, tamoxifen, raloxifene, phenol red, estrogen, diethylstilbestrol, clomiophene citrate</td>
</tr>
<tr>
<td>INSR</td>
<td>Multi-organ</td>
<td>Direct</td>
<td>Metformin</td>
</tr>
<tr>
<td>JUN</td>
<td>Multi-organ</td>
<td>Direct</td>
<td>Nordhydroguaiaretic acid</td>
</tr>
<tr>
<td>PDGFRB</td>
<td>Multi-organ</td>
<td>Direct</td>
<td>Imatinib</td>
</tr>
</tbody>
</table>

ORCID

Dhammapal Bharne: https://orcid.org/0000-0001-6729-1699

Conflicts of Interest

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

Acknowledgments

I acknowledge Bioinformatics Infrastructure Facility (BIF) at the Department of Biotechnology and Bioinformatics, University of Hyderabad for providing the necessary facilities.

Supplementary Materials

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

References


Comparative study: nonsynonymous and synonymous substitution of SARS-CoV-2, SARS-CoV, and MERS-CoV genome

Vipan Kumar Sohpal*

Department of Chemical & Bio Engineering, Beant College of Engineering & Technology, Gurdaspur 143521, India

The direction of evolution can estimate based on the variation among nonsynonymous to synonymous substitution. The simulative study investigated the nucleotide sequence of closely related strains of respiratory syndrome viruses, codon-by-codon with maximum likelihood analysis, z selection, and the divergence time. The simulated results, dN/dS > 1 signify that an entire substitution model tends towards the hypothesis’s positive evolution. The effect of transition/transversion proportion, Z-test of selection, and the evolution associated with these respiratory syndromes, are also analyzed. Z-test of selection for neutral and positive evolution indicates lower to positive values of dN-dS (0.012, 0.019) due to multiple substitutions in a short span. Modified Nei-Gojobori (P) statistical results also favor multiple substitutions with the transition/transversion rate from 1 to 7. The divergence time analysis also supports the result of dN/dS and imparts substantiating proof of evolution. Results conclude that a positive evolution model, higher dN-dS, and transition/transversion ratio significantly analyzes the evolution trend of severe acute respiratory syndrome coronavirus 2, severe acute respiratory syndrome coronavirus, and Middle East respiratory syndrome coronavirus.

Keywords: divergence time, respiratory syndrome, synonymous and nonsynonymous change, transition/transversion ratio

Introduction

Coronaviruses (CoVs) generally disturb human beings’ respiratory tract and other mammals that causes severe respiratory infections. A previous study reveals that two extremely pathogenic human CoVs, including severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), growing from viral family, have led to worldwide epidemics at different times [1,2]. According to the World Health Organization, as of 15 Jan 2021, CoV had 13.1 million diagnosed cases causing 1.98 million deaths throughout the world [3]. A comparative study indicates that the ORF8 protein of SARS-CoV attained from host genes during evolution [4]. Comprehensive genomic analysis of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) strains and its closely related coronavirus strains shows that the ratio of nucleotide to amino acid substitutions of the spike gene is higher [5]. Lower R0 values of MERS-CoV full genome assist in tracking transmission and multiple mutations [6]. Similarly, the complete genomic sequence of the porcine hamagglutinating encephalomyelitis virus discloses the
existence of a truncated group ns2 gene [7]. Turkey coronavirus that has divergence within spike protein and provided the evidence of recombination that can directly lead to new coronaviruses [8]. The analysis was performed on 93 complete genomes of SARS-CoV-2 from the GISAID (Global Initiative for Sharing All Influenza Data) database to explore the evolution and human-to-human transmissions of SARS-CoV-2 [9]. The recombination study proved that the newly discovered MERS-CoV had obtained their spike genes from a bat coronavirus, which provide significant proof that bats represent MERS-CoV evolutionary origins [10]. Simulated 219 camels and human MERS-CoV genome sequences available in GenBank for phylogenetic analysis showed that clade B divided into B1 to B6 (each containing both human and camel strains) [11]. Viral spike glycoprotein, which recognizes a cell surface receptor, supports that the SARS-CoV-2 a recombinant of the bat coronavirus and an unknown origin coronavirus. Higher sequence homology found between SARS-CoV-2 to bat CoV RaTG13 suggested that the Chinese chrysanthemum bat is the origin of SARS-CoV2 in China [12,13]. The Vietnam B-CoV, Cuban, and Chinese strains have high nucleotide sequence similarity due to same cluster [14]. Experimental data show that SARS-CoV-2 originate from bat CoV RaTG13 under the positive selection hypothesis model having common ancestor [15]. The inequality between nonsynonymous to synonymous changes simulates using capsid protein of human herpes virus to understand the evolution [16]. MEGA software implements numerous statistical techniques for nucleotide substitution models and estimates evolutionary rates [17]. The nonsynonymous mutations of SARS-CoV-2 were isolated and evaluated for surface glycoprotein spike for amino acid alterations [18]. The protein sequence similarity of pangolin-hCoV and bat-hCoV with human coronavirus was higher than their nucleotide similarity, denoting the occurrence of more synonymous mutations in the genome [19]. It has been observed that, due to the structure of the genetic code, nonsynonymous transitions are recessive than transversions to leads radical changes in amino acids [20]. In this simulative study, four nucleotide substitution methods simulated for revealing dN/dS and improved method, assumes that the substitution rate is not equal.

The purpose of the present work is to estimate the synonymous and nonsynonymous substitution in the genome of MERS-CoV, SARS-CoV, and SARS-CoV-2 with an objective to (1) normalization dN-dS value using codon through HyPhy using maximum likelihood approach (2) Z-test of selection for estimation of positive or neutral evolution using different transition/transversion ratio (tr/tv) (3) estimate the effect of different substitution models on divergence time.

Methods

Methodology and benchmark data

Nucleotide-coding sequences of SARS-CoV-2, SARS-CoV, and MERS-CoV evaluate using the synonymous to nonsynonymous nucleotide substitutions of evolution model. The execution and assessment of substitution models assess using different statistical and simulated models under discrete conditions. MEGA used to simulated sequences SARS-CoV-2, SARS-CoV, and MERS-CoV viruses that were deriving from the National Centre for Biotechnology Information (NCBI). The genomic data for substitution analysis of SARS-CoV-2 (NC_045512.2), MERS-CoV (NC_019843.3), and SARS-CoV (FJ588686.1) were retrieved from the NCBI GenBank database and filtered using BLAST. A genomic database that contains codons of SARS-CoV-2, SARS-CoV, and MERS-CoV is assuming that an approximate statistical and nucleotide substitution method must not diverge from dN/dS ratio. Furthermore, dN/dS ratio, additional parameters were also set in the simulation, including the normalized p-value, tr/tv, and divergence time.

Results

Phylogenetic analysis

The 20 different strains of SARS-CoV and MERS-CoV are firstly assessed to detect the preliminary outgroup and similarity in genome sequence through MEGA software. The evolutionary history is inferred using the Minimum Evolution method as Fig. 1. The optimal tree with the sum of branch length is observed 60.85502865. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The minimum-evolution tree was searched using the Close-Neighbor-Interchange algorithm. The neighbor-joining algorithm is also used to generate the initial tree and analysis involved 20 nucleotide sequences. Codon positions included were 1st + 2nd + 3rd + Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 1,692 positions in the final dataset.

The evolutionary tree inferences that severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Wuhan and USA strains) are extremely different severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) (Fig. 1). The genetic evolution can detect in divergent species. So, from the phylogenetic analysis its clear distinction between three selected strains SARS-CoV-2, SARS-CoV, and MERS-CoV for synonymous-to-synonymous substitution.
Selective strength, replicated in dN/dS, differs from biological sequence to sequence, nucleotide substitution model used for genomic analyses for the selection at codons level via HyPhy. GTR (General Time Reversible model), T-Nei (Tamura-Nei model), Fels1981 (Felsenstein 1981 model), and HKY (Hasegawa-Kishino-Yano model) models simulated to analyze the effects using sets of three biological sequences of respiratory syndrome virus. 1.136, 1.541, and 1.46 average values estimated for Fels1981, HKY, T-Nei, respectively. GTR model has the same value of dN/dS as estimates for the T-Nei. The average of normalized P, dN, and dS values for three respiratory syndrome sequences computed with the same four nucleotide substitution models and estimated values of dN and dS observed revealed in Table 1.

High dN/dS ratio observed for all nucleotide substitution models (Fels1981, HKY, T-Nei, GTR models), at codons level is larger than one tends towards to positive Darwinian selection. The values

**Table 1. Simulative estimate of S, N, dS, dN, and normalized (dN-dS) average value using codon through HyPhy using ML approach of SARS-CoV-2, SARS-CoV, and MERS-CoV**

<table>
<thead>
<tr>
<th>Model</th>
<th>Syn sites (S)</th>
<th>Nonsyn sites (N)</th>
<th>dS</th>
<th>dN</th>
<th>dN-dS</th>
<th>Normalized dN-Ds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fels1981</td>
<td>1,104.58</td>
<td>4,157.41</td>
<td>1,666.33</td>
<td>1,893.72</td>
<td>227.38</td>
<td>6.7254</td>
</tr>
<tr>
<td>HKY</td>
<td>1,037.75</td>
<td>4,510.25</td>
<td>1,363</td>
<td>2,100.25</td>
<td>737.25</td>
<td>1.1325</td>
</tr>
<tr>
<td>T-Nei</td>
<td>1,010.41</td>
<td>4,657.58</td>
<td>1,507.7</td>
<td>2,202.14</td>
<td>694.43</td>
<td>2.2069</td>
</tr>
<tr>
<td>GTR</td>
<td>1,010.41</td>
<td>4,657.58</td>
<td>1,507.7</td>
<td>2,202.14</td>
<td>694.43</td>
<td>2.2069</td>
</tr>
</tbody>
</table>

ML, maximum likelihood; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SARS-CoV, severe acute respiratory syndrome coronavirus; MERS-CoV, Middle East respiratory syndrome coronavirus; Fels1981, Felsenstein 1981 model; HKY, Hasegawa-Kishino-Yano model; T-Nei, Tamura-Nei model; GTR, General Time Reversible model.
lie in the range from 1.136 to 1.541 as reproduced in Fig. 2. Similarly, higher dN-dS and positive values of statistic condition show an overabundance of nucleotide substitutions. The peak value dN-dS observe 737.25, and the lowest 227.38 corresponding to the HKY model and Fels1981. An intermediate value of dN-dS was found 694.43 for GTR and T-Nei model. A similar outcome obtains for normalized dN-dS, which assesses the expected substitutions per site. Results point out towards data sets of respiratory syndrome virus, normalized dN-dS, dN/dS ratio, and dN-dS for validating the positive selection hypothesis.

**Z-test of selection**

Higher value of nonsynonymous to synonymous difference and ratio plays a significant role in the assessed positive evolution model. Z-test of selection is comparing the relative abundance of nonsynonymous to synonymous substitutions that occurred in the codon site. Results point out towards data sets of respiratory syndrome virus, normalized dN-dS, dN/dS ratio, and dN-dS for validating the positive selection hypothesis.

**Table 2.** Average estimations (dN-dS) of Z-codon-based test selection for all sequence pairs

<table>
<thead>
<tr>
<th>Evolutionary model</th>
<th>Mathematical correlation</th>
<th>dN = dS Neutral</th>
<th>dN &gt; dS Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>NG (P) {pS = Sd/Sk, pN = Sd/SR}</td>
<td>dN = -3/4 ln*(1-4/3pN)</td>
<td>-1.397</td>
<td>-1.419</td>
</tr>
<tr>
<td>NG (JC) {pS = Sd/Sk, pN = Sd/SR}</td>
<td>dS = -3/4 ln*(1-4/3pS)</td>
<td>0.034</td>
<td>0.043</td>
</tr>
<tr>
<td>LWL (K-2)</td>
<td>dS = 3L2A1 + L0(A1 + B0)</td>
<td>0.022</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>(L2 + 3L0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>dN = 3L2B1 + L0(A0 + B0)</td>
<td>0.012</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>(2L2 + 3L0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PBL (K2)</td>
<td>dS = B4(L2A2 + L2A0)</td>
<td>0.012</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>(L2 + L0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>dN = A0(L0B0 + L0B2)</td>
<td>0.012</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>(L0 + L2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kumar method (K-2)</td>
<td>P0 = SdS + SdN</td>
<td>0.811</td>
<td>0.831</td>
</tr>
<tr>
<td></td>
<td>L0 + L2C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P1 = SdS + SdS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L2S + L2C</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of synonymous differences (Sd) and nonsynonymous (Nd) differences. Normalized number of synonymous sites (S) and nonsynonymous sites (N), numbers of synonymous (Sd) to normalize using the number of potential synonymous sites (Sk), L0, L2, and L4 are the number of 0-fold, 2-fold, and 4-fold degenerate sites, respectively. L0, L2, and L4 are the numbers of 0-fold, simple 2-fold, complex 2-fold, and 4-fold degenerate sites.

NG (P), Nei-Gojobori method (proportion); NG (JC), Nei-Gojobori method (Juke Cantor); LWL, Li-Wu-Luo method (K-2); PBL, Pamilo-Bianchi-Li method (K-2).
(dN) to synonymous distance (dS), under the assumption that synonymous substitutions in the coding region are selectively neutral. In NG (P) methods, the main reason for this negative correlation is that there was a substantial proportion of codons with non-zero nonsynonymous difference but zero synonymous difference. The negative correlation appears to result mainly from three factors: (1) the occurrence of nucleotide differences varies stochastically across codons, (2) selection of purifying selection eliminating number nonsynonymous mutations, the rate of occurrence of nonsynonymous differences is much lower than that of synonymous differences, and (3) the overall rate of substitution is much less than one substitution per site. In the case of LWL (K-2), PBL (K 2), and Kumar method (K-2), 0-fold, 2-fold, and 4-fold degenerate codons, the proportion of codons with a non-zero nonsynonymous difference and the non-zero synonymous difference is higher.

**Effect of tr/tv**

The effects of the tr/tv with a set of three sequences of SARS-CoV-2, SARS-CoV, and MERS-CoV using MEGA were also investigated. The dN-dS distance calculates with the MNJ (P) method for two hypotheses of evolution plotted in Fig. 3. The tr/tv was assessed from 1.0 to 7.0. Positive evolution is less biased than neutral evolution, but both have approximately equivalent inclinations with an increasing tr/tv.

MNJ (P) gives the dN-dS value ranges from 9.5 to 30.21 and 9.397 to 29.66 for null and positive evolution models. The maximum dN-dS values observe for both methods, when tr/tv approach 6. With an increasing tr/tv (lower to higher), the results are changed moderately in both the evolutionary models. The tr/tv for genome restricted to 1 to 6, exceed than the upper limit no influence observes on dN-dS. We have made several observations to avoid divergence and to maintain consistency in evolution models. This is quite remarkable that neutral and positive evolution models show a comparatively similar performance.

**Effect of divergence time**

The divergence time $t = f(dS/dN)$ of nucleotide substitutions per codon. Three taxa of SARS-CoV-2, SARS-CoV, and MERS-CoV having an average of 1,181 codons that were used for simulation and tested four different models. The average estimates of divergence time against Fels 1981, HKY, T-Nei, and GTR are shown in Fig. 4. T-Nei and GTR models have nearly equal divergence times. Fels1981 is closely linked to the HKY model having a difference in divergence time of about 0.87% only, whereas the HKY model departs 5.58% from the T-Nei and GTR model in terms of divergence time value.

With increasing divergence time, dN-dS for positive evolution support by GTR and T-Nei, and biased approximation for neutral evolution model selection, while purifying evolution model remains unaffected by divergence. Figs. 2 and 4 compare the co-relation between dN/dS ratio raise and divergence time and directly proportional to each other. The outcomes of dN/dS indicate high divergence time, force towards the positive evolution model.

**Discussion**

Numerous substitution models provide the platform for the comparative study of evolution for SARS-CoV-2, SARS-CoV, and MERS-CoV strains. The critical area is to analyze the synonymous

---

https://doi.org/10.5808/gi.20058
and nonsynonymous substitution, divergence time, and the substitution ratio of nucleotide for the concept of evolution. Comparative analyses point out that overlook of substitution rate close to estimates of dN/dS parameters effect to evolutionary analysis. HKY model is robust in simulations and real datasets, especially concerning the Fels1981 model. The whole models have dN/dS > 1 expect diversity at the codon level is favored, to the positive evolution using the mutations. In Z-test selection of codon observed that without nucleotides substitution rate, positive evolution pronounced in respiratory syndrome. Since transitions are more dominant than transversions, this indicates positive selection in a short span due to multiple substitutions. Consequently, it is essential to take account of transition/transversion rates to correctly confine the evolutionary information when unequal transitional ratios among respiratory syndrome sequences exist. SARS-CoV-2, SARS-CoV, and MERS-CoV simulation results demonstrate that lower to higher transition/transversion ratios favor the positive evolution due to synonymous substitution as recessive compared to nonsynonymous. The proximity-based on the p-distance between viral strains implies SARS-CoV-2 have substitution rate higher than SARS-CoV, while MERS-CoV stands in midway. Divergence time, dN, and dS output parameters are data mining from 4-fold degenerate sites at the 3-codon positions and noncoding sites. The consequences explain that dN-dS are inversely proportional to divergence time, and dN substitution favors positive evolution. The above discussion is based on the estimate selection for HyPhy and Z-test results for three different respiratory syndrome strains that also agree with positive evolution and the average distance between them.

Compared nucleotide substitution models, especially HKY and T-Nei, Fels1981, and GTR for nucleotide sequences and establish that dN/dS ratio goes beyond than one, tend towards positive selection. It has been concluded that in the lack of transition/transversion, synonymous and nonsynonymous substitution rates tending to positive evolution for entire the methods NG (JC), LWL, PBL, and Kumar. The comparative results obtained from consistent and continuous analysis signify that higher transition/transversion rates show lesser dN-dS under positive evolution and reject the null hypothesis. Moreover, it is also observed that substitution models and dN-dS having a high impact on divergence time.

**ORCID**

Vipan Kumar Sohpal: https://orcid.org/0000-0003-1009-4920

**Conflicts of Interest**

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

**Acknowledgments**

This work was supported by Beant College of Engineering & Technology, Gurudaspur, India, and Dr. Sudhir Kumar, University of Arizona, US for MEGA X software. It is available free of charge on http://www.megasoftware.net

**References**


A network–biology approach for identification of key genes and pathways involved in malignant peritoneal mesothelioma

A. M. U. B. Mahfuz1*, A. M. Zubair-Bin-Mahfuj2, Dibya Joti Podder3

1Department of Biotechnology & Genetic Engineering, Faculty of Life Science, University of Development Alternative, Dhaka 1209, Bangladesh
2Department of Oral and Maxillofacial Surgery, Dhaka Dental College, Dhaka 1216, Bangladesh
3Department of General Surgery, Sher-E-Bangla Medical College, Barishal 8200, Bangladesh

Even in the current age of advanced medicine, the prognosis of malignant peritoneal mesothelioma (MPM) remains abysmal. Molecular mechanisms responsible for the initiation and progression of MPM are still largely not understood. Adopting an integrated bioinformatics approach, this study aims to identify the key genes and pathways responsible for MPM. Genes that are differentially expressed in MPM in comparison with the peritoneum of healthy controls have been identified by analyzing a microarray gene expression dataset. Gene Ontology and Kyoto Encyclopedia of Genes and Genomes pathway analyses of these differentially expressed genes (DEG) were conducted to gain a better insight. A protein–protein interaction (PPI) network of the proteins encoded by the DEGs was constructed using STRING and hub genes were detected analyzing this network. Next, the transcription factors and miRNAs that have possible regulatory roles on the hub genes were detected. Finally, survival analyses based on the hub genes were conducted using the GEPIA2 web server. Six hundred six genes were found to be differentially expressed in MPM; 133 are upregulated and 473 are downregulated. Analyzing the STRING generated PPI network, six dense modules and 12 hub genes were identified. Fifteen transcription factors and 10 miRNAs were identified to have the most extensive regulatory functions on the DEGs. Through bioinformatics analyses, this work provides an insight into the potential genes and pathways involved in MPM.

Keywords: differentially expressed genes (DEGs), hub genes, malignant peritoneal mesothelioma, miRNA, PPI network, transcription factor

Introduction

Peritoneum is the serous membrane covering the abdominal cavity and organs and is lined by a layer of simple squamous epithelium. These epithelial cells are called mesothelium [1]. Malignant peritoneal mesothelioma (MPM) is the malignancy of peritoneal mesothelial cells and is a relatively rare disease but with a very poor prognosis [2]. The 5-year relative survival rate of MPM patients is only 10% [3]. MPM is related to industrial pollutants and mineral exposure (asbestos accounts for 33%-50% of cases, others include erionite, thorium, and mica). However, the duration between exposure and disease causation is variable. Other risk factors for MPM include familial Mediterranean fever and diffuse lymphocytic lymphoma. It usually spreads and remains within the abdominal cavity but in few cases may metastasize outside the abdomen [2,4]. MPM patients pres-
ent with atypical symptoms in most cases and this is responsible for the delayed diagnosis (average time to diagnose takes 4–6 months) of this fatal disease. Abdominal symptoms include abdominal distension, ascites, abdominal pain, early satiety, intestinal obstruction, and intestinal perforation. Other non-specific symptoms include weight loss, anorexia, nausea, night sweats, and unexplained fever [2,4]. MPM may also be responsible for paraneoplastic syndrome exhibiting thrombocytosis, venous thrombosis, hypoglycemia, paraneoplastic hepatopathy, wasting of muscles, and adipose tissue [2]. Surgical intervention is the first line of treatment for MPM patients. Cytoreductive surgery with heated intraperitoneal chemotherapy is currently the most preferred treatment option. Systemic chemotherapy is chosen for those MPM patients who are unable to undergo surgery [4].

In recent years, there has been rapid advancement in microarray and RNA-sequencing (RNA-seq) technologies, and analysis of the large amount of data obtained from them has shed light on complex biological processes in an unprecedented manner. Key genes and pathways involved in different cancers and diseases have been identified by analyzing these data. These key genes can be used as biomarkers and can be utilized for early diagnosis, survival prediction, drug target identification, and drug response observation [5-11]. Bioinformatics analyses have been employed to identify important genes and pathways involved in the disease process of malignant pleural mesothelioma [12-14]. Bioinformatics approach has also identified upregulation of spliceosomal genes, especially SF3B1 [15] and haploinsufficiency of BAPI gene [16] associated with MPM. In this study, we have identified the significantly overexpressed and underexpressed genes in MPM by bioinformatics analyses.

Methods

Retrieval of microarray data

Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) [17], a database of gene/microarray profiles with unrestricted public access hosted by NCBI was searched for MPM data. The following filters were applied while searching: ‘Expression profiling by array’ as the study type, ‘Homo sapiens’ as the organism, and the publication date to be within the last eight years. Only one MPM dataset (accession No. GSE112154) was retrieved from the search. Sciarrillo et al. [15] deposited this dataset to the GEO. They utilized Illumina HumanHT-12 V4.0 expression bead-chip to obtain this gene expression profiling dataset and the dataset is based on the GPL10558 platform. Series matrix file of GSE112154 was downloaded for subsequent analyses.

Identification of differentially expressed genes

NetworkAnalyst (https://www.networkanalyst.ca/) [18], a web tool dedicated to the analysis of gene expression data was employed for analyzing the dataset of our interest. Forty-seven samples (45 malignant peritoneal mesothelioma samples and two peritoneal mesothelioma cell lines) were classified as ‘MPM’ and the rest three healthy peritoneal samples were classified as ‘Normal Mesothelium’ to make them compatible for analysis. Illumina probe IDs were converted by NetworkAnalyst to their corresponding Entrez gene IDs and official gene symbols. Probes corresponding to unannotated genes were filtered out and for multiple probes mapped to the same genes, their average expression values were considered. To obtain statistically significant results, data with the lowest 15th percentile expression and data with relative abundance lower than the 5th percentile were discarded from downstream analyses using the ‘variance’ and ‘low abundance’ filters. The dataset was quantile normalized followed by box and whisker plot visualization. After quantile normalization, data quality was assessed through three-dimensional principal component analysis (PCA).

Limma (linear models for microarray data) [19], an R package for differential expression analysis of microarray data, embedded in the NetworkAnalyst server was exploited to identify the differentially expressed genes (DEGs). Genes having an adjusted p-value (Benjamini-Hochberg method) < 0.05 and a |log2FC| value ≥ 1.5 were considered as differentially expressed.

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes pathway enrichment analyses:

Enrichment analyses of the DEGs were carried out using Enrichr (https://maayanlab.cloud/Enrichr/). Enrichr provides a wide range of annotations curated from other databases and annotation tools for the submitted genes [20]. A list containing official gene symbols of the DEGs was used as the input. Gene Ontology (GO) biological process, molecular function, cellular component, and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways annotations for the DEGs were retrieved through Enrichr for an insight into the roles played by these DEGs. GO [21,22] provides annotations of gene products regarding their complex biological processes, molecular functions, and cellular distributions. KEGG provides molecular-level information about large-scale biological data obtained from genome sequences and other high-throughput experiments [23]. KEGG pathway database provides maps of molecular interaction, reaction, and relation networks relevant to cellular metabolism, genetic and environmental information processing, cellular processes, organismal systems, human diseases, and drug development. GO and KEGG pathway annotations having an adjusted p-value < 0.05 were considered statistically significant.
Gene set enrichment analysis
Gene set enrichment analysis (GSEA) can be performed on DNA microarray or RNA-seq data to identify biologically linked enriched gene sets. It is a widely used approach where a priori gene sets grouped on the basis of their common biological functions, proximity in chromosomal locations, or participation in identical biological pathways are used to detect enriched gene clusters differentially expressed in two different conditions or cell types. Here the focus is placed on sets of genes rather than on individual genes [24]. The DEGs identified in our study were ranked according to their [logFC] values in a .rnk file and GSEA was performed employing the GSEA Preranked module of GenePattern platform (https://www.genepattern.org/) using this file [25]. A priori annotated gene sets for conducting GSEA were retrieved from the Molecular Signatures Database (MsigDB) [26]. The c2.cp.kegg.v7.4.symbols.gmt dataset was selected as the reference gene set database. The number of gene set permutations was set to 1,000 and other parameters were used as default. In the ‘collapse dataset’ option ‘No_Collapse’ was chosen since we used official gene symbols. Gene sets having a false discovery rate (FDR) q-value < 0.25 were considered significantly enriched.

Protein-protein interaction network construction and identification of significant modules and hub genes
The network of interactions among the protein products of the DEGs was obtained utilizing STRING (Search Tool for the Retrieval of INteracting Genes) (https://string-db.org/cgi/input?page_input_active_form=multiple_sequences) [27]. The STRING database was searched with medium confidence (interaction score cutoff was 0.4) and the protein-protein interaction (PPI) network was visualized using Cytoscape 3.6.1 [28]. Applying the degree cutoff, node score cutoff, K-core, and maximum depth as 2, 0.2, 2, and 100, respectively, Molecular Complex Detection (MCODE) [29], a Cytoscape plug-in, was used to identify the modules with significant densities in the PPI network. Modules having an MCODE score ≥ 4 were considered important. CytoHubba [30], another Cytoscape plug-in was utilized for topological analysis to identify the nodes representing the hub proteins in the PPI network. CytoHubba allows to apply different methods of calculation for identifying hub nodes. 10 methods available in CytoHubba, namely ‘Betweenness,’ ‘Bottle-Neck,’ ‘Closeness,’ ‘Degree,’ ‘EcCentricity,’ ‘EPC’ (Edge Percolated Component), ‘MCC’ (Maximal Clique Centrality), ‘MNC’ (Maximum Neighborhood Component), ‘Radiality,’ and ‘Stress’ were applied to detect the top 50 hub nodes. Twelve proteins were ultimately identified as hub proteins from the consensus of all methods and their corresponding genes were considered as the hub genes.

Identification of transcription factors acting on DEGs and miRNAs acting on DEGs and transcription factors
Transcription factors (TF) and microRNAs (miRNAs) are the two master regulators of gene expression. Cellular levels of TFs and miRNAs are influenced by each other in normal cells and their complex interplay controls the expression of common gene targets through feedback and feedforward loops [31]. miRNAs can bind to 3’ untranslated retion (UTR), 5’ UTR, promoter region, or even coding sequence of a gene to either suppress gene expression or induce expression. miRNAs can cause gene silencing by binding to 3’ UTR, 5’ UTR, and coding sequence, and can induce transcription by binding to the promoter region. They can also regulate gene expression within the nucleus at the time of or after transcription but the detailed mechanism of this intra-nuclear regulation by miRNAs is not yet fully understood [32]. In our study, TF-DEG, miRNA-DEG, and miRNA-TF interaction networks were identified using the miRNet web server (https://www.mirnet.ca/) [33]. miRNet is a curated database of miRNA interactions from 14 different miRNA databases. Official gene symbols of the hub genes were used as inputs and ChEA (ChIP Enrichment Analysis) was chosen as the TF database. ChEA provides data on genome-wide target specific TFs deduced from the chromatin immunoprecipitation (ChIP) followed by microarray hybridization, ChIP followed by high-throughput sequencing, ChIP with paired-end tag sequencing, and DNA adenine methyltransferase identification [34].

Identification of enriched kinases
Kinase enzymes phosphorylate proteins by transferring a phosphate group and phosphatase enzymes can dephosphorylate proteins, thus reversing the function of kinases. Their coordinated actions make possible many normal cellular processes. Dysregulated kinases and deactivated phosphatases have significant roles in different malignancies, and kinase inhibitors are promising anticancer drugs [35]. Kinases that can phosphorylate the top TFs regulating the DEGs and thereby affect their expression level were identified employing the KEA2 (Kinase Enrichment Analysis 2) webserver (https://www.maayanlab.net/KEA2/) [36]. A list of the top TFs was submitted as input to the KEA2 server. KEA2 hosts various phosphosite and protein level libraries that are either manually curated from kinase-substrate interactions in the literature or experiment-derived data. A kinase enrichment analysis can be performed against these libraries to prioritize kinases phosphorylating the query proteins. ‘Literature based kinase-substrate library’ was chosen to identify enriched kinases.
Survival analysis of hub genes
Correlation between hub gene expression and survival rate was analyzed employing GEPIA2 (http://gepia2.cancer-pku.cn; index) [37], a web server for gene expression analysis based on the RNA-seq data of 9,736 tumors and 8,587 normal samples from the TCGA (The Cancer Genome Atlas) and the GTExs (Genotype-Tissue Expression) projects. Since TCGA provides data for only malignant pleural mesothelioma [38] and GTEx doesn’t contain gene expression information of mesothelium, we choose the MESO (malignant pleural mesothelioma) dataset available in the GEPIA2 server to extrapolate the relationship between the expression level of hub genes and prognosis from pleural mesothelioma data. Hub gene names were used as inputs to GEPIA2. During survival analyses, median values were chosen as the group cutoff values, hazard ratios (HRs) were calculated based on Cox Proportional-Hazards Model and the analyses were conducted with a 95% confidence interval.

Identification of protein-drug interactions
Approved drugs, investigational and experimental compounds that can interact with the identified hub proteins were identified by searching the DrugBank knowledgebase (version 5.1.7, released 2020-07-02; available at https://go.drugbank.com/) [39]. This is a rich database that currently provides information on 13,791 drug entries of which 2,653 are approved small molecule drugs, 1,417 are approved biologics, 131 are nutraceuticals and more than 6,451 are experimental drugs. Information about 5,236 non-redundant protein (i.e., drug target/enzyme/transporter/carrier) sequences related to these entries is also available through DrugBank. Official symbols of the hub proteins were used as inputs.

Results
Identification of differentially expressed genes
The means of the microarray samples were found uniform after quantile normalization (Supplementary Fig. 1) and the PCA plot showed MPM and normal mesothelium samples arrange in different clusters (Supplementary Fig. 2). A total number of 608 DEGs were identified from the analysis. Among these 608 DEGs, the records of two genes (Entrez gene ID: 100302207 and 100302173) have been withdrawn by the HGNC (HUGO Gene Nomenclature Committee) and these genes were excluded from subsequent analyses. Among the rest 606 DEGs, 133 genes are upregulated and 473 genes are downregulated. Fig. 1 depicts a volcano plot representing the up- and downregulated genes. This volcano plot was generated using VolcaNoseR (https://huygens.science.uva.nl/VolcaNoseR/) [40].

GO and KEGG pathway enrichment analyses
GO and KEGG pathway enrichment analyses were performed using the Enrichr web server. Biological processes, molecular functions, cellular locations, and biological pathways enriched in DEGs are shown in Supplementary Tables 1-4. From the GO biological processes, it was found that the DEGs significantly participate in the regulation of angiogenesis, negative regulation of cell proliferation, sprouting angiogenesis, negative regulation of cellular processes, regulation of cell proliferation, negative regulation of angiogenesis, regulation of vasculature development, regulation of inflammatory response, negative regulation of blood vessel morphogenesis, and in negative regulation of cell adhesion. Angiogenesis, cellular proliferation, inflammation, and cellular adhesion are critical events for tumorigenesis and metastasis. From the GO molecular functions, it was found that the DEGs mainly involve in cytokine activity, calcium ion binding, integrin binding, metal ion binding, oxidoreductase activity, chemokine activity, coreceptor activity involved in the Wnt signaling pathway, planar cell polarity pathway, lipoprotein particle binding, and chemokine receptor binding. From the GO cellular component, it was found that the proteins encoded by the DEGs are chiefly distributed in lipid droplets, membrane rafts, microvilli, actin-based cell projections, cytoskeleton, endoplasmic reticulum lumen, integral components of the plasma membrane, sarcoplasmic reticulums, perinuclear regions of cytoplasm, and caveolae. From the KEGG pathway enrichment analysis, the DEGs were found to be enriched in Malaria, PPAR signaling pathway, lipolysis regulation in adipocytes, cytokine-cytokine receptor interaction, AMPK signaling pathway, AGE-RAGE signaling pathway in diabetic complications, pathways in cancer, cell adhesion molecules, longevity regulating pathway, thyroid hormone synthesis, glycerolipid metabolism, and PI3K-Akt signaling pathway. The top 10 GO terms and enriched pathways according to p-value are depicted in Fig. 2.

Gene set enrichment analysis
GSEA was performed to corroborate the enrichment results from Enrichr. Only one pathway (cytokine-cytokine receptor interaction) was identified as downregulated from the GSEA with an FDR < 25% (FDR q = 0.0788609 and nominal p = 0.043296088) (Fig. 3). Among the 24 genes identified by Enrichr to be involved in cytokine-cytokine receptor interaction pathway (ACVRL1, CCL14, GDF10, CXCL8, OSM, LIFR, INHBB, PPBP, NGF, CXCL14, CXCL2, CXCL1, BMP6, CXCL5, BMP5, GHR, IL-1RL1, IL6, IL18RAP, ACVR1C, CCL8, LEP, LEPR, and IL17D), 17 are included in the GSEA result (OSM, LEPR, ACVRL1, CXCL5, LIFR, IL18RAP, CCL14, INHBB, GHR, PPBP, CX3CL1, CXCL8, LEP, CCL8, CXCL2, CXCL14, and IL6). Two underex-
pressed hub genes (IL6 and CXCL8) are included in the GSEA outcome which further emphasizes their importance.

**PPI network construction and identification of significant modules and hub genes**

STRING constructed the PPI network for the protein products of all the 606 DEGs. The number of nodes and edges in the network are 570 and 2,573, respectively. The average node degree of the network is 9.03, the average local clustering coefficient is 0.41 and the PPI enrichment p-value is < 1.0e-16. The PPI network was found to have significantly more interactions than expected which is an indication of their probable biological inter-connections as a group. The PPI network was visualized in Cytoscape and the significantly dense modules were detected by the MCODE plug-in of Cytoscape. Six dense modules having an MCODE score ≥ 4 were identified. The first module has 38 nodes, 217 edges, and has an MCODE score of 11.73. The second and third modules have 51 and 29 nodes, respectively, and 224 and 79 edges, respectively. Their MCODE scores are 8.96 and 5.643, respectively. The fourth and fifth modules have five nodes, 10 edges, and their MCODE score is 5.0. The sixth module has 17 nodes, 35 edges and its MCODE score is 4.375. Proteins in each module are listed in Table 1 and are graphically presented in Fig. 4.

CytoHubba plug-in of Cytoscape was next employed to detect the top 50 hub proteins in the network. Ten available calculation methods in CytoHubba for detecting hub proteins were used. Multiple List Comparator (http://www.molbiotools.com/list-compare.html) was utilized to identify their intersections. Twelve
proteins were identified as hub proteins by all the methods and their encoding genes were considered as the high-confidence key genes (Table 2). Only two of the 12 hub genes were found upregulated (CDH1 and GAPDH) and the rest 10 hub genes were found downregulated. CXCL8, PTPGS2, and FGF2 (fibroblast growth factor 2) were found to be present in MCODE module 1, IL6, CDHS, VWF, TEK, MYC, and CDH1 in module 2, PPARG and GAPDH in module 3, and ADIPOQ in module 6. The hub proteins were next submitted to STRING to identify their interactions, which were then visualized by Cytoscape. It was found from STRING analysis that the hub proteins have significant interactions among themselves. Among the 12 hub proteins, IL6, CXCL8, FGF2, and GAPDH are each connected with the rest 11 hub proteins (Fig. 5).

Identification of TFs acting on DEGs and miRNAs acting on DEGs and TF

TFs that can act on the DEGs and miRNAs that can regulate the DEGs and TFs were identified, and the DEG-TF-miRNA interaction network was constructed and visualized through miRNet (Fig. 6A). miRNet identified 197 TFs and 2,305 miRNAs. Fifteen top TFs with a degree cutoff of 180 and 10 top miRNAs with a degree cutoff of 150 were identified. Among the 12 hub proteins, IL6, CXCL8, FGF2, and GAPDH are each connected with the rest 11 hub proteins (Fig. 5).

Identification of enriched kinases

Enriched kinases having interactions with the identified TFs were detected through the KEA2 web server. A total number of 20 kinases having an adjusted p-value (FDR) < 0.1 were detected. Among them, significant kinases are GSK3B (6 substrates), MAPK14 (6 substrates), MAPK1 (5 substrates), CSNK2A1 (4 substrates), MAPK8 (4 substrates), HIPK2 (3 substrates), PRKACB (3 substrates), PRKACB (3 substrates), and PDK1 (3 substrates), PRKACB (3 substrates), PRKACB (3 substrates), and PDK1 (3 substrates).


caption: Fig. 3. Gene set enrichment analysis identified cytokine-cytokine receptor interaction pathway is significantly disturbed in malignant peritoneal mesothelioma. KEGG, Kyoto Encyclopedia of Genes and Genomes.
Table 1. Significant modules in the PPI having an MCODE score ≥ 4

<table>
<thead>
<tr>
<th>Module</th>
<th>Nodes</th>
<th>Edges</th>
<th>MCODE score</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38</td>
<td>217</td>
<td>11.73</td>
<td>TF, GPC3, PDK4, MSLN, ACACB, CLDN5, NES, FGF2, CIDEA, SDC2, CDH2, CEBPA, AGPAT2, CXCL8, CXCL5, PNP, LA2, CX3CL1, CIDEA, SOD2, LPL, GPT, PLIN1, SPARC1, PLIN2, HRC, GPAM, SCG, FABP4, LIPE, TMEM132A, PTGS2, DGA2, ACSL1, ABHD6, CD36, MLXIP, MFGE8, PCK1</td>
</tr>
<tr>
<td>2</td>
<td>51</td>
<td>224</td>
<td>8.96</td>
<td>VTN, MYC, SLCA2A, SELP, MCAM, HBEGF, NAMPT, CDH5, PTPRB, HMOX1, CAV1, KRT8, MRAP, SNAI2, CDC20, SQX17, FDE3B, STOM, ALDH1B1, UBE2C, CAT, KLHL21, LPAR5, S1PR3, CXCL2, MYLIP, FBXO2, ZBTB16, SPSB1, VAMP8, S1PR1, FBXO7, HECV2, PPBP, NQF, LDLR, AGTR1, PECAM1, VWF, ADCY4, ANGPT1, CDH1, RBP4, MMP25, IL6, STAT5A, CDF, SELE, LEPR</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>79</td>
<td>5.643</td>
<td>THRSP, SOD3, KLF4, GPX3, GPX8, ADRB2, PTGER4, NOX4, CA9, THBD, RAM2, INSIG1, NOP3, ASPM, NUSAP1, PPARG, TOP2A, KRT19, CEP55, NEK2, PARBP2, CALCRL, VIPR1, S100A4, GAPDH, PTH1R, COL1A1, GPBAR1, LEP</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>10</td>
<td>5</td>
<td>PIK3R2, NMB, PROK2, EDNRB, GPR4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>10</td>
<td>5</td>
<td>HBA2, HBB, HBA1, HBG2, HBO1</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>35</td>
<td>4.375</td>
<td>ADAMTS5, HSPH1, DNAJB1, HSPB7, ATF4, FOXO1, BAG3, PTX3, TIMP1, ANG, MMP24, HSPAA1B, ADIPOQ, HSP-A12B, DNAJA4, OSM, EBF1</td>
</tr>
</tbody>
</table>

PPI, protein-protein interaction; MOCDE, Molecular Complex Detection.

Table 2. Twelve hub genes with their respective log_{2}FC and adjusted p-values (FDR)

<table>
<thead>
<tr>
<th>Hub gene</th>
<th>Log_{2}FC</th>
<th>Adj. p-value (FDR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDH1</td>
<td>4.1017</td>
<td>0.00074857</td>
</tr>
<tr>
<td>GAPDH</td>
<td>1.5141</td>
<td>0.0020854</td>
</tr>
<tr>
<td>FGF2</td>
<td>-1.7636</td>
<td>0.001524</td>
</tr>
<tr>
<td>MYC</td>
<td>-2.0953</td>
<td>0.020572</td>
</tr>
<tr>
<td>PTGS2</td>
<td>-2.4941</td>
<td>0.014657</td>
</tr>
<tr>
<td>TEK</td>
<td>-2.7234</td>
<td>1.47E-05</td>
</tr>
<tr>
<td>VWF</td>
<td>-2.7452</td>
<td>0.025154</td>
</tr>
<tr>
<td>CDH5</td>
<td>-2.7455</td>
<td>0.0011496</td>
</tr>
<tr>
<td>CXCL8</td>
<td>-3.0163</td>
<td>0.002552</td>
</tr>
<tr>
<td>ADIPOQ</td>
<td>-3.0367</td>
<td>1.04E-15</td>
</tr>
<tr>
<td>PPARG</td>
<td>-3.0666</td>
<td>1.10E-07</td>
</tr>
<tr>
<td>IL6</td>
<td>-4.1746</td>
<td>0.0001007</td>
</tr>
</tbody>
</table>

FDR, false discovery rate.

Survival analysis of hub genes

To elucidate the relationship between hub gene expression level and patient survival, survival analyses for the hub genes were performed through GEPIA2 (Fig. 7). From the survival analyses, it was found that with a log-rank p < 0.05, increased expression levels of one upregulated hub gene, GAPDH (HR, 2.3; p = 0.00061), and three downregulated hub genes, namely, TEK (HR, 1.8; p = 0.016), VWF (HR, 2; p = 0.0056) and CDH5 (HR, 1.9; p = 0.0089) are associated with markedly decreased overall survival duration. Moreover, Cox regression analyses of these genes indicated that these genes have high HRs (2.3, 1.8, 2, and 1.9) and can be considered as prognostic factors. These hub genes (GAPDH, TEK, VWF, and CDH5) can serve as survival biomarkers for MPM also.

Identification of hub protein-drug interactions

The approved drugs or the compounds that can interact with the hub proteins were identified through searching DrugBank. A total number of 237 drug or drug-like compounds were found to act on the hub proteins. Seven compounds were found to act on GAPDH, eight compounds with FGF2, 2 compounds with MYC, 109 compounds with PTGS2, seven compounds with TEK, 11 compounds with VWF, two compounds with CDH5, four compounds with CXCL8, 74 compounds with PPARG, and 13 compounds with IL6. No drug/compound was found to act on CDH1 and ADIPOQ. Among the identified compounds, there are agonists, antagonists, and compounds with still unknown pharmacological actions. For the two upregulated hub proteins (CDH1 and GAPDH), no antagonist/inhibitor was found. A full list of the identified compounds is available in Supplementary Table 5 and the hub protein-drug interaction network can be found in Supplementary Fig. 3.

Discussion

MPM is an aggressive disease and its prognosis is usually very poor. Identification of biomarkers of this disease can help in early diagnosis and treatment, and observing responses to ongoing treatment. For this purpose, in this study, we have analyzed GSE112154, a microarray dataset containing gene expression information of normal peritoneum and MPM. We have found a total...
number of 606 genes are differentially expressed in MPM (133 genes are upregulated and 473 genes are downregulated) with an adjusted p < 0.05 and a $|\log_2 FC| \geq 1.5$. These DEGs were next subjected to GO and KEGG pathway enrichment analyses followed by GSEA. From the GSEA, it was found that many genes involved in the cytokine-cytokine receptor interaction pathway are significantly downregulated in MPM, thereby interfering with the normal functions of this pathway.

Six significant modules and 12 hub genes (CDH1, GAPDH, FGF2, MYC, PTGS2, TEK, VWF, CDH5, CXCL8, ADIPOQ, PPARG, and IL6) were identified from our analyses. Two of these hub genes, CDH1 and GAPDH, are overexpressed and the rest 10 hub genes are underexpressed. CDH1 or E-cadherin’s main role is in cell-cell adhesion. CDH1 is considered a tumor suppressor gene [41] but was found overexpressed in epithelioid malignant pleural mesothelioma [42]. GAPDH (glyceraldehyde-3-phosphate dehydrogenase) is a glycolysis enzyme that can be found in all tissues. Glyceraldehyde-3-phosphate is converted to 1,3-diphosphoglycerate in the presence of GAPDH. In addition to this catalytic conversion, GAPDH also takes part in various other complex biological processes like replication and repair of DNA, export of tRNA from the nucleus, exo- and endocytosis, cytoskeletal organization, etc. GAPDH overexpression is associated with different types of lung cancer, renal cell carcinoma, glioma, breast cancer, hepatocellular
Fig. 5. (A) Heatmap showing hub gene expressions in different samples of GSE112154. (B) Interactions among the common hub genes identified by 10 calculation methods of CytoHubba. (A, B) The downregulated genes are colored red and the upregulated genes are colored green. (B) The nodes are colored in a continuous manner according to their $|\log_{2} FC|$ values. (C) Box plots showing expression levels of different hub genes in MPM (n = 47) and normal mesothelium (n = 3) samples.

https://doi.org/10.5808/gi.21019
carcinoma, pancreatic cancer, gastric and colorectal cancer, melanoma, prostate, and urinary bladder cancer. In some of these cancers, GAPDH overexpression was responsible for chemotherapeutic resistance \[43\]. GAPDH is also overexpressed in malignant pleural mesothelioma \[44\]. From survival analysis of pleural mesothelioma, it can be extrapolated that GAPDH upregulation would also be associated with reduced survival in peritoneal mesothelioma. FGF2 protein has important roles in cellular proliferation, motility, and differentiation. FGF2 gene has a suppressive effect on CDH1. Overexpression of the FGF2 gene in ovarian cancer cells was associated with downregulation of CDH1, upregulation of Slug (SNAI2) and ZEB1, and increased invasiveness \[45\].

Table 3. Top transcription factors and miRNAs

<table>
<thead>
<tr>
<th>Transcription factor</th>
<th>Degree</th>
<th>Betweenness</th>
<th>miRNA</th>
<th>Degree</th>
<th>Betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOX2</td>
<td>293</td>
<td>14,615.15</td>
<td>hsa-mir-124-3p</td>
<td>241</td>
<td>75,109.76</td>
</tr>
<tr>
<td>MYC</td>
<td>283</td>
<td>20,679.36</td>
<td>hsa-mir-16-5p</td>
<td>236</td>
<td>75,393.02</td>
</tr>
<tr>
<td>SUZ12</td>
<td>229</td>
<td>7,521.775</td>
<td>hsa-mir-1-3p</td>
<td>209</td>
<td>57,945.4</td>
</tr>
<tr>
<td>EGR1</td>
<td>226</td>
<td>8,460.82</td>
<td>hsa-mir-27a-3p</td>
<td>195</td>
<td>49,269.11</td>
</tr>
<tr>
<td>STAT3</td>
<td>211</td>
<td>7,224.139</td>
<td>hsa-mir-129-2-3p</td>
<td>184</td>
<td>36,612.51</td>
</tr>
<tr>
<td>HNF4A</td>
<td>211</td>
<td>7,456.627</td>
<td>hsa-mir-34a-5p</td>
<td>170</td>
<td>36,162.74</td>
</tr>
<tr>
<td>NANOG</td>
<td>209</td>
<td>6,106.242</td>
<td>hsa-mir-155-5p</td>
<td>167</td>
<td>39,779.95</td>
</tr>
<tr>
<td>SPI1</td>
<td>208</td>
<td>8,373.637</td>
<td>hsa-mir-146a-5p</td>
<td>166</td>
<td>32,701.75</td>
</tr>
<tr>
<td>TP63</td>
<td>207</td>
<td>7,455.71</td>
<td>hsa-mir-374a-5p</td>
<td>159</td>
<td>30,642.76</td>
</tr>
<tr>
<td>AR</td>
<td>206</td>
<td>8,356.316</td>
<td>hsa-let-7b-5p</td>
<td>155</td>
<td>40,806.08</td>
</tr>
<tr>
<td>PPARG</td>
<td>204</td>
<td>9,683.092</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUNX1</td>
<td>193</td>
<td>6,815.653</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP53</td>
<td>187</td>
<td>7,209.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MITF</td>
<td>185</td>
<td>5,649.56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POU5F1</td>
<td>181</td>
<td>4,757.409</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Kinases acting on the TFs regulating the hub genes

<table>
<thead>
<tr>
<th>Kinase name</th>
<th>Adj. p-value (FDR)</th>
<th>No. of substrate TFs</th>
<th>Substrates (TF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIPK2</td>
<td>0.0003872</td>
<td>3</td>
<td>RUNX1, TP53, STAT3</td>
</tr>
<tr>
<td>MAPK1</td>
<td>0.0146</td>
<td>5</td>
<td>PPARG, TP53, STAT3, MYC, AR</td>
</tr>
<tr>
<td>MAPK14</td>
<td>0.0445</td>
<td>6</td>
<td>PPARG, RUNX1, HNF4A, STAT3, TP53, MYC</td>
</tr>
<tr>
<td>GSK3A</td>
<td>0.0722</td>
<td>2</td>
<td>TP53, MYC</td>
</tr>
<tr>
<td>CSNK2A1</td>
<td>0.0745</td>
<td>4</td>
<td>EGR1, TP53, SPI1, MYC</td>
</tr>
<tr>
<td>MAPK8</td>
<td>0.0745</td>
<td>4</td>
<td>TP53, MYC, PPARG, STAT3</td>
</tr>
<tr>
<td>GSK3B</td>
<td>0.0745</td>
<td>6</td>
<td>MITF, RUNX1, TP53, PPARG, STAT3, MYC</td>
</tr>
<tr>
<td>EPHA3</td>
<td>0.0745</td>
<td>1</td>
<td>STAT3</td>
</tr>
<tr>
<td>DYRK2</td>
<td>0.0745</td>
<td>1</td>
<td>STAT3</td>
</tr>
<tr>
<td>IRAK1</td>
<td>0.0745</td>
<td>1</td>
<td>STAT3</td>
</tr>
<tr>
<td>PRKACB</td>
<td>0.0745</td>
<td>3</td>
<td>EGR1, TP53, MITF</td>
</tr>
<tr>
<td>CDK5</td>
<td>0.0745</td>
<td>2</td>
<td>STAT3, TP53</td>
</tr>
<tr>
<td>PLK3</td>
<td>0.0784</td>
<td>1</td>
<td>TP53</td>
</tr>
<tr>
<td>CDK9</td>
<td>0.0784</td>
<td>1</td>
<td>TP53</td>
</tr>
<tr>
<td>VRK1</td>
<td>0.0806</td>
<td>1</td>
<td>TP53</td>
</tr>
<tr>
<td>FGFR4</td>
<td>0.0806</td>
<td>1</td>
<td>STAT3</td>
</tr>
<tr>
<td>MAPKAPK5</td>
<td>0.0806</td>
<td>1</td>
<td>TP53</td>
</tr>
<tr>
<td>MAPK11</td>
<td>0.0863</td>
<td>1</td>
<td>HNF4A</td>
</tr>
<tr>
<td>MSK1</td>
<td>0.0863</td>
<td>1</td>
<td>STAT3</td>
</tr>
<tr>
<td>FGFR3</td>
<td>0.0909</td>
<td>1</td>
<td>STAT3</td>
</tr>
</tbody>
</table>

TF, transcription factor; FDR, false discovery rate.

miRNA, microRNA,
have found an inverse relationship between CDH1 and FGF2 in MPM. We have found CDH1 is significantly upregulated and FGF2 is downregulated in MPM. We have also found SNAI2 is downregulated and ZEB1 is not differentially expressed in MPM. This suggests a vice versa regulatory relationship may exist between CDH1 and FGF2 in MPM. Increased FGF2 protein level also causes a receptor-independent upregulation of the IL6 gene [46]. This implies a directly proportional relationship between FGF2 and IL6 exists. We have found both FGF2 and IL6 are underexpressed in MPM. MYC is a proto-oncogene and MYC protein has roles in cellular proliferation, differentiation, apoptosis, cellular senescence, DNA damage responses, biosynthesis of ribosome, glycolysis, and mitochondrial functions. It can initiate events that lead to either hyperproliferation of cancer cells or prevention of tumorigenesis. It was further observed that tumor cells having poor blood supply becomes metabolically inactive and MYC level is decreased in these cells. MYC helps survive these cells under hypoglycemic and hypoxic conditions. MYC is usually overexpressed in different cancers. However, it has been reported that MYC is underexpressed in adrenocortical cancers. [47-49]. We have also found it is underexpressed in MPM. PTGS2 (also known as COX-2) is overexpressed in many solid tumors, for example, breast, colorectal, lung, pancreatic, liver, and ovarian cancers [50]. But we have found this gene is downregulated in MPM. TEK is a receptor tyrosine kinase protein. It is involved in angiogenesis through TEK-angiopoietin 1 (ANGPT1) and TEK-angiopoietin 2 (ANGPT2) signaling. ANGPT1 is an agonist of TEK whereas ANGPT2 can play as both an agonist and an antagonist. TEK is downregulated in meta-

![Fig. 6. (A) Differentially expressed gene-transcription factor (TF)-miRNA interaction network. This network was obtained after filtering nodes of the original network with a degree cutoff of 150 to avoid hairball effect. In this image, genes are shown in blue circles, TFs are shown in green circles, and miRNAs are shown in red squares. (B) Network showing kinases that interact with TFs.](https://doi.org/10.5808/gi.21019)

![Fig. 7. Overall survival analyses of the hub genes (p < 0.05). 95% Confidence interval is shown as dotted lines. HR, hazard ratio.](https://doi.org/10.5808/gi.21019)
static clear cell renal cell carcinoma and is associated with poor prognosis [51]. We have found both TEK and ANGPT1 are downregulated in MPM. Zhang et al. (2020) [12] showed that CXCL8/IL8, PPARG, ADIPOQ, and IL6 are upregulated in malignant pleural mesothelioma. But we have found from our analyses that these genes are downregulated in MPM. These genes might have different roles in the pathogenesis of malignant pleural mesothelioma and MPM. We have found VWF and CDHS are downregulated in MPM. These two genes are also downregulated in non-small cell lung cancer [52]. Apart from the hub genes, we have found the most upregulated genes in MPM are KRT19, KRT18P55, MSLN, KRT8, SLPI, CGN, and CXADR (Coxsackie and adenovirus receptor). KRT19 (keratin 19) and KRT8 (keratin 8) are upregulated in lung adenocarcinoma where they are associated with poor prognosis. High KRT19 expression is also associated with liver and breast cancer. High KRT8 expression was also observed in clear cell renal cell carcinoma and gastric cancer [53]. KRT8 was also found to be upregulated in rat models of mesothelioma [54]. KRT18P55 (keratin 18 pseudogene 55) encodes a long intergenic noncoding RNA (KRT18P55), is overexpressed in intestinal-type gastric cancer and correlates with its progression [55]. This implies that this long noncoding RNA has a role common in intestinal-type gastric cancer and MPM. MSLN (mesothelin) in healthy individuals is expressed in pleura, pericardium and peritoneum. However, it is upregulated in all types of mesothelioma, pancreatic adenocarcinoma, ovarian cancers, lung adenocarcinoma, and cholangiocarcinoma [56,57]. SLPI (secretory leukocyte protease inhibitor) up-regulation has been observed in breast, lung, stomach, and colorectal cancers [58]. CXADR is a receptor for Coxsackie B viruses and adenoviruses 2 and 5 [59]. It was shown that CXADR maintains survival and growth of oral squamous cell carcinoma by translocating CDH1 from cytoplasm to cell membrane [60]. We have found that both CXADR and CDH1 are upregulated in MPM and postulate a similar role played by CXADR in MPM.

TFs and miRNAs maintain spatiotemporal gene expression. TFs and miRNAs that can regulate the DEGs were also identified. Among the identified miRNAs, mir-34a was reported to be downregulated in MPM in comparison with normal peritoneum. Re-expression of mir-34a in MPM cells exhibited oncosuppressive events both in vitro and in vivo. This suggests downregulation of this miRNA has a possible role in the pathogenesis of MPM [61]. The roles of other miRNAs are yet to be elucidated.

In this study, adopting a biological network analysis approach, we have identified the potential pathways and genes involved in MPM. These candidate genes and pathways need to be validated in further in vitro and in vivo experiments and in MPM samples to confirm their active roles and to manipulate them for clinical usefulness.

**ORCID**

A. M. U. B. Mahfuz: https://orcid.org/0000-0001-9592-430X

A. M. Zubair-Bin-Mahfuj: https://orcid.org/0000-0003-2237-855X

Dibya Joti Podder: https://orcid.org/0000-0003-1270-6476

**Authors’ Contribution**

Conceptualization: AMUBM. Data curation: AMUBM. Formal analysis: AMUBM. Methodology: AMUBM, AMZBM, DJP. Writing - original draft: AMUBM, AMZBM. Writing - review & editing: AMUBM, DJP.

**Conflicts of Interest**

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

**Supplementary Materials**

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

**References**


8. Xu M, Zeng B, Kuang BH. Identification of hub genes, miRNAs and regulatory factors relevant for Duchenne muscular dys-

https://doi.org/10.5808/gi.21019


https://doi.org/10.5808/gi.21019
Analysis of H3K4me3-ChIP-Seq and RNA-Seq data to understand the putative role of miRNAs and their target genes in breast cancer cell lines

Aneesh Kotipalli, Ruma Banerjee, Sunitha Manjari Kasibhatla, Rajendra Joshi*

HPC-Medical and Bioinformatics Applications Group, Centre for Development of Advanced Computing, Pune 411008, India

Introduction

Breast cancer is one of the leading causes of cancer in women all over the world and accounts for ~25% of newly observed cancers in women. Epigenetic modifications influence differential expression of genes through non-coding RNA and play a crucial role in cancer regulation. In the present study, epigenetic regulation of gene expression by in-silico analysis of histone modifications using chromatin immunoprecipitation sequencing (ChIP-Seq) has been carried out. Histone modification data of H3K4me3 from one normal-like and four breast cancer cell lines were used to predict miRNA expression at the promoter level. Predicted miRNA promoters (based on ChIP-Seq) were used as a probe to identify gene targets. Five triple-negative breast cancer (TNBC)-specific miRNAs (miR153-1, miR4767, miR4487, miR6720, and miR-LET7I) were identified and corresponding 13 gene targets were predicted. Eight miRNA promoter peaks were predicted to be differentially expressed in at least three breast cancer cell lines (miR4512, miR6791, miR330, miR3180-3, miR6080, miR5787, miR6733, and miR3613). A total of 44 gene targets were identified based on the 3′-untranslated regions of downregulated mRNA genes that contain putative binding targets to these eight miRNAs. These include 17 and 15 genes in luminal-A type and TNBC respectively, that have been reported to be associated with breast cancer regulation. Of the remaining 12 genes, seven (A4GALT, C2ORF74, HRCT1, ZC4H2, ZNF512, ZNF655, and ZNF608) show similar relative expression profiles in large patient samples and other breast cancer cell lines thereby giving insight into predicted role of H3K4me3 mediated gene regulation via the miRNA-mRNA axis.

Keywords: breast neoplasms, ChIP-Seq, luminal-A/triple-negative, miRNA, RNA-Seq
moters is known to play a crucial role in breast cancer regulation [4]. Histone modifications wrapped around genes play an important role in gene regulation by providing access to transcription factors, RNA-polymerases, and other regulatory mechanisms [5]. There are several histone post-translational modifications identified earlier with a unique regulatory function for each of them. Using chromatin immunoprecipitation followed by sequencing (ChIP-Seq), one can identify the position of targeted protein binding (transcription factors, histone modifications) regions in the genome [6].

Epigenetic gene regulation occurs in three major ways namely DNA methylation of CpG islands, histone modifications, and non-coding RNA mediated [7]. Each regulatory level has a crucial role in normal cell development and diseases such as cancer and other non-communicable diseases. Gene level histone modifications can be used to predict the status of a gene whether it is active or inactive. H3K4me3 modification at the promoter level provides information about active genes, whereas H3K27me3 and H3K9me3 modifications provide the inactive status of the gene [8]. Previous studies provided evidence for correlation of histone modifications (H3K36me3, H3K9ac, H3K27ac, and H3K4me1) with gene expression. Combinations of two or more histone modifications at gene promoter and gene body level provides more resolution to predict gene activity [8].

Non-coding (nc) RNAs (LncRNAs, long intervening/intergenic noncoding RNA, miRNA, and small interfering RNA) play a major role in gene regulation of different biological processes such as cell cycle and proliferation along with developmental and metabolic processes [9]. The most widely studied ncRNAs are miRNAs which are small ncRNA ~22nt in length, evolutionarily conserved, and have a wide regulatory role in development and diseases. They play an important role in gene regulation by targeting complementary binding sites in untranslated region (UTR) regions of gene transcripts or by targeting promoters or other miRNA or LncRNA [10]. Interestingly, miRNAs are also involved in the upregulation of specific genes via binding to their promoters in the nucleus and thereby controlling gene expression [11]. An actively transcribed miRNA is able to regulate ~100–1,000 genes by complementary binding to their targeted genes. Based on conserved base-pairing homology, it has been predicted that ~60% of human genes are targeted by miRNAs [12]. In many cancers dysregulation of miRNA causes cancer progression and drug resistance. The role of miRNA deregulation in breast cancer was first reported in 2005 wherein the role of miR548 as an oncogenic regulator in breast cancer was elaborated [13]. There are other miRNAs such as let-7, miR145, miR200, and miR497 with a definitive role in breast cancer [14].

The present study is limited to normal-like, luminal-A and TN-BC-claudin subtypes of breast cancer based on hormonal receptor expression. One each of normal subtype, ER positive subtype (Luminal-A) and ER negative (TNBC-claudin) were chosen for analysis. In the current study, ChIP-Seq data corresponding to histone modification H3K4me3 for one normal-like (MCF10A) and four breast cancer cell lines (luminal-A [MCF7, ZR751], TNBC [MB231, MB436]) were chosen to understand role of miRNA-gene promoter regulation of miRNA-mRNA axis. An attempt has been made to map the epigenetic expression patterns (histone H3K4me3) with RNA-sequencing (RNA-Seq) expression of miRNA targeted genes. Analysis of such combined data promises to provide insights into understanding epigenetic gene regulation (chromatin) as well as gene expression [15].

Methods

ChIP-Seq data were downloaded from Gene Expression Omnibus (GEO) for breast cancer pertaining to six cell lines, viz., normal-like (MCF10A and 76NF2V), luminal-A subtype (MCF7 and ZR751), and TNBC subtype (MB231 and MB436), each with one activation histone modification H3K4me3 and two replicates (sequenced as Illumina single-end reads) (Supplementary Tables 1, 2) [16]. RNA-Seq data for the above-mentioned cell lines with four replicates were also downloaded from GEO (Supplementary Table 3) [16].

The raw reads were checked for quality using FastQC (version 0.11.7) [17]. BWA-MEM (Burrows-Wheeler alignment–Maximal Exact Matches) version 0.7.17 was used for alignment with reference genome build hg38 [18]. Samtools (version 1.8) was used to manage replicates and for sam to bam conversion [19]. ChIP-Seq analysis was done using model-based analysis of ChIP-Seq (MACS2, version 2.1.1.20160309) [20]. Peak calling was done using narrowpeak as H3K4me3 generates narrow histone marks [21]. p-value thresholds for peak calling were set to 0.001 for all samples and all replicates (Fig. 1A) [22]. Raw ChIP-Seq data of both H3K4me3 and corresponding input sequence was used for peak calling. To identify the reproducibility within the biological replicates IDR2.0.3 tool was used with a threshold of 0.05 to obtain statistically significant peaks [23,24]. Pseudo replicate analysis was carried out to identify low reproducible replicates which satisfy the criteria of N1/N2 ≥ 2 and Np/Nt ≥ 2 (where N1 represents the number of replicate 1 self-consistent peaks, and N2 represents the number of replicate two self-consistent peaks; Np represents the number of peaks consistent between pooled pseudoreplicates, and, Nt represents the number of peaks consistent between true replicates) [24]. Peaks were annotated using HOMER (v3.12) tool [25]. Peaks corresponding to miRNA promoter genes were extract-
Fig. 1. (A) Chromatin immunoprecipitation sequencing (ChIP-Seq) peak calling workflow for miRNA promoter prediction. (B) ChIP-Seq and RNA sequencing (RNA-Seq) data integration workflow for prediction of miRNA-mRNA interaction via 3′-untranslated region (3′-UTR) binding target prediction.
ChIP-Seq analysis

All the ChIP-Seq datasets passed the quality check (Supplementary Fig. 2) and > 86% of reads were mapped to the reference genome for all replicates of H3K4me3 (for each cell line) used in the study. The number of peaks in the biological replicates varied from 27875 to 64652 for different cell lines (Fig. 2). Reproducibility analysis of peaks (obtained for the replicates) enabled identification of statistically significant peaks (threshold 0.05) (Table 1, Supplementary Figs. 3, 4) which are common between biological replicates [21]. In the normal cell line (MCF10A) and luminal-A subtype (cell lines MCF7 and ZR751), 16,601 peaks (59.7%), 13,008 peaks (53.4%) and 10,158 peaks (48.1%) passed the reproducibility threshold respectively. In the triple-negative subtype (cell lines MB231 and MB436), 14,339 peaks (65.7%) and 16,549 peaks (61.2%) passed the threshold. The highest percentage of overlapping peaks between the replicates was observed in cell line MB231 (65.7%) whereas the least percentage of peaks that passed the threshold was seen in cell line ZR751 (48.1%). All cell lines except 76NF2V cell line generated reproducibility ≥ 2, which indicated that the peaks

![Figure 2](https://doi.org/10.5808/gi.21020)
were reproducible and statistically significant. Hence, cell line 76NF2V was not used for further analysis because of the low reproducibility of replicates, N1/N2 was 3.370 (Supplementary Table 4).

Chromosomal-level distribution of ChIP-peaks is available in Supplementary Fig. 5.

miRNA promoter prediction analysis

miRNA promoter regions were identified for each cell-line (Table 2, Supplementary Table 5). Peaks corresponding to miRNA-gene promoters that are common and unique between normal versus cancerous cell lines were identified (normal vs. TNBC, normal vs. luminal-A, and TNBC vs. luminal-A) (Tables 3, 4, Fig. 3). The majority of the miRNAs predicted have been reported to have a role in breast cancer (Supplementary Table 6).

Cell line-specific miRNAs obtained in this study have been listed in Table 3. Few of these miRNAs have been validated previously [34]. It is interesting to note that there are no common miRNAs between both the luminal-A cell lines used in this study; however, five TNBC-specific miRNAs viz., miR153-1, miR4767, miR4487, miR6720, and miR-LET7I were exclusively found in both the TNBC cell lines. Identification of target genes belonging to TNBC-specific miRNAs was carried out (Supplementary Tables 7, 8). It is to be mentioned that with the cutoff criteria for target-gene identification used in this study (refer to Methods section), no targets were found for miR153-1. Of the five miRNA promoters found to be upregulated in the TNBC cell lines, 3 miRNAs, viz., miR-153-1, miR-6720, and miR-LET7I were found to have similar relative expression in TCGA data samples. Of these, miR-LET7I was found to have higher expression in TNBC.

![Fig. 3. Common and unique miRNAs predicted across different cell lines.](https://doi.org/10.5808/gi.21020)

Eight miRNAs obtained are found to be common across two cancer subtypes (Table 4). miR4512 was observed in all cancerous cell lines, both luminal-A (MCF7 and ZR751) and TNBC (MB231 and MB436) subtypes. miR3180-3 was observed in luminal-A (MCF7 and ZR751) and TNBC subtypes (MB231 and MB436). miR6791 and miR330 were observed to be common in three cancer cell lines, two luminal-A (MCF7 and ZR751) and one TNBC (MB231) subtypes. miR5787, miR6733, and miR3613 were observed in three cancer cell lines, two TNBC (MB231 and MB436) and luminal-A (ZR751) subtypes. miR6080 was observed to be present in three cancer cell lines, two TNBC (MB231 and MB436) and luminal-A (MCF7) subtypes. All the eight miRNAs listed above have been used for further downstream analysis to identify their putative gene targets based on mRNA expression data. Of the eight miRNA promoters found to be upregulated across breast cancer cell lines, the relative expression of three miRNAs, viz., miR-
The relative expression of the other five miRNAs in this resource was found to be insufficient to draw any conclusion.

RNA-Seq analysis

All the RNA-Seq datasets passed the quality check (> 28) and hence were retained for further analyses (Supplementary Fig. 8). About 96% reads mapped for cell lines MCF10A, MCF7, and MB231 whereas, for cell lines ZR751 and MB436 > 93% mapping was observed (Supplementary Table 9). In normal-like vs. luminal-A type, a total of 1,189 genes were upregulated (Supplementary Table 10) and 687 genes were downregulated (Supplementary Table 11). In normal-like vs. TNBC type, a total of 954 genes were upregulated (Supplementary Table 12) and 167 genes were downregulated (Supplementary Tables 13, 14, Supplementary Fig. 9). Five miRNAs specific to the TNBC cell lines were further studied to identify their binding to downregulated 3′-UTR gene targets (Supplementary Tables 8, 15).

TNBC-specific miRNA target analysis of the downregulated genes helped in the identification of 13 genes (Supplementary Fig. 7). The relative expression of the other five miRNAs in this resource was found to be insufficient to draw any conclusion.

Table 3. Predicted cell line specific miRNAs

<table>
<thead>
<tr>
<th>Cell line</th>
<th>No. of unique miRNAs</th>
<th>miRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCF10A - Normal-like</td>
<td>9</td>
<td>miR4790, miR4687, miR4530, miR4692, miR4520-1, miR548AJ1, miR4279, miR1470, miR3675</td>
</tr>
<tr>
<td>MCF7 - Luminal-A</td>
<td>7</td>
<td>miR4734, miR4520-2, miR4521, miR4519, miR4497, miR4477B, miR1244-3</td>
</tr>
<tr>
<td>ZR751 - Luminal-A</td>
<td>11</td>
<td>miR6650, miR4761, miR200C, miR4738, miR1282, miR4781, miR7706, miR4756, miR6090, miR375, miR6515</td>
</tr>
<tr>
<td>MB231 - TNBC</td>
<td>5</td>
<td>miR1260B, miR1258, miR7704, miR574, miR4651</td>
</tr>
<tr>
<td>MB436 - TNBC</td>
<td>12</td>
<td>miR34B, miR1184-3, miR6875, miR6790, miR11401, miR4482, miR6743, miR148A, miR5448, miR4799, miR4466, miR9-3</td>
</tr>
</tbody>
</table>

Table 4. Predicted TNBC and luminal-A specific miRNAs common across (≥3) cancer cell lines

<table>
<thead>
<tr>
<th>Cell line</th>
<th>No. of common miRNAs</th>
<th>miRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MB231 (TNBC)</td>
<td>1</td>
<td>miR4512</td>
</tr>
<tr>
<td>MB436 (TNBC)</td>
<td>2</td>
<td>miR6791, miR330</td>
</tr>
<tr>
<td>MCF7 (Luminal-A)</td>
<td>1</td>
<td>miR3180-3</td>
</tr>
<tr>
<td>ZR751 (Luminal-A)</td>
<td>1</td>
<td>miR6080</td>
</tr>
<tr>
<td>MB231 (TNBC)</td>
<td>3</td>
<td>miR5787, miR6733, miR3613</td>
</tr>
</tbody>
</table>

TNBC, triple-negative breast cancer.

330, miR-3613, and miR-6733 were found to be complementary in studies reported in TCGA data samples using UALCAN webserver (Supplementary Fig. 7). The relative expression of the other five miRNAs in this resource was found to be insufficient to draw any conclusion.

Kotipalli A et al. • ChIP-Seq analysis of breast cancer and miRNA role
Genes STC2, CPA4, and NUPR1 were found to have a relatively higher expression amongst the 13 target genes of TNBC cell lines. In larger breast cancer samples obtained from TCGA, with the exception of SPOCK2, CPA4, C1orf228, and NFE2, the other target genes are found to have relatively low expression in TNBC as compared to normal samples (Supplementary Fig. 11). Relative expression of these genes in other cancer subtypes hints at the down-regulatory effect of TNBC-specific miRNAs. Survival plots of most of the downregulated genes (with the exception of NUPR1, CPA4, EPHA3, ADAMTSL1, and ATP13A4) were found to be associated with poor patient survival (Supplementary Fig. 12).

Eight miRNA promoters that are common across more than three cancer cell lines were also used as probes to identify the gene targets (Table 4, Supplementary Tables 16-18). A total of 44 downregulated gene targets were identified across luminal-A and TNBC subtypes. In normal-like (MCF10A) vs. luminal-A (MCF7 and ZR751) downregulated genes, 17 genes have been predicted and their role in breast cancer has been reported earlier (rer, Ig-FBP6, Spata18, Ax1, Bmf, Fxyd5, Ptf1, Runx2, Ugt8, Csf3, Heg1, Plau, Pter, S100a3, Snurf, and Wipf1) [35-51] (Fig. 4, Supplementary Tables 6, 19, Supplementary Fig. 13). In normal-like (MCF10A) vs. TNBC (MB231 and M436) downregulated genes, 15 genes have been predicted in this study and their role in breast cancer have also been previously reported (Tnfsf10, Temem47, Igqap2, Fat4, Nupr1, Hoxc13, Prrx1, STC2, Ac108941.2, Adamtsl1, Arhgef5, Bnc1, Cp4, Ppl, and Tnfrsf10d) [52-66] (Fig. 5, Supplementary Tables 5, 6, 20, Supplementary Fig. 14).

Of the remaining 12 target genes identified, nine genes in luminal-A were identified to be regulated by their corresponding miRNAs (gene A4GALT targeted by miR3180-3, miR5787, and miR6791; gene C10orf55 targeted by miR330, miR3180-3, miR5787, and miR6791; gene C2orf74 targeted by miR330 and miR5787; gene Zc4h2 targeted by miR330 and miR5787; gene Znf512 targeted by miR330, miR3180-3, miR5787, and miR6791; gene Znf655 targeted by miR5787; gene Znf71 targeted by miR5787 and miR6791; gene Hcg2042738 targeted by miR6791; gene Hrct1 targeted by miR4512 and miR5787) (Table 5). Similarly, three genes in TNBC were also identified to be regulated by their corresponding miRNAs (gene Hist3h2a targeted by miR5789; Znf608 targeted by miR5787; Elov14 targeted by miR5787) (Supplementary Table 18). Comparison of these 12 target genes to other breast cancer cell lines from the CCLE database revealed that all of them have low expression as compared to the Actb control gene (Supplementary Fig. 15). Genes Hist3h2a and C2orf74 were found to have a relatively higher expression amongst the 12 target genes. In larger datasets of breast cancer, with the exception of Znf71 and Hist3h2a, all other gene targets were found to be downregulated (Supplementary Figs. 13, 14, 16–18). This observation supports the probable role of miRNA-mRNA axis in gene regulation. The down-regulation of A4GALT, C2orf74, Hrct1, Zc4h2, Znf512, Znf655, Znf608, and Hist3h2a genes were found to be independently associated with poor survival in breast cancer patients (Table 5, Supplementary Fig. 19). It

<table>
<thead>
<tr>
<th>Table 5. Predicted gene targets of differentially regulated miRNAs in breast cancer cell lines (TNBC and luminal-A) proposed using ChIP-Seq–RNA-Seq integrated analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Luminal-A</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>TNBC</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

TNBC, triple-negative breast cancer; ChIP-Seq, chromatin immunoprecipitation sequencing; RNA-Seq, RNA-sequencing; TCGA, The Cancer Genome Atlas; CCLE, Broad Institute Cancer Cell Line Encyclopedia.
Fig. 4. Relative gene expression (The Cancer Genome Atlas [TCGA] breast cancer samples) of luminal-A downregulated gene targets (12 of the total 17 genes) previously reported in breast cancer that correlate with predicted miRNA binding analysis: (A) PTER, (B) HEG1, (C) SPATA18, (D) PTRF, (E) SNURF, (F) RERG, (G) AXL, (H) FXYD5, (I) WIPF1, (J) CSF3, (K) UGT8, and (L) IGFBP6.
Fig. 5. Relative gene expression (TCGA breast cancer samples) of triple-negative breast cancer downregulated gene targets (12 of the total 15 genes) previously reported in breast cancer that correlate with predicted miRNA binding analysis: (A) PPL, (B) ADAMTS11, (C) TMEM47, (D) TNFSF10, (E) FAT4, (F) TNFRSF10D, (G) ARHGEF5, (H) BNC1, (I) PRRX1, (J) NUPR1, (K) STC2, and (L) IQGAP2.
needs to be mentioned that relative expression data and survival plots for gene HCG2042738 could not be obtained due to insufficient annotation.

Discussion

The interplay between epigenetic gene regulation through histone modifications and other regulatory mechanisms like ncRNA is of great interest in cancer biology. In the present analysis, the role of H3K4me3 in miRNA expression based on promoter level peaks has been studied using ChIP-Seq and RNA-seq data integration. To achieve the same, a novel approach of mapping data derived from ChIP-Seq (miRNA promoter peaks) and RNA-Seq (targets of 3′-UTRs of genes binding to miRNA) was used to understand epigenetic regulation that may aid in the identification of subtype and cell line specific miRNAs [15,16].

In normal-like cell line MCF10A, of the nine unique miRNAs identified, miR4530 was found to have a role in the suppression of cell proliferation, promote angiogenesis and induce apoptosis by targeting gene VASH1 (Vasohibin 1) in breast carcinoma [67]. Hence, promoter-level epigenetic regulation of miR4530 by H3K4me3 may have a protective role in normal-like subtypes. miR34B was observed to be present only in cell line MB436 (TNBC subtype). miR34B has high expression in TNBC tumors compared to normal types. Expression of miR34B highly correlates with clinical outcome of patients. Notch2 (notch receptor 2) gene that has a role controlling cell differentiation, is a direct target for miR34B [68]. miR6875 was observed in TNBC cell line MB436. According to previous reports, a high expression of miR6875 was observed in early breast cancer patients [69]. miR574-5p attenuates proliferation, migration, and epithelial mesenchymal transition (EMT) in TNBC cells by targeting genes BCL11A (BAF chromatin remodeling complex subunit) and SOX2 (SRY-Box transcription factor 2) to inhibit the SKIL (SKI like proto-oncogene)/TAZ (Tafazzin)/CTGF (connective tissue growth factor) axis [70].

Of the five TNBC subtype-specific miRNAs, miR153, miR6720, and miR-LET7I were found to be upregulated in larger breast cancer datasets belonging to TCGA. miR153 has been reported to have a tumor suppressor role and has been suggested as a prognostic marker for TNBC [34].

The majority of the predicted gene targets (total 44) overlap with previous experimental studies and include 32 gene targets (Figs. 4, 5) of eight miRNAs (miR4512, miR6791, miR330, miR3180-3, miR6080, miR5787, miR6733, and miR3613) which are identified in more than three breast cancer cell lines and absent in normal-like cell lines. Overexpression of miR330-3p in breast cancer cell lines has been reported earlier, which results in greater invasiveness in vitro, and miR330-3p–overexpressing cells also metastasize more aggressively ex-ovo [71]. Gene CCBE1 (collagen and calcium binding EGF domains 1) is a direct target of miR330-3p, and knockout of CCBE1 results in a greater invasive capacity [71]. Exosomal expression of miR3613-3p promotes breast cancer cell proliferation and metastasis. It has been previously reported that miR3613-3p levels were negatively correlated to SOCS2 (suppressor of cytokine signaling 2) expression in breast cancer tissues [72]. Few genes were observed to be targeted by multiple miRNAs (like A4GALT and FOXL2 targeted by three miRNAs each) as it is known that miRNAs can regulate multiple targets based on seed match and sequence similarity between miRNA-mRNA [10].

Of the remaining 12 gene targets, relative gene expression of genes A4GALT, C20orf74, HRCT1, ZC4H2, ZNFS12, ZNF65S, and ZNF608 agree with the proposed hypothesis of H3K4me3 regulated miRNA-mRNA axis in large patient data (TCGA samples) along with their relative expression in other breast cancer cell lines (CCLE database). These genes were associated with poor survival based on KM plots (Human Protein Atlas). The proposed methodology of miRNA-mRNA regulation when analyzed in the context of other histone modifications like H3K27me3, H3K4me1, H3K9me3 will enable better insights into the underlying mechanism of breast cancer regulation.

Acknowledgments

The authors thank the BRAF facility of C-DAC for HPC infrastructure. The authors would also like to deeply thank Dr. Janaki C.H. for her scientific review and critical comments. This research was supported by the Ministry of Electronics and Information Technology, Government of India under the Grant No. 286/2018-19/EMR-II. The authors also thank the BRAF facility of C-DAC for HPC infrastructure. The authors would also like to deeply thank Dr. Janaki C.H. for her scientific review and critical comments. This research was supported by the Ministry of Electronics and Information Technology, Government of India under the Grant No. 286/2018-19/EMR-II.
work is funded by the National Supercomputing Mission (NSM) of the Government of India. The authors thank anonymous reviewers for their valuable suggestions and constructive criticism in improving the manuscript.

Supplementary Materials

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

References


Implications of the simple chemical structure of the odorant molecules interacting with the olfactory receptor 1A1

S. June Oh*

Department of Pharmacology, Inje University College of Medicine, Busan 47392, Korea

G protein–coupled receptors (GPCRs), including olfactory receptors, account for the largest group of genes in the human genome and occupy a very important position in signaling systems. Although olfactory receptors, which belong to the broader category of GPCRs, play an important role in monitoring the organism’s surroundings, their actual three-dimensional structure has not yet been determined. Therefore, the specific details of the molecular interactions between the receptor and the ligand remain unclear. In this report, the interactions between human olfactory receptor 1A1 and its odorant molecules were simulated using computational methods, and we explored how the chemically simple odorant molecules activate the olfactory receptor.

Keywords: electron transfer, homology modeling, molecular docking, olfactory receptor

Introduction

It is very important for animals to be able to perceive their surroundings accurately, and this ability is directly linked to survival. Sight and smell are among the senses involved in the initial recognition of external signals by an animal’s biological system, and G protein–coupled receptors (GPCRs) are responsible for these pivotal functions.

Since the first GPCR structure with a good resolution (2.8 Å) was reported in 2000 [1], more than 400 GPCR entries produced by methods such as X-ray crystallography or cryogenic electron microscopy have been registered in the RCSB Protein Data Bank (PDB) [2]. Many of the three-dimensional (3D) structures of these seven-transmembrane receptors (7TMRs), including visual receptors, have been revealed through the methods described above; however, the 3D structure of olfactory receptors (ORs) has not yet been reported. Accordingly, the actual binding mode of ORs and their ligands has not been known. This gap in scientific knowledge has hampered research on the actual activation mechanism of ORs by odorant molecules.

It is known that ORs are widely expressed in the olfactory organs, where they carry out their main functions; however, recent reports have suggested that ORs also appear to be active throughout the animal body [3,4]. In addition to discovering the important functions and roles of ORs, it is essential to determine and understand the actual activation mechanisms of functional proteins, including 7TMRs. Signal transduction in 7TMR systems is known to be mediated by structural changes in receptors triggered by ligands. Therefore, it is of the utmost importance to investigate the initial interactions between ligands and receptors.
The simpler the chemical structure of a ligand, the more straightforward it is to characterize the binding and action relationship between the membrane receptor and the ligand through their interactions. In recent research, human olfactory receptor 1A1 (OR1A1) activation was observed in 11 different tumor types by the analysis of single-cell transcriptomes [5]. The molecular weight of the ligands binding to OR1A1 was around 140 Da, which is relatively small, and the ligands have simple structures as biomolecules. The present study was designed to establish a homology model of OR1A1 and to determine the functional group involved in the interaction between the model and the ligand from the binding mode by a computational simulation.

Methods

Dataset for OR1A1 ligands
In order to use experimentally validated data for ligands of OR1A1, 3D structural data for a total of 106 chemical compounds (Supplementary Table 1)—53 agonists and 53 non-agonists—were applied to the analyses [6]. Three-dimensional files in the structure-data file (SDF) format for the 106 compounds were downloaded from the PubChem Compound Database at the National Center for Biotechnology Information (NCBI) and subjected to geometric optimization, molecular vibrational pattern analysis, docking simulation, and further study.

Molecular vibration calculation and data formulation of OR1A1 ligands
To explore the site of the ligand that contacts the OR1A1 model, we adopted the revised corralled intensity of molecular vibrational frequency (CIMVF) of ligands as the molecular descriptor. Most of the analysis process followed previous reports [7,8], and a brief description of the procedure is as follows.

We utilized the CIMVF as the characteristic of each ligand molecule. Since the calculation of molecular vibrational frequencies requires a 3D structure of a given molecule, the geometric optimization of the ligand was carried out using the SDF file of each molecule. The theoretical 3D conformer SDF of each ligand molecule was modeled as a single low-energy conformation by using the Becke three-parameter Lee-Yang-Parr (B3LYP) density functional theory and the standard split-valence basis set 6-31G(d,p). Then, the result of geometric optimization was subjected to vibrational frequency calculation. The calculations of the geometric optimization and normal modes of molecular vibration were performed using the GAMESS program package [9]. When it was necessary to check the wavenumber of a substructure in a ligand molecule, we utilized MacMolPlt [10].

The wavenumbers of calculated molecular vibrations in a ligand molecule were sorted in increasing order and taken into each corral with a fixed step size (5 cm⁻¹). As a molecular descriptor of a ligand, the intensity sum of each corral was arrayed in a one-dimensional vector containing 800 elements representing the wavenumber range of 0–4,000 cm⁻¹ [8].

Feature selection by information gain
One of the challenges encountered when dealing with high-dimensional and sparse datasets such as CIMVFs of small molecules is that the number of important or informative features that must be grasped in order to understand the underlying mechanism of a particular phenomenon is very small. Feature selection using information gain (IG) is a process for reducing meaningless or less informative features.

Scoring with IG involves separately counting the occurrences of a feature in the agonist and non-agonist training examples, and then computing an equivalent function. The IG yielded from a dataset is given by the relative entropy between the prior and posterior probabilities [11]. When the information available is the presence of a feature and the corresponding class distribution, IG measures the amount of information about the class prediction in bits [12].

We adopted IG-based feature selection to identify the corral of molecular vibrational frequency as the most informative features among the 800 elements for the classification of OR1A1 ligands into two types (agonists and non-agonists). Since most of the OR1A1 ligands have a small molecular weight, the number of corals with molecular vibration is not large, so the density of the features will be quite sparse. We trained and tested the procedure by applying leave-one-out cross-validation to each of 106 ligands using the Weka machine learning package [13].

Homology modeling of OR1A1
Since molecular docking requires a homology model of the target protein, it must be preceded by homology modeling of the corresponding protein. The amino acid sequence for OR1A1 (UniProt: Q9P1Q5) was obtained from UniProt KnowledgeBase (https://www.uniprot.org/). To choose the amino acid sequences of 7TMRs that have a higher BLAST score (bits) than 45 or a lower E-value than 1e-07, we analyzed the amino acid sequence of OR1A1 against the locally-built BLAST database of 7TMR amino acid sequences registered in the RCSB PDB. Based on their similarity results for OR1A1, four PDB entries of four 7TMRs were selected as the experimental templates for the homology modeling of OR1A1: human β2-adrenergic receptor (ADRA2A), human adenosine receptor (AdoRA2A), bovine rhodopsin (Rho), and turkey β1-adrenoceptor (ADRB1) PDB models.

Oh SJ • Simple chemical structure of OR1A1 odorants

Genomics & Informatics

https://doi.org/10.5808/gi.21033
In order to apply the activated structure of the receptor as the template, we adopted the PDB model of each receptor binding to an agonist or the corresponding G protein: 7B22 (PDB code for ADRA2A), SG53 (AdoRA2A), STE3 (Rho), and 6H7J (ADRB1). The multiple sequence alignment (MSA) of these four PDB models was executed using T-Coffee (Rel. 11) in “slower and more accurate” mode [14]. We applied the MSA result to MODELLER (Rel. 9.25 [15]) and the application automatically combined these four templates to build the model for OR1A1 using information from multiple templates to build the three-dimensional (3D) structural model of OR1A1. After confirming the 3D model of OR1A1 obtained here with a Ramachandran plot, it was used for the subsequent docking experiment.

In addition to confirmation of the 3D model of OR1A1, Phobius was used to determine the terminal regions of the transmembrane α-helices of the OR [16].

**Molecular docking and scoring**

In recent years, molecular docking has frequently been used as a practical computational methodology to predict the binding structure between a ligand and a receptor. There are several freely available programs for molecular docking analysis, such as smina [17]. Smina was created as a fork of AutoDock Vina optimized to support high-throughput and user-specified scoring. With reference to a report that smina has relatively high performance and is convenient to handle relative to several freely available docking programs, such as AutoDock4, AutoDock Vina, and idock [18], the subsequent docking and scoring experiments were performed using the smina program.

In order to simulate the binding of a ligand to a protein in a molecular docking tool, it is necessary to designate a spatial region within the protein to which the ligand can bind. We used AutoDockTools4 (ADT4), which accompanies AutoDock4, to prepare practical conditions for the 3D docking space in the modeled OR1A1 [19]. The binding site grid box was visually defined for the model of OR1A1 by employing the grid setting feature of ADT4.

To explore the structural conformation of each receptor-ligand set, smina was executed using the default parameters with the exception of the 3D coordinates of the search space so that the program outputs nine docking poses for each run. The 3D SDF files of each ligand were downloaded from the PubChem of NCBI since smina receives SDF files as input for the docking experiment with the corresponding receptor. The subsequent processes were carried out under the default conditions of smina.

### Results

**Molecular vibrational patterns of ligands and IG ranking**

According to the results of the molecular vibrational frequency, 372 corrals of agonists and 366 corrals of non-agonists had vibrational intensities of 0, out of the 800 ones of CIMVF. The remaining corrals were regarded as features containing partial characteristics of the ligand with molecular vibration patterns. In the IG score calculations, only 11 out of over 400 features showed IG scores exceeding 0. Among them, nine wavenumbers that are physicochemically meaningful are shown in Table 1. Only informative features with IG scores larger than 0 are listed in the table.

To view the distribution and intensity of molecular vibrational frequencies as a whole, we also plotted the mean vibrational intensities of OR1A1 agonists and non-agonists according to their molecular vibrational frequency (Fig. 1).

There were three major areas of features with mean intensities greater than the mean value: around 1,200, 1,800, and 3,100 cm⁻¹. Among them, the first region around 1,200 cm⁻¹ seemed to be marked in a way that could distinguish agonists and non-agonists. This can also be confirmed in the features from the IG ranking data: F241 and F243 (Table 1). F230, the feature with the highest frequency, has a wavenumber only 50 cm⁻¹ away from them.

The wavenumber 1,204.48 cm⁻¹, belonging to F241, corresponds to the molecular vibrational frequency of the cyclohexanone ring in (+)-dihydrocarvone, and F230, which includes a wavenumber of 1,146.86 cm⁻¹, is a representative feature of the molecular vibration of quinoline. Since the structure of the odorant molecules used in the analysis is relatively simple, a limitation is that the distinction of features corresponding to each substructure does not seem to have much meaning. However, molecular vibrational information of a specific substructure of a ligand is useful for identifying the sites where the receptor interacts. The relationship between agonism on

<table>
<thead>
<tr>
<th>IG feature No.</th>
<th>Wave number (cm⁻¹)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>F601</td>
<td>3,000–3,005</td>
<td>106</td>
</tr>
<tr>
<td>F286</td>
<td>1,425–1,430</td>
<td>106</td>
</tr>
<tr>
<td>F230</td>
<td>1,145–1,430</td>
<td>106</td>
</tr>
<tr>
<td>F108</td>
<td>535–540</td>
<td>106</td>
</tr>
<tr>
<td>F764</td>
<td>3,815–3,820</td>
<td>100</td>
</tr>
<tr>
<td>F241</td>
<td>1,200–1,205</td>
<td>99</td>
</tr>
<tr>
<td>F243</td>
<td>1,210–1,215</td>
<td>95</td>
</tr>
<tr>
<td>F368</td>
<td>1,835–1,840</td>
<td>48</td>
</tr>
<tr>
<td>F347</td>
<td>1,730–1,735</td>
<td>48</td>
</tr>
</tbody>
</table>

Table 1. The features with high frequencies and scores in the information gain (IG) analysis
the receptors and molecular vibrations of ligands needs further research.

**Homology model of OR1A1**

The homology model of human OR1A1 generated by MODELLER with 4 PDB 7TMR models is shown in Fig. 2. The most conserved structure in ORs is the seven-helix bundle of transmembrane α-helical amino acids. In the figure, the color of each helix denotes the direction from the N-terminus (blue) to the C-terminus (red). In Fig. 2B and D, we can better recognize the structure of the receptor’s binding pocket, which are formed by helices 3, 5, 6, and 7 with the schematic representation of the OR1A1 model.

To compare the amino acid sequences in the terminal regions of the seven-transmembrane α-helices generated in the OR1A1 model to those produced by the sequence prediction tool, we used Phobius (Fig. 3).

As shown in Fig. 3, the transmembrane region of the homology model is aligned properly to that predicted using Phobius, supporting the validity of the 3D model of OR1A1.

**Docking dimensions between OR1A1 receptor model and ligands**

In order to determine which site of each ligand binding to the previously generated OR1A1 model interacts with which site of the receptor, the distance between each ligand and the amino acid residue of the receptor located closest to it was determined. As shown in Table 2, the distance between the element group constituting the backbone of the ligand and the functional group of the nearest receptor amino acid residue was observed to be around 4 Å.

![Fig. 1. The mean intensities of CIMVF of the two ligand groups (agonists and non-agonists) for OR1A1. CIMVF, corralled intensity of molecular vibrational frequency; OR1A1, olfactory receptor 1A1.](image)

![Fig. 4](image)

**Discussion**

In animal systems, ORs are the first proteins to detect signals from outside, and the mechanism of OR activation by ligands has not yet been elucidated in detail. It is known that odorant molecules bind to specific receptors through conventional molecular interactions, causing a conformational change in the receptor that initiates intracellular signals. However, this hypothesis was unable to distinguish or predict the properties of odorant molecules because a significant number of odorant molecules bind to a single OR. Therefore, it was not possible to design odorant molecules in consideration of the receptor’s characteristics [20].

If the ligand of a 7TMR has a large and complex structure, such as angiotensin with a molecular weight of 1,000 Da or more, it is very difficult to predict the mechanism of binding between the receptor and the ligand. In contrast, if the ligand has a small molecular weight and a chemically simple structure like the ligand of the OR1A1 receptor, it would be possible to track the process by which the ligand acts, binds to, and activates the receptor.

When electron tunneling was first described, it was exclusively the purview of physicists [21], but it is now also a very important
theme in chemistry and biology. Recently, efforts have been made to interpret similar bitter almond odors between hydrogen cyanide, benzaldehyde, and nitrobenzene using the emission spectra of electron tunneling [22]. Studies of electron transfer in proteins such as electron hopping in polypeptides, electron transfer in peptides such as amino acid relays, and light harvesting systems in photosynthesis have been reported several times [23]. Quantum coherence must also be considered in relation to the Fenna-Matthews-Olson light-harvesting complex [24].

It is known that oxidoreductase proteins mediate tunneling of electrons at rates far faster than the substrate redox reactions they support. Electrons can travel up to 14 Å between redox centers through the protein medium [25]. Electron tunneling for a distance longer than 14 Å is possible through interventions such as

Fig. 2. The homology model of human OR1A1 generated by MODELLER: (A) solid ribbon (side-view), (B) schematic (side-view), (C) solid ribbon (top-view), and (D) schematic (top-view). The colored transmembrane domains are shown in blue to red from the N-terminus to the C-terminus. The side-view and top-view stand for the view from the cell membrane parallel and outside the cell membrane, respectively. OR1A1, olfactory receptor 1A1.
It has been reported that (+)-dihydrocarvone, with a molecular weight of 152.23 Da, is an agonist of OR1A1. The calculated distance between the oxygen atom of (+)-dihydrocarvone and the nearest oxygen of the Tyr251 residue of OR1A1 model was 3.260 Å in the docking simulation model. Therefore, the distance between the OR and the ligand is not an obstacle to electron transfer. Even if the odorant molecule interacts with another amino acid residue of other structures in OR1A1, the context of the above point is not expected to change significantly. This is because, as shown in Fig. 5, most of the ligands of OR1A1 do not have cofactors. The average distance between the molecular backbones of six ligands (Fig. 4) and the nearest amino acid residue of OR1A1 was 3.660 Å (Table 2). It has been reported that (+)-dihydrocarvone, with a molecular weight of 152.23 Da, is an agonist of OR1A1. The calculated distance between the oxygen atom of (+)-dihydrocarvone and the nearest oxygen of the Tyr251 residue of OR1A1 model was 3.260 Å in the docking simulation model (Fig. 4). Therefore, the distance between the OR and the ligand is not an obstacle to electron transfer. Even if the odorant molecule interacts with another amino acid residue of other structures in OR1A1, the context of the above point is not expected to change significantly. This is because, as shown in Fig. 5, most of the ligands of OR1A1 do not have cofactors.

### Table 2. The shortest distance between each ligand and the nearest amino acid residue of the OR1A1 model

<table>
<thead>
<tr>
<th>Name</th>
<th>CAS No.</th>
<th>CID</th>
<th>Distance (Å)</th>
<th>B.A.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(+)-Carvone</td>
<td>2244-16-8</td>
<td>16724</td>
<td>4.132</td>
<td>-6.4</td>
</tr>
<tr>
<td>Allyl phenyl acetate</td>
<td>1797-74-6</td>
<td>15717</td>
<td>3.574</td>
<td>-6.2</td>
</tr>
<tr>
<td>Citral</td>
<td>5392-40-5</td>
<td>638011</td>
<td>3.468</td>
<td>-6.0</td>
</tr>
<tr>
<td>D-Limonene</td>
<td>5989-27-5</td>
<td>440917</td>
<td>3.489</td>
<td>-6.0</td>
</tr>
<tr>
<td>Helional</td>
<td>1205-17-0</td>
<td>64805</td>
<td>3.542</td>
<td>-6.5</td>
</tr>
<tr>
<td>Quinoline</td>
<td>91-22-5</td>
<td>7047</td>
<td>3.755</td>
<td>-6.1</td>
</tr>
</tbody>
</table>

OR1A1, olfactory receptor 1A1; B.A., binding affinity.
Fig. 5. The chemical structures of seven agonists of OR1A1 [27,28]: (A) (+)-carvone, (B) allyl phenyl acetate, (C) citral, (D) D-limonene, (E) helional, (F) quinoline, and (G) (+)-dihydrocarvone. OR1A1, olfactory receptor 1A1.

a reactive group that would be capable of sending and receiving protons. In fact, odorant molecules do not change chemically when they bind to ORs.

The fact that such a chemically very simple ligand changes the structure of the receptor makes it undeniable that there is something very minute between them. It is thought that it would be impossible for odorant molecules to exert a strong enough influence to change the structure of a relatively large OR without any kind of physicochemical transportation. For example, if electron flow occurs, this phenomenon is likely to affect a salt bridge, such as an ionic lock [29], eventually changing the structure of the receptor. In this case, the conformational shift of TM6, a member of the ionic lock, reaches about 5 Å [30]. As an explanation for this, vibration-assisted tunneling can be proposed [31]. Once the electrons of the ligand are transferred in any way to the amino acid residues of the receptor, then there arises a possibility of affecting the ionic lock, either through single-step tunneling or a multi-step hopping process.

Research on the activation mechanism of ORs has been steadily progressing, but no model can fully explain this phenomenon. In particular, regarding the specific process through which receptors are activated by odorant molecules, whether the chemical structure of the molecule is the main factor or whether activation is due to the vibration of the molecule remains unclear. However, this inconsistency can be viewed as involving phenomena that are complementary to each other rather than mutually inconsistent possibilities. This is because the phenomenon through which the receptor recognizes and binds the ligand and the phenomenon through which the receptor is activated by the bound ligand may be different processes. More scrupulous and detailed follow-up studies are needed to clarify this issue.

**ORCID**

S. June Oh: https://orcid.org/0000-0001-7607-870X

**Conflicts of Interest**

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

**Supplementary Materials**

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

**References**

2. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, et al. RCSB Protein Data Bank: powerful new tools for exploring...


https://doi.org/10.5808/gi.21033
Plant height is an important component of plant architecture and significantly affects crop breeding practices and yield. We studied DNA variations derived from F5 recombinant inbred lines (RILs) with 96.8% homozygous genotypes. Here, we report DNA variations between the normal and dwarf members of four lines harvested from a single seed parent in an F6 RIL population derived from a cross between Glycine max var. Peking and Glycine soja IT182936. Whole genome sequencing was carried out, and the DNA variations in the whole genome were compared between the normal and dwarf samples. We found a large number of DNA variations in both the dwarf and semi-dwarf lines, with one single nucleotide polymorphism (SNP) per at least 3.68 kb in the dwarf lines and 1 SNP per 11.13 kb of the whole genome. This value is 2.18 times higher than the expected DNA variation in the F6 population. A total of 186 SNPs and 241 SNPs were discovered in the coding regions of the dwarf lines 1282 and 1303, respectively, and we discovered 33 homogeneous nonsynonymous SNPs that occurred at the same loci in each set of dwarf and normal soybean. Of them, five SNPs were in the same positions between lines 1282 and 1303. Our results provide important information for improving our understanding of the genetics of soybean plant height and crop breeding. These polymorphisms could be useful genetic resources for plant breeders, geneticists, and biologists for future molecular biology and breeding projects.

Keywords: dwarf, RIL population, SNP, soybean, whole genome sequencing, wild type

Introduction

Soybean is one of the most important leguminous crops worldwide due to its use in human food and oil production. Currently, the United States, Brazil, and Argentina account for more than 80% of the worldwide production of soybean [1]. In Southeast Asian countries, particularly Korea, China, and Japan, soybean is used in multiple life stages as a rich source of protein, and it is considered one of the five major grains [2]. Plant height is an important trait that has a direct impact on yield and lodging resistance. Extremely tall
plants can be affected by lodging, which may reduce yield and quality [3]. Dwarfism in crops has played a major role in the “Green Revolution,” in which semi-dwarf varieties were chosen for further cultivation, first in wheat and then in rice [4]. Many studies of plant height inheritance have successfully cloned dwarf genes [5-7]. In soybean, some high-yielding and lodging-resistant dwarf varieties have been developed [8].

The rapid development of next-generation sequencing (NGS) technology and instruments has supported quick and efficient genomics research [9,10]. The key advantage of NGS is that it can produce a large amount of data at low cost, and it is currently being applied to a number of plants [11-14]. Single nucleotide polymorphisms (SNPs) are ubiquitous in genomes and have emerged as a marker of choice, especially in sequenced plants [15,16]. Plants adapt to different environments by various mechanisms, one of which is allelic variation. The identification of these variations is the first step in the in-depth study of the genes and alleles involved in plant evolution and environmental adaptation.

A previously reported study, where the wild soybean Glycine soja was sequenced and compared to the Glycine max reference, found 2.5 megabases (Mb) of substituted sequences, 4.6 kilobases (kb) of indels, 32.4 Mb of deletions and 8.3 Mb of new sequences in a total of 915.5 Mb of genome sequence [14]. Although a great deal of information is available from whole genome sequencing, resequencing strategies have become an important tool to study allelic variations. There have been studies in other plants, such as rice [17], maize [18], Arabidopsis [13], and sorghum [19], as well as resequencing studies in soybean, where both wild and commercial varieties have been analyzed [20-22]. In this study, we compared two inbred lines obtained from a cross between G. max var. Peking and G. soja IT182936 in the F6 generation. A few lines segregated for a dwarf phenotype and continued to show the same phenotype in the next generation. Resequencing analysis revealed many SNPs and indels in both genic and non-genic regions, which are explained in this study.

Methods

Plant materials
Recombinant inbred lines (RILs) were developed from a cross between G. max var. Peking and G. soja IT182936. The soybeans used in this experiment were harvested in Chuncheon city (Gangwon-do, South Korea). All the plants were grown in field condition. Two RILs exhibiting normal and dwarf phenotype from F6 generation were selected for whole genome variant analysis and designated 1282NF6 and 1303NF6 for normal plants and 1282DF6 and 1303DF6 for dwarf plants. Two more RILs, 1214 and 1290, exhibiting semi-dwarf phenotypes, were also selected for comparison with the dwarf lines. Three leaves were collected from each plant before flowering, frozen immediately in liquid nitrogen and stored at −80°C.

DNA isolation and Illumina sequencing
Genomic DNA was isolated from the leaf tissues using the modified CTAB method [23]. DNA purification was carried out using QIAquick Purification Kit (28104, Qiagen, Beijing, China). Adaptor ligation and DNA clustering preparations were done by Solexa sequencing using Illumina HiSeq 4000 sequencing platform according to the manufacturers’ protocol by the National Instrumentation Center for Environmental Management (NICEM) at Seoul National University. The sequencing libraries were prepared by random fragmentation of the DNA sample, followed by 5’ and 3’ adaptor ligation. Low-quality reads (ratio of reads that have phred quality score of < 20), reads with adaptor sequences, and duplicate reads were eliminated. The remaining high-quality data were used for mapping. We used the published genome sequence of G. max version 1 as a reference [24].

Identification of DNA variations in normal and dwarf lines
The raw Illumina sequencing data were filtered and compared to characterize the genotype of normal and dwarf samples using the Bowtie2 (v2.3.4.3) aligner [25]. The SNPs were qualified by GATK (version 2.3.9 Lite) [26] and biallelic filtering. GATK filtering was performed with the options MQ0 ≥ 4 && ((MQ0/(1.0*DP)) > 0.1), QUAL < 30, QD < 5.0, and FS > 200.0 [27]. The biological function of each SNP and indel locus were identified using Snpeff software [28]. Paired-end reads were mapped against the TAIR10 reference genome sequence [29]. The DNA variations common to the two sets of dwarf lines were discovered by comparing the SNPs discovered at the sample loci between the normal and dwarf samples.

Results

In silico mapping of resequencing reads to reference and variant calls
The parental lines (G. max var. Peking and G. soja IT182936) did not exhibit dwarfism, but few plants in F3 generation appeared dwarf (Fig. 1). Although F3 dwarf lines didn’t produce any seeds, there were five dwarf lines in next generation (F3). Two out of five produced seeds and continued to produce dwarf phenotype in their next generation. Two of such samples, 1282 normal and dwarf (labeled 1282NF6 and 1282DF6) and 1303 normal and dwarf (labeled 1303NF6 and 1303DF6) from F6 generation were
chosen and used for genomics variation analysis using Illumina sequencing method. Additional two samples that exhibited semi-dwarf phenotype were also used for analysis viz. (labeled 1214NF6 and 1214DF6 and 1290NF6 and 1290DF6). The total number of reads obtained for 1282NF6 was 159,196,424, which accounted for 24 G bps. The GC content was 35.65% with a Q20 value of 94.82%. Likewise, the total reads for 1282DF6 were 172,215,288, accounting for 26 Gbps, and in line 1303, the numbers of total reads for the normal and dwarf samples were 150,712,004 and 139,484,320, accounting for 22 Gbps and 21 Gbps, respectively (Supplementary Table 1). The GC content for all the lines was above 35%, and the Q20 average percentage was above 94%. More than 93%, on average, was mapped to the reference genome. There were 5,597,100 variant calls in both genotypes (Table 1). The raw data was deposited in NCBI SRA database with an accession number PRJNA665611.

To estimate the number of variants that can be obtained in a hybrid of *G. max* and *G. soja* we need to know the total number of SNPs and indels in them. G. Ramakrishna et al. identified a total 77,339 SNPs and 451,522 indels in *G. max* whereas 215,932 SNPs and 697,295 indels in *G. soja*, with comparison to reference post-filling [30]. Among them, the number of common variants for both species was 10,873 SNPs and 80,078 indels. So if we exclude the common SNPs and indels from both species, we would be left with 282,398 SNPs and 1,069,739 indels making the total count of variations into 1,351,137. As per the Mendelian genetics, the cross be-

---

**Table 1.** Variant calling statistics as compared to reference genome

<table>
<thead>
<tr>
<th>Raw variants (SNP + INDEL)</th>
<th>SNP</th>
<th>INDEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw variants</td>
<td>Filtered</td>
<td>Filtered</td>
</tr>
<tr>
<td>5,597,100</td>
<td>4,645,717</td>
<td>944,240</td>
</tr>
<tr>
<td>4,108,601</td>
<td>904,602</td>
<td></td>
</tr>
</tbody>
</table>

SNP, single nucleotide polymorphism.
tween *G. max* and *G. soja* will distribute the variants into half from each parent to its offspring F1 into 1:2:1 ratio of both parents, then subsequently to the next F2 generation and so on [31].

With respect to that, theoretically the distribution of SNP in resulting cross should be 1 SNP per 0.748 kb in F1 generation (genome size 1,013,200 kb/total variation 1,351,137) [31]. Therefore in F2 generation the distribution is 1 SNP per 1.49 kb and so on. Consequently in theory in F6 generation there should be 1 SNP per 23.99 kb. From the filtered SNPs, we obtained average 217,764 and 57,403 homozygous SNP in dwarf lines and semi-dwarf lines of F6 generation (Table 2). Implying there is 1 SNP per 3.68 kb region of dwarf lines and 11.13 kb region of semi-dwarf lines (Fig. 2).

### Distribution of SNPs in coding regions of the reference genome

The genes that directly affect the growth of plants viz. plant defense, phytohormones, and photosynthesis were considered while comparing the SNPs in dwarf and normal lines. Furthermore we focused exclusively on missense SNPs on coding regions as they may cause base changes in protein sequence and alter the gene function. We observed 503 and 485 SNPs among dwarf and normal of 1,282 and 1,303, respectively, out of which 98 were common to both genotypes when comparison to reference genome. The highest number of SNPs was observed in NB-ARC (nucleotide-binding APAF-1 R proteins and CED-4) domain–containing disease resistance protein (75%), followed by disease resistance protein (TIR-NBS-LRR [toll interleukin 1 receptor nucleotide-binding site leucine-rich repeat resistance proteins] class) family (35%) and auxin-like 1 protein (27%) (Fig 3, Supplementary Table 2). The distribution was highest on chromosome (Chr) 18, followed by Chr 16, for both genotypes. When considering the individual nonsynonymous SNP distribution, we observed that Chr 16 had the maximum number of SNPs in both 1282 (74) and 1303 (67), followed by Chr 7, with 58 and 63 in 1282 and 1303, respectively (Supplementary Table 3). We discovered a total of 33

### Table 2. Number of candidate variants after filtering with reference genome

<table>
<thead>
<tr>
<th>Sample</th>
<th>Different between normal and dwarf (SNP)</th>
<th>Homozygous in dwarfism sample (SNP)</th>
<th>Different between normal and dwarf (INDEL)</th>
<th>Homozygous in dwarfism sample (INDEL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1282</td>
<td>458,209</td>
<td>182,497</td>
<td>108,277</td>
<td>51,622</td>
</tr>
<tr>
<td>1303</td>
<td>337,001</td>
<td>253,032</td>
<td>100,267</td>
<td>63,184</td>
</tr>
<tr>
<td>1214</td>
<td>116,136</td>
<td>65,599</td>
<td>55,767</td>
<td>30,225</td>
</tr>
<tr>
<td>1290</td>
<td>111,052</td>
<td>57,098</td>
<td>56,000</td>
<td>29,152</td>
</tr>
</tbody>
</table>

SNP, single nucleotide polymorphism.

![Figure 2](https://doi.org/10.5808/gi.21024)
homogeneous nonsynonymous SNPs that occurred at the same loci in each set of dwarf and normal soybean derived from normal soybean. We then identified the homogeneous SNPs across all the dwarf samples, with the highest representation of SNPs on Chr 16 and Chr 7 in both genotypes. SNPs that were common to both genotypes were highest on Chr 4, followed by Chr 7 and Chr 15 (Supplementary Table 3).

There were three nonsense SNPs found in both genotypes: one was on Chr 2, another was on Chr 13, and the last was on Chr 20; the SNP from Chr 20 was homogeneous in the dwarf lines (Supplementary Table 4). The three nonsense SNPs obtained in the genic regions included proteins with the gene functions single-stranded DNA (ssDNA)-binding transcriptional regulator, UDP-glycosyltransferase protein (UGT) and PIF1 helicase. Twenty frameshift SNPs were common to both genotypes (Supplementary Table 4). The frameshift mutations observed in the genic regions of the normal and dwarf individuals are shown in Supplementary Table 3. Out of 20 mutated genes, five were leucine-rich receptor-like protein family genes; three were proteins of unknown function (DUF647); two were NADH-ubiquinone/plastoquinone oxidoreductase chain 4L, 2 cytochrome P450, family 78, subfamily A, polypeptide 5; and one each was MUTS homolog 6, unknown protein and RNA helicase-like 8.

![Fig. 3. Distribution of missense single nucleotide polymorphisms (SNPs) in the soybean genome in genic regions. NB-ARC, nucleotide-binding APAF-1 R proteins and CED-4; TIR-NBS-LRR, toll interleukin 1 receptor nucleotide-binding site leucine-rich repeat resistance proteins.](https://doi.org/10.5808/gi.21024)

### Table 3. Chromosome wise distribution of SNP among normal and dwarf lines

<table>
<thead>
<tr>
<th>Chromosome No.</th>
<th>1282 Normal and dwarf</th>
<th>1303 Normal and dwarf</th>
<th>1282 + 1303 Common SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>54</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>36</td>
<td>71</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>23</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>27</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>22</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>13</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>186</td>
<td>241</td>
<td>5</td>
</tr>
</tbody>
</table>

SNP, single nucleotide polymorphism.

#### Distribution of SNPs among normal and dwarf plants

When we considered the SNP among normal and dwarf lines and not with the reference, 426 SNPs were obtained, among which five were common to both samples (1282 and 1303) (Table 3). The
SNPs were distributed on all chromosomes except 3, 4, 5, 16, and 17. Sample 1282 had the highest number of SNPs on Chr 7 (36), followed by Chr 1 (29) and Chr 13 (27). Likewise, sample 1303 had the highest number of SNPs on Chr 7 (71), followed by Chr 2 (54) and Chr 13 (28). The gene functions of the SNPs that were common to both normal and dwarf samples in both lines (1282 and 1303) included matrixin family protein, zinc finger protein, transcription factor and cytochrome P450 protein polypeptide (Table 4).

**Discussion**

RIL lines were developed from *G. max* and *G. soja* and none of the parent exhibited dwarf phenotype. From F3 generation we obtained few lines that exhibited dwarf phenotype which continued to produce dwarf lines in subsequent generation. We chose two lines from F6 that had both phenotypes (dwarf and normal) to survey the variance among them. Genome variant analysis was performed in such lines from F6 generation exhibiting dwarf and semi-dwarf phenotype. We observed that the dwarf lines in this study had a higher number of SNPs and indels than the semi-dwarf lines (Table 2). This implies that higher number of variations could cause changes in gene function causing dwarf phenotype. We obtained 5M of total variation in our data (Table 1). The number although is much smaller than that obtained in a recent study in *G. soja* [32] which was more than 15M SNPs and 14M indels. The reason could be due to the fact that they have used more than 26 accessions to determine variants whereas we have used only one Wm82 genome to look for the differences. Initially, we compared the number of SNPs with respect to reference *G. max* genome where we observed a smaller number of SNPs in semi-dwarf lines and a higher number in dwarf lines. The homozygosity in F5 and F6 RILs are 94% and 97%, respectively [31]. This can explain the higher numbers of variants which may also impact the phenotype. Although the number of SNPs observed in the F6 populations in our study was greater between the dwarf and normal plants (1 SNP per 3.68 kbp for dwarf and 11.13 kbp for semi-dwarf plants) (Fig. 2). This is much higher than the normal SNP frequency in the F6 generation, which is 1 per 23.99 kbp [33].

Natural variations in the genome, such as SNPs in coding regions, can alter amino acid sequences and modify the post-translation products, which may affect gene function [34,35]. Notably, disease resistance protein genes exhibited a high number of variants in our study. Similar studies have shown that changes in protein function (gain or loss of function) may contribute to dwarf phenotypes. Dwarfism due to a gain-of-function mutation in a TIR-NB-LRR protein was reported in *Arabidopsis* as one of the mechanisms underlying enhanced disease resistance [36]. Moreover, changes in these proteins cause autoimmunity in plants, and one of the useful features of autoimmune mutants is their dwarf phenotype [37]. Temperature and humidity are known to play important roles in dwarfism, although the exact mechanism is still unknown [38-42]. The RILs in our study were all grown in the same environmental conditions with the same temperature and humidity. This suggests that the dwarf phenotypes observed in our study were not due to temperature or humidity.

The SNPs were mostly in the NB-ARC domain-containing disease resistance protein, followed by disease resistance protein and auxin-like protein (Fig. 4). In *Arabidopsis*, NB-ARC mutants induced autoimmunity in plants, of which dwarfism is one of the typical phenotypes [37]. Common SNPs occurring in gene coding regions could affect the phenotype, whether or not they are combined with other genes. There is not much known about the matrixin family protein in plants, but the members of the matrix metalloproteinase family are thought to be involved in remodeling of the extracellular matrix during plant growth and development [43]. Thus, an SNP in a gene encoding a member of this protein family might have caused an alteration in protein function leading to dwarfing. Mutations in the transcription factor jumonji (jmjC) have been reported to complement plant growth defects and expression changes [44]. Overexpression of TTF-type zinc finger protein in *Arabidopsis* resulted in divergent physiological and metabolic phenotypes, some of which were significant for improved plant performance [45]. The SNPs occurring in these vital genes

Table 4. Common SNP among normal and dwarf lines

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome No.</th>
<th>1282NF6</th>
<th>1282DF6</th>
<th>1303NF6</th>
<th>1303DF6</th>
<th>TAIR TOP hit function</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLYMA01G04370</td>
<td>1</td>
<td>G/G</td>
<td>A/G</td>
<td>G/G</td>
<td>A/G</td>
<td>Matrixin family protein</td>
</tr>
<tr>
<td>GLYMA01G06671</td>
<td>1</td>
<td>C/C</td>
<td>T/T</td>
<td>T/T</td>
<td>C/C</td>
<td>TTF-type zinc finger protein with HAT dimerization domain</td>
</tr>
<tr>
<td>GLYMA19G14700</td>
<td>19</td>
<td>A/G</td>
<td>A/A</td>
<td>A/G</td>
<td>A/A</td>
<td>Transcription factor jumonji (jmjC) domain–containing protein</td>
</tr>
<tr>
<td>GLYMA19G14700</td>
<td>19</td>
<td>G/G</td>
<td>G/A</td>
<td>G/A</td>
<td>A/A</td>
<td>Transcription factor jumonji (jmjC) domain–containing protein</td>
</tr>
<tr>
<td>GLYMA1057S00200</td>
<td>20</td>
<td>T/C</td>
<td>T/T</td>
<td>T/T</td>
<td>T/C</td>
<td>Cytochrome P450, family 76, subfamily C, polypeptide 4</td>
</tr>
</tbody>
</table>

SNP, single nucleotide polymorphism.
may have contributed to the dwarf phenotypes found in the RILs. There were nonsense SNPs in three loci: ssDNA-binding transcriptional regulator, UGT superfamily protein, and PIF1 helicase. ssDNA-binding transcriptional regulators are known to function as positive and negative regulators in leaf senescence [46]. UGTs also act as major contributors to plant growth, including development, disease resistance, and interaction with the environment, by interacting with various substrates, such as flavonoids, terpenes, auxins, cytokinin, and many others [47]. PIF1 helicases are enzymes that are essential in DNA replication, repair, and recombination in all organisms. Likewise, frameshift mutations were found in genes with important functions, such as leucine-rich receptor-like protein family, protein of unknown function (DUF647), NADH-ubiquinone/plastoquinone oxidoreductase chain 4L, cytochrome P450 family 7-subfamily A polypeptide 5, MUTS homolog 6, unknown protein and RNA helicase-like 8. All these proteins are major regulators and contributors to plant growth and development.

The SNPs obtained in our study (missense, nonsense SNP, and frameshift mutations) were in vital genes but may or may not have impacted plant growth. We were unable to derive any particular conclusion about the dwarf phenotype based on our results, but this study will provide a basis to analyze further and evaluate which SNPs affect plant growth type. The identification of functional SNPs in genes and analysis of their effects on phenotype may lead to a better understanding of their impacts on gene function and thus support varietal improvement.

**ORCID**

Neha Samir Roy: https://orcid.org/0000-0002-4529-4861
Yong-Wook Ban: https://orcid.org/0000-0002-1912-0374
Hana Yoo: https://orcid.org/0000-0002-3451-4646
Rahul Vasudeo Ramekar: https://orcid.org/0000-0003-3461-3617
Eun Ju Cheong: https://orcid.org/0000-0002-2576-5435
Nam-Il-Park: https://orcid.org/0000-0001-6725-0758
Jong Kuk Na: https://orcid.org/0000-0003-2616-4890
Kyong-Cheul Park: https://orcid.org/0000-0002-3737-815X
Ik-Young Choi: https://orcid.org/0000-0003-4168-0471

**Authors’ Contribution**

Conceptualization: KCP, NIP, IYC. Data curation: RVR, NSR, JKN. Funding acquisition: IYC. Methodology: YWB, HY, EJC, KCP. Writing - original draft: YWB, NSR, IYC. Writing - review & editing: NSR, IYC.

**Conflicts of Interest**

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

**Acknowledgments**

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2017R1A2B4011198)

**Supplementary Materials**

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

**References**


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 (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.

**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)
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
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://kci.go.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).

Author can archive publisher’s version/PDF.

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

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, informed consent form, 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 xox@yxx. 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>
</tbody>
</table>

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 programmatically 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.

Title of article

Author(s)

Author’s signature

Date

Taesung Park
Editor in Chief
Genomics & Informatics
Korea Genome Organization (KOGO)
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.