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 2020, the Korea Genome Organization

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

Editor’s introduction to this issue (G&I 18:4, 2020)
Taesung Park

Review articles

The nature of triple-negative breast cancer classification and antitumoral strategies
Songmi Kim, Dong Hee Kim, Wooseok Lee, Yong-Moon Lee, Song-Yi Choi, Kyudong Han

The ceRNA network of lncRNA and miRNA in lung cancer
Danbi Seo, Dain Kim, Yeonsoo Chae, Wanyeon Kim

Original articles

Transcriptome analysis of iBET-151, a BET inhibitor alone and in combination with paclitaxel in gastric cancer cells
Sun Kyoung Kang, Hyun Joo Bae, Woo Sun Kwon, Jingmin Che, Tae Soo Kim, Hyun Cheol Chung, Sun Young Rha

Investigation of gene–gene interactions of clock genes for chronotype in a healthy Korean population
Mira Park, Soon Ae Kim, Jieun Shin, Eun-Jeong Joo

Prediction of hub genes of Alzheimer’s disease using a protein interaction network and functional enrichment analysis
Jia Jin Wee, Suresh Kumar

Development of reverse transcription loop-mediated isothermal amplification assays for point-of-care testing of avian influenza virus subtype H5 and H9
Songzi Zhang, Juyoun Shin, Sun Shin, Yeun-Jun Chung

An information-theoretical analysis of gene nucleotide sequence structuredness for a selection of aging and cancer-related genes
David Blokh, Joseph Gitarts, Ilia Stambler

Comparative analysis of commonly used peak calling programs for ChIP-Seq analysis
Hyeongrin Jeon, Hyunj Lee, Byunghhee Kang, Insoon Jang, Tae-Young Roh

Druggability for COVID-19: in silico discovery of potential drug compounds against nucleocapsid (N) protein of SARS-CoV-2
Manisha Ray, Saurav Sarkar, Surya Narayan Rath

Spike protein D614G and RdRp P323L: the SARS-CoV-2 mutations associated with severity of COVID-19
Subrata K. Biswas, Sonchita R. Mudi

Research communication

Updated confidence intervals for the COVID-19 antibody retention rate in the Korean population
Md. Kamruzzaman, Catherine Apio, Taesung Park

Application note

BaSDAS: a web-based pooled CRISPR-Cas9 knockout screening data analysis system
Young-Kyu Park, Byoung-Ha Yoon, Seung-Jin Park, Byung Kwon Kim, Seon-Young Kim
Editor’s introduction to this issue (G&I 18:4, 2020)

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

In this issue, there are 12 articles: two Review Articles, eight Original Articles, one Research Communication, and one Application Note. The first review by Kim et al. (Dankuk University, Cheonan, Korea) is about the nature of triple-negative breast cancer (TNBC) classification and antitumoural strategy. It is well known that TNBC, characterized by estrogen receptor, progesterone receptor genes, and human epidermal growth factor receptor 2 genes, has highly aggressive clinical outcomes. The authors provided a further classification into specific subtypes according to their genomic mutation and cancer immunogenicity with its promising treatments. The second review by Seo et al. (Korea National University of Education, Cheonju, Korea) provides an overview about the roles of long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) in lung cancer cells and those of the lncRNA/miRNA competing endogenous RNA networks in carcinogenesis and therapeutic resistance of lung cancer.

Among eight original articles, there are three articles related to clinical studies. The first original article by the group of Rha (Yonsei University College of Medicine, Seoul, Korea) is about gastric cancer. The authors performed transcriptome analysis of iBET-151, a BET inhibitor, which reduces the expression of oncogenes. Analysis of RNA sequencing of gastric cancer cells treated with iBET-151 and/or paclitaxel identified the differentially expressed genes associated with possible mechanisms of synergistic effect. Although additional functional studies are needed, it is the first evidence of the synergistic effect between iBET-151 and paclitaxel in regulating the transcriptome of gastric cancer cells. The second original article is about for chronotype which is an important moderator of psychiatric illnesses. Park et al. (Eulji University, Korea) investigated genetic associations and gene-gene interactions for chronotype. The clock genes such as BHLHB2, CLOCK, CSNK1E, NR1D1, PER1, PER2, PER3, and TIMELESS were successfully identified from 1,293 healthy Korean individuals via regression analysis and the quantitative multifactor dimensionality reduction method.

The third original article by Wee and Kumar (Management and Science University, Shah Alam, Malaysia) is about Alzheimer’s disease. The authors identified the hub genes of Alzheimer’s disease. Through network analysis of protein-protein interactions, the authors successfully identified the top 10 hub genes associated with Alzheimer’s disease: PTGER3, C3AR1, NPY, ADCY2, CXCL12, CCR5, MTNR1A, CNR2, GRM2, and CXCL8.

Next, Chung (The Catholic University, Seoul, Korea)’s group developed reverse transcription loop-mediated isothermal amplification (LAMP) assays for the point-of-care testing of avian influenza virus subtype H5 and H9 by adopting LAMP technology. The new sets of reverse transcription LAMP are shown to be approximately four times quicker than conventional reverse-transcription polymerase chain reaction.

The next two original articles are about an algorithm of bioinformatics and comparison studies for the bioinformatics tools. Stambler (Bar Ilan University, Ramat Gan, Israel) proposed an algorithm for the construction and analysis of autocorrelation (information)
functions to identify aging and cancer-related genes based on the normalized mutual information. Roh’s group (Pohang University of Science and Technology, POSTECH, Pohang, Korea) performed a comparative study for commonly used peak calling programs for ChiP-Seq (chromatin immunoprecipitation coupled with high-throughput DNA sequencing) analysis such as five peak callers (CisGenome, MACS1, MACS2, PeakSeq, and SISSRs) using 12 publicly available ChiP-Seq data.

The coronavirus disease 2019 (COVID-19) is a highly contagious disease and created widespread mortality and morbidity. Recent efforts on COVID-19 pandemics have produced large volume of research articles. In this issue, there are three articles related to COVID-19. The first research article by Rath’s group (Odisha University of Agriculture and Technology, Bhubaneswar, India) analyzed the phylogenetic relationship of N protein sequence divergence with other 49 coronavirus species and identified the conserved regions according to protein families through conserved domain search. The authors established antiviral drug glycyrrhizic acid and the phytochemical theaflavin as possible drug compounds against target N-protein of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) with lower binding affinities. The second article by Biswas and Mudi (Bangabandhu Sheikh Mujib Medical University, Dhaka, Bangladesh) presents the comparison results of mutation profiles of SARS-CoV-2 isolated from mildly affected and severely affected COVID-19 patients and explores relationship between mutation profile and disease severity.

Through the genomic sequences of SARS-CoV-2, spike protein D614G and RNA-dependent RNA polymerase P323L mutations in SARS-CoV-2 were shown to be associated with severity of COVID-19. The final article on COVID-19 is a short research communication by Kamruzzaman et al. (Seoul National University, Seoul, Korea). It is the updated 95% confidence intervals COVID-19 antibody rate for the Korean population using three recently performed antibody tests in Korea. The most conservative 95% confidence interval estimation shows that as of 00:00 November 23, 2020, at least 69,524 people were infected but not confirmed. More positive cases were found among the young male in their twenties (0.22%), three times that of the public (0.051%).

The one article of application note by Kim’s group (Korea Research Institute of Bioscience and Biotechnology and University of Science and Technology, Daejeon, Korea) developed the BaSDAS (Barcode-Seq Data Analysis System), a GUI-based pooled knockout screening data analysis system with a user-friendly web interface. The BaSDAS facilitates the analysis of pooled knockout screen data easily and supports the analysis of various pooled screening libraries including yeast, human and mouse with many useful statistical and visualization functions.

ORCID

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

Breast cancer is one of the most common cancers in women and is a leading cause of cancer-related deaths. In 2018, ~266,000 women were expected to be diagnosed with invasive breast cancer in the United States, and 40,920 women were expected to die from it [1]. Triple-negative breast cancer (TNBC), which accounts for 15% of all breast cancers, is characterized by the lack of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) expression, which means that TNBC patients do not benefit from hormonal therapy or trastuzumab, which targets HER2 [2-4]. TNBC generally occurs in younger women and compared to other types of breast cancer, it shows a high histologic grade, a high propensity to metastasize to distant organs, and a poor outcome with a high recurrence rate after adjuvant therapy, which mainly consists of systemic cytotoxic chemotherapy [4]. Therefore, a significant challenge is to identify new targets and associated biomarkers for its treatment. Compared to other breast cancers, TNBC is a heterogeneous disease with different subtypes determined by distinct biological features. Luminal breast cancer is characterized by a relatively high expression of estrogen receptor (ER) and progesterone receptor (PR) genes, which are expressed in breast luminal cells. In ~25% of invasive breast cancers, human epidermal growth factor receptor 2 (HER2) is overexpressed; these cancers are categorized as the HER2 type. Triple-negative breast cancer (TNBC), in which the cancer cells do not express ER/PR or HER2, shows highly aggressive clinical outcomes. TNBC can be further classified into specific subtypes according to genomic mutations and cancer immunogenicity. Herein, we discuss the brief history of TNBC classification and its implications for promising treatments.

Keywords: classification, gene expression, immune checkpoint blockade, microbiome, subtype, triple-negative breast cancer
cancer types, TNBC is characterized by a higher mutational load, which makes the tumor immunogenic and amenable to immunotherapeutic treatment [4]. Thanks to the promising results of immunotherapy in some cancers such as malignant melanoma and non-small cell lung cancer [5,6], several clinical trials are currently assessing immunotherapeutic approaches in TNBC patients [4,7-11]. The U.S. Food and Drug Administration (FDA) approval of nanoparticle albumin-bound paclitaxel (nab-paclitaxel) combined with atezolizumab in 2019 presented an innovative therapeutic development for TNBC patients. In this review, we discuss a brief history of TNBC classification based on gene expression patterns, as well as promising anticancer strategies for TNBC.

What Is TNBC?

In 1999, a molecular classification of breast cancer was first proposed by the National Cancer Institute (NCI) [12]. In 2000, Perou et al. [13] introduced a classification of breast cancer into four types: luminal, basal-like, HER2, and normal breast-like. The mammary epithelium consists of two layers: luminal cells and basal (myoepithelial) cells. The term “luminal” refers to the part containing mammary lobules and ductal structures, which are the major targets of estrogen and progesterone and mature to produce milk. In contrast, the term “basal” denotes another part of the mammary epithelium that supports the lobular and ductal structures. Sorlie et al. [14] further subdivided the luminal type into luminal A and B. The luminal A type demonstrated relatively high expression of luminal epithelial gene cluster, including ER, GATA3, X-box binding protein 1, trefoil factor 3, hepatocyte nuclear factor 3a, FOXA1, and LIV-1. In contrast, the luminal B type showed high expression of proliferation-related genes and low to moderate expression of luminal-specific genes [14]. In a following study, breast cancer was classified into three major types based on the expression or absence of ER, PR, and HER2, namely ER+/PR+/HER2– (luminal A type), HER2+, and TNBC. The HER2+ type can be further divided into two subtypes: ER+/PR+/HER2+ (luminal B type) and ER–/PR–/HER2+ (HER2+ type) [15]. TNBC often occurs in under 40-year-old women and has a mortality rate of 40% within the first 5 years after diagnosis, a median survival time after metastasis of only 13.3 months, and a recurrence rate after surgery of 25%. Therefore, gene expression profiling analysis has begun to enable a deeper understanding of this disastrous disease, which is not sensitive to hormone therapy or HER2 treatment [2].

The Classification Begins!

Classification and analysis are milestones for scientific understanding. Now, it is TNBC’s turn. In 2011, Lehmann et al. [9] classified TNBC into six subtypes using 587 TNBC patients based on transcriptome profiling data as follows:

1. The basal-like 1 subtype is involved in cell cycle and cell division pathways and shows overexpression of the AURKA, AURKB, BIRC5, BUB1, CENPA, CENPF, CCNA2, MYC, NRAS, PRC1, PLK1, and TTK genes. In addition, the expression of DNA repair (ATR/BRCA pathway)–related genes, such as CHEK1, FANCA, FANCG, RAD54BP, RAD51, NBN, EXO1, MSH2, MCM10, RAD21, and MDC1, is significantly increased.

2. The basal-like 2 subtype is associated with increased expression of growth factor signaling pathways, including the EGF, NGF, MET, Wnt/β-catenin, and IGFIR pathways.

3. The immunomodulatory subtype (IM subtype) shows high activation of immune signaling pathways (CTLA4, natural killer (NK)-cell, Th1/Th2, NFKB, TNE, T-cell, JAK/STAT, ATR/BRCA), and cytokine signaling pathways, such as the interleukin (IL)-12 and IL7 pathways.

4. The mesenchymal subtype (M subtype) exhibits significantly lower expression levels in the immune signal transduction pathway, unlike the IM subtype. The M subtype also shows profound activation of cell migration-related signaling pathways, extracellular matrix receptor interaction pathways, and cell differentiation pathways, such as the Wnt pathway, anaplastic lymphoma kinase pathway, and transforming growth factor (TGF)-β pathway. These molecular changes result in sarcomatous morphological features.

5. The mesenchymal stem-like subtype (MSL subtype) features high expression of stemness-related pathways, including the inositol phosphate metabolism pathway, G-protein-coupled receptor pathway, and calcium signaling pathway. In addition, the MSL subtype displays high expression of angiogenesis pathways such as KDR, TEK, TIE1, and EPAS1, but very low expression of the proliferative pathway. Moreover, this subtype is accompanied by high expression of stem cell markers (ABCA8, PROC, ENG, ALDHA1, PER1, ABCB1, TERT2IP, and BCL2) and mesenchymal stem cell-specific markers (BMP2, ENG, IITV, KDR, NGFR, NTSE, PDGFR, THY1, and VCAM1).

6. The luminal androgen receptor subtype (LAR subtype) displays high expression of hormonal-related signaling pathways, including steroid synthesis, porphyrin metabolism, and an-
Drugs/estrogen metabolism.

Subsequently, Burstein et al. [16] proposed four molecular subgroups using gene expression profiling of 198 TNBC cases as follows.

1) The LAR subgroup is characterized by gene expression for hormone-related signaling pathways, including prolactin signaling and estrogen/androgen metabolism. The tumors within this subgroup show androgen receptor, ER, prolactin, and ErbB4 signaling, but are ERα-negative by immunohistochemistry (IHC) staining. ESR1 and other estrogen-regulated genes such as PGR, FOXA1, XBPI, and GATA3 are expressed. This group demonstrates ER activation despite belonging to the category of ER-negative tumors by IHC, suggesting that traditional anti-estrogen therapies and anti-androgen therapies might be useful [9].

2) The mesenchymal subgroup shows activation of pathways related to the complement system, prothrombin activation, coagulation system, leukocyte extravasation signaling, and hepatic stellate cell activation signaling. In addition, this subgroup shows down-regulation of several signaling pathways, including cell cycle, mismatch repair, and hereditary breast carcinoma signaling pathways. In general, genes exclusive to osteocytes (OGN) and adipocytes (ADipoQ, and PLIN1), and insulin-like growth factors (IGF-1) are highly expressed in this subgroup.

3) The basal-like immune-suppressed (BLIS) subgroup is characterized by down-regulation of B cell, T cell, and NK cell immune-regulating pathways and cytokine pathways. Activation of the cell cycle and DNA repair-related signaling pathways has also been identified in this subgroup.

4) The basal-like immune-activated (BLIA) subgroup, unlike the BLIS subgroup, shows up-regulation of B cell, T cell, and NK cell immune-regulating pathways. Additionally, the expression levels of STAT genes are elevated, and STAT transcription factor-mediated pathways are highly activated in this subgroup.

Recently, Liu et al. [17] proposed four new subtypes after a classification analysis of the gene expression profile combined with mRNAs and long noncoding RNAs in 165 TNBC samples, as follows.

1) The IM subtype has high expression levels of genes related to innate immune response T-cell co-stimulation and the immune response, such as CCR2, CXCL13, CXCL11, CD1C, CXCL10, and CCL5.

2) The LAR subtype, despite being ER-negative on IHC staining, shows activation of the ER signaling pathway. Steroid biosynthesis, porphyrin metabolism, androgen/estrogen metabolism, and peroxisome proliferator-activated receptor signaling pathways are highly activated in this subtype.

3) The mesenchymal-like subtype is enriched with various gene ontology category members and signaling pathways, such as extracellular matrix-receptor interactions, gap junctions, TGF-β, growth factor pathways, and the adipocytokine signaling pathway. Contrariwise, the mesenchymal-like subtype shows down-regulation of cell proliferation-related genes (cell division process, mitotic cell cycle, mitotic prometaphase, and mitosis).

4) The BLIS subtype is highly enriched in cell division and cell cycle-related signaling pathways, including DNA replication, DNA repair, mitotic cell cycle, mitotic prometaphase, and the M phase of the mitotic cell cycle. The BLIS subtype has high expression of proliferation-related genes, such as CENPF, BUB1, and PRC1, but this subtype is characterized by significant down-regulation of immune cell signaling pathways, immune response, and complement activation processes. The TNBC subtypes discussed above are summarized in Table 1.

### Immune Checkpoint Blockade Therapy in Certain TNBC Subtypes

Targeting cytotoxic T-lymphocyte-associated antigen 4 (CTLA-4), programmed cell death protein 1 (PD-1), and its ligand (PD-L1) has revolutionized cancer treatment. CTLA-4 signaling is more prevalent in lymph nodes, and PD-1/PD-L1 is involved in multiple processes in the tumor microenvironment, such as suppression of T-cell responses and T-cell anergy. Blockade of PD-1/PD-L1 can significantly induce T-cell proliferation and activity, generating antitumor immune responses [6,18]. Thanks to this, immune checkpoint blockade (ICB) therapies have become a mainstay of treatment. According to Karn et al. [19], TNBC can be divided into immune-rich or immune-poor subtypes based on metagenes for a high lymphocyte infiltration (major histocompatibility complex class II) gene signature and low inflammation markers, such as interleukin-8 and vascular endothelial growth factor. Tumors with high immune gene expression had a better prognosis due to strict immunosurveillance and were associated with a low level of clonal heterogeneity, somatic copy number alteration level, mutations, and neoantigen load [19,20]. Of the TNBC subtypes, the IM and BLIA subtypes are classified as immune-rich.

CD8+ T cells are activated by antigen-presenting cells that present neoantigens on major histocompatibility complex class I and II
They are directly activated by their receptor and further regulated by a complex interplay of co-stimulatory and co-inhibitory signals, known as immune checkpoints [8]. Through these mechanisms, tumors can hijack physiological immune responses and cause immune tolerance. Sequential clinical trials on the treatment of TNBC with anti–PD-1/PD-L1 monoclonal antibodies (pembrolizumab and atezolizumab), significantly increased overall survival rates. In 2019, the U.S. FDA approved the use of nab-paclitaxel in combination with atezolizumab for PD-L1+ TNBC. At the moment, the primary challenge is to improve the response of patients with TNBC to anti–PD-1/PD-L1 treatment and to convert non-responders into responders.

What about the Microbiome?

Strong evidence from recent studies has suggested that the gut microbiota can affect the host’s antitumor immunity and that the composition of the intestinal microbiome may modulate the efficacy of ICB therapy in mice and humans [22-28]. Previous studies revealed that specific components of the gut microbiome might influence the efficacy of ICB therapy in patients, and primary resistance to ICB therapy could be overcome through treatment with ICB-promoting bacteria [29-33]. Notably, Mager et al. [30] demonstrated that *Bifidobacterium pseudolongum*, isolated from ICB-treated tumors, modulated enhanced ICB effects through production of the soluble metabolite inosine. Elevated systemic inosine levels and activated antitumor T cells are caused by the induction of decreased gut barrier function, resulting from ICB therapy in colorectal adenocarcinoma, urothelial carcinoma, and melanoma mice models. Inosine binding to T cell--specific adenosine 2A receptor (A2AR) promotes Th1 cell activation [30]. Previous studies have demonstrated that inosine and A2AR binding exert an inhibitory effect on Th1 differentiation *in vitro* and antitumor immunity *in vivo* [34-36]. Emerging studies have implicated the involvement of the gut microbiome in response to breast cancer treatment. Recently, in 2018, Banerjee et al. [22] discovered a predominant microbial signature in TNBC, which includes the families Caulobacteriaceae, Actinomycetaceae, Enterobacteriaceae, Prevotellaceae, Sphingobacteriaceae, Brucellaceae, Flavobacteriaceae, and Bacillaceae. In conclusion, some types of gut microbiota and their metabolites are available to develop microbial-based adjuvant therapies that enhance the effectiveness of ICB therapy in TNBC patients (Fig. 1).

![Fig. 1. Schematic relations among ICB therapies and gut microbiota. The well-established anticancer mechanisms of anti–PD-1, anti–PD-L1, and anti–CTLA-4 antibodies (bold text), as well-known ICB therapies, are illustrated; the efficacy of these therapies can be augmented by gut microbiota. The cytotoxic T-cell recognizing tumor antigen (red circle) causes TNBC cell death by blocking calm-down signaling using ICBs. Gut microbiota metabolites can strengthen the anticancer effects of anti–CTLA-4 antibody and present the tumor antigen to T-cells interacting with dendritic cells (upper right). ICB, immune checkpoint blockade; PD-1, programmed cell death protein 1; PD-L1, programmed death-ligand 1; CTLA4, cytotoxic T-lymphocyte-associated antigen 4; TNBC, triple-negative breast cancer; TCR, T-cell receptor.](https://doi.org/10.5808/GI.2020.18.4.e35)
Conclusion

TNBC is the most aggressive type of breast cancer and has a poor prognosis. Because it lacks the expression of hormone receptors and HER2 receptors, therapy is primarily based on systemic chemotherapy, rather than targeted agents. Although TNBC has a higher mutational load than other breast cancer types, it generally responds well to ICB therapy. This led to FDA approval for the use of nab-paclitaxel in combination with atezolizumab in TNBC patients. Thanks to new promising treatments such as ICB mono-therapy and ICB therapy combined with conventional systemic chemotherapy, we currently face exciting new perspectives. As discussed above, many efforts are being made to increase the antitumor effects of ICB, and studies are intensively investigating the individual gut microbiota. Collectively, this research program will provide a profound understanding of TNBC pathology and insights into antitumor mechanisms, the first step in developing therapeutic strategies for this devastating type of breast cancer.

ORCID

Songmi Kim: https://orcid.org/0000-0002-5497-1174
Dong Hee Kim: https://orcid.org/0000-0002-5056-7406
Wooseok Lee: https://orcid.org/0000-0002-5437-1814
Yong-Moon Lee: https://orcid.org/0000-0002-4302-9263
Song-Yi Choi: https://orcid.org/0000-0002-8496-5613
Kyudong Han: https://orcid.org/0000-0001-6791-2408

Authors’ Contribution

Conceptualization: SYC, KH. Data curation: DHK. Formal analysis: SK. Funding acquisition: KH. Methodology: WL. Writing - original draft: SK, YML. Writing - review & editing: SYC, KH.

Conflicts of Interest

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

Acknowledgments

This research was supported by the Basic Science Research Capacity Enhancement Project through Korea Basic Science Institute (National Research Facilities and Equipment Center) grant funded by the Ministry of Education (Grant No. 2019R1A6C1010033).

References


Table 1. Summary of the molecular classification of triple-negative breast cancer by gene expression

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of cases</td>
<td>587 (21 public datasets)</td>
<td>198</td>
<td>165</td>
</tr>
<tr>
<td>Subtype/Dysregulated pathway</td>
<td>6 subtypes</td>
<td>4 subtypes</td>
<td>4 subtypes</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>BL1</td>
<td>BL1</td>
<td>BL1</td>
</tr>
<tr>
<td>DNA repair</td>
<td>BL2</td>
<td>BL2</td>
<td>BL2</td>
</tr>
<tr>
<td>Immune signaling</td>
<td>IM</td>
<td>BU A</td>
<td>IM</td>
</tr>
<tr>
<td>EMT signaling</td>
<td>M</td>
<td>MES</td>
<td>ML</td>
</tr>
<tr>
<td>Stemness-related signaling</td>
<td>MSL</td>
<td>MSL</td>
<td>MSL</td>
</tr>
<tr>
<td>Hormone-related signaling</td>
<td>LAR</td>
<td>LAR</td>
<td>LAR</td>
</tr>
<tr>
<td>Modality</td>
<td>Gene expression profile</td>
<td>Gene expression profile</td>
<td>Gene expression profile</td>
</tr>
<tr>
<td></td>
<td>copy number variation</td>
<td>(mRNA + lncRNA)</td>
<td></td>
</tr>
</tbody>
</table>

BL1, basal-like 1; BLIS, basal-like immune-suppressed; BL2, basal-like 2; IM, immunomodulatory; BU A, basal-like immune-activated; M, mesenchymal; MES, mesenchymal; ML, mesenchymal-like; MSL, mesenchymal stem-like; LAR, luminal androgen receptor; lncRNA, long noncoding RNA.


The ceRNA network of IncRNA and miRNA in lung cancer

Danbi Seo¹, Dain Kim¹, Yeonsoo Chae¹,², Wanyeon Kim¹,³*

¹Department of Science Education, Korea National University of Education, Cheongju 28173, Korea
²Department of Science Education, Chungbuk Science High School, Cheongju 28189, Korea
³Department of Biology Education, Korea National University of Education, Cheongju 28173, Korea

Since lung cancer is a major causative for cancer-related deaths, the investigations for discovering biomarkers to diagnose at an early stage and to apply therapeutic strategies have been continuously conducted. Recently, long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) are being exponentially studied as promising biomarkers of lung cancer. Moreover, supportive evidence provides the competing endogenous RNA (ceRNA) network between lncRNAs and miRNAs participating in lung tumorigenesis. This review introduced the oncogenic or tumor-suppressive roles of lncRNAs and miRNAs in lung cancer cells and summarized the involvement of the lncRNA/miRNA ceRNA networks in carcinogenesis and therapeutic resistance of lung cancer.

**Keywords:** competing endogenous RNA, long non-coding RNA, lung tumorigenesis, miRNA, therapeutic resistance

**Introduction**

Lung cancer is a malignant disease of the lungs and accounts for a large proportion of all cancer-attributable deaths [1]. Lung cancer is classified as non-small cell lung cancer (NSCLC) or small cell lung cancer (SCLC), and NSCLC accounts for around 80% of all lung cancers [2]. Lung cancer is one of the most difficult cancers to diagnose at an early stage because it has few initial symptoms [3]. However, early diagnosis and surgical treatment provide the best strategy in terms of increasing survival rates [4], and thus, medically applicable biomarkers are required for early diagnosis.

Only 2% of RNA transcribed from human DNA encodes proteins, and the remaining 98% is referred to as non-coding RNA (ncRNA) [5]. Although not translated into proteins, ncRNAs perform various functions within cells [6,7] and have potential use as biomarkers for the diagnosis of lung cancer. Long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) are representative examples of ncRNA [8,9] and are being actively studied as potential biomarkers. Furthermore, accumulating evidence indicates competing endogenous RNA (ceRNA) networks of IncRNA and miRNA affect carcinogenesis. This review addresses the association between the ceRNA network of IncRNAs and miRNAs and the development of lung cancer and provides an overview of the effect of this network on the survival, proliferation, motility, and radiation and drug resistances of lung cancer cells.
The Roles of miRNAs and LncRNAs in Lung Cancer

MiRNA is around 22 nucleotides in length and an evolutionarily conserved type of ncRNA that suppresses gene expression by interacting directly with DNA and RNA [10,11]. In cytoplasm, miRNA inhibits translation by destabilizing and causing the degradation of mRNA, regulating transcriptional stability in the nucleus, and recruiting epigenetic remodeling factors to induce gene silencing [12]. MiRNA is involved in various biological processes such as cell cycle development, cell differentiation, and does so by regulating the expressions of target genes. Furthermore, abnormal miRNA expression is associated with many diseases including cancer [13].

In particular, miRNA may be directly involved in carcinogenesis. Genomic instability is one of the hallmarks of cancer and facilitates tumorigenic process [14]. Certain miRNA genes are placed in chromosomes which are susceptible to damage and mutation, and physical disruption of these miRNAs may be responsible for a wide range of abnormalities in the expressions of genes that play critical roles in the cell cycle, DNA repair, and apoptosis. In cancer, miRNAs are classified as oncogenic miRNAs (onco-miRs) or tumor-suppressive miRNAs. Onco-miRs target tumor-suppressive mRNAs, CDK inhibitors, and pro-apoptotic members of the Bcl-2 family, and thus, promote tumor growth and anti-apoptotic signaling [15]. In contrast, tumor-suppressive miRNAs disrupt the expressions of oncogenic mRNAs such as those of cyclins, CDKs, and genes that are directly and indirectly involved in growth factor-mediated signaling pathways and inhibit cell proliferation and survival [16]. Various miRNAs have been shown to be involved in cell proliferation and death in lung cancer. In NSCLC cells, miR-21 affected cell growth and invasion by targeting the PTEN transcript [17], whereas miR-451a regulated the migration and invasion of lung cancer cells by targeting ATF2 [18]. Interestingly, these miRNAs are potentially involved in the development of lung cancer and can be regulated by interaction with lncRNAs.

LncRNAs, like miRNAs, represent a major group within the ncRNA family. LncRNAs are considerably longer than miRNAs; they contain around 200 nucleotides and are more than 100 kb long and some have poly A tails [19,20]. In cells, lncRNAs play a variety of roles, such as activating signaling pathways, modifying chromatin, and regulating transcription and translation [21]. In particular, lncRNAs can regulate mRNA expression by competing with miRNA in cytoplasm [22]. It was suggested some lncRNAs have sponge-like effects on miRNAs that attenuate the effects of mRNAs, which is referred to as the ceRNA hypothesis [23,24]. In fact, many lncRNAs have miRNA-binding sites that regulate the expressions of genes encoding proteins [25]. LncRNA, which functions as a ceRNA, sequesters miRNA and prevents them regulating the translations of target mRNAs (Fig. 1). In 2014, lncRNA AK048451, which is called cardiac hypertrophy related factor, was first identified as a ceRNA of miR-489 and found to inhibit miR-489 expression by direct binding in a sequence-specific manner [20]. Furthermore, abnormal expressions, mutations, and single nucleotide polymorphisms of lncRNA have been associated with tumor formation and metastasis [26], and accumulating evidence indicates networks of lncRNAs, miRNAs, and mRNAs importantly contribute to the epithelial-to-mesenchymal transition (EMT), onset and progression of cancer [27,28]. For example, lncRNAs (MEG3, MIAT, and LINC00115) were found to play important roles in carcinogenesis by regulating miRNA-mRNA networks in lung cancer [29].

Oncogenic LncRNAs Acting as ceRNAs in Lung Cancer

Oncogenic lncRNAs are generally upregulated in lung cancer cells and tissues and bind directly to tumor-suppressive miRNAs. Direct lncRNA to miRNA binding upregulates the expressions of oncogenic mRNAs (a target of miRNAs), and thus, promotes cancer cell growth and development. Several lncRNAs that function as oncogenes in lung cancer have been identified (Table 1), for example, lncRNA H19 is highly expressed in the A549, H1299, H23, and SPC-A1 lung cancer cell lines, and inhibits miR-200a, miR-196b, and miR-29b-3p [30-32]. The interaction between lncRNA H19 and miR-200a (a tumor-suppressive miRNA downregulated in patients with a high lung cancer stage) regulates the expressions of ZEB1 and ZEB2 [30]. LIN28B is a target of miR-196b and can function as a proto-oncogene, and lncRNA H19 can upregulate LIN28B by ‘sponging’ miR-196b [31]. MiR-29b-3p is involved in the regulation of apoptosis, the cell cycle, and metastasis, and its targeting by lncRNA H19 transforms STAT3 (signal transducer and activator of transcription 3), and thus, promotes the survival and EMT of lung adenocarcinoma cells [32].

MALAT1 (metastasis associated in lung adenocarcinoma transcript 1) is another representative oncogenic lncRNA and is highly expressed in the A549 and H1299 lung cancer cell lines, in which miR-124 is downregulated. MiR-124 is a direct target of MALAT1 and inhibits the expression of STAT3 [33], and the expression of MALAT1 has also been reported to be correlated with the expressions of miR-200a-3p and programmed death-ligand 1 (PD-L1) [34]. MALAT1 acts as a sponge for miR-200a-3p, and thus, increases the expression of PD-L1 (a direct target of miR-200a-3p), inhibits apoptosis, and promotes the metastasis of NSCLC cells.
Table 1. Interactions of oncogenic lncRNAs with tumor-suppressive miRNAs in lung cancer

<table>
<thead>
<tr>
<th>lncRNA</th>
<th>miRNA (direct interaction with lncRNA)</th>
<th>The number of miRNA-binding sites</th>
<th>mRNA (target of miRNA)</th>
<th>Effects of lncRNA in cells</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>H19</td>
<td>miR-200a</td>
<td>-</td>
<td>ZEB1, ZEB2</td>
<td>Promoting cell proliferation, migration and invasion</td>
<td>[30]</td>
</tr>
<tr>
<td></td>
<td>miR-196b</td>
<td>7mer-m8</td>
<td>LINC28B</td>
<td>Promoting cell proliferation</td>
<td>[31]</td>
</tr>
<tr>
<td></td>
<td>miR-29b-3p</td>
<td>8mer</td>
<td>STAT3</td>
<td>Promoting cell proliferation and metastasis, inhibiting apoptosis</td>
<td>[32]</td>
</tr>
<tr>
<td>MALAT1</td>
<td>miR-124</td>
<td>7mer-m8</td>
<td>STAT3</td>
<td>Promoting cell proliferation</td>
<td>[33]</td>
</tr>
<tr>
<td></td>
<td>miR-200a-3p</td>
<td>7mer-m8</td>
<td>PD-L1</td>
<td>Promoting metastasis, inhibiting apoptosis</td>
<td>[34]</td>
</tr>
<tr>
<td>DANCR</td>
<td>miR-216a</td>
<td>8mer</td>
<td>EIF4B, JAK2</td>
<td>Promoting cell proliferation</td>
<td>[35]</td>
</tr>
<tr>
<td>LINCO0336</td>
<td>miR-6852</td>
<td>-</td>
<td>CBS</td>
<td>Inhibiting ferroptosis</td>
<td>[36]</td>
</tr>
<tr>
<td>MNX1-AS1</td>
<td>miR-527</td>
<td>-</td>
<td>BRF2</td>
<td>Promoting cell proliferation, migration and invasion</td>
<td>[37]</td>
</tr>
<tr>
<td>LINCO0673</td>
<td>miR-150-5p</td>
<td>-</td>
<td>ZEB1</td>
<td>Promoting cell proliferation, EMT, migration and invasion</td>
<td>[38]</td>
</tr>
<tr>
<td>SNHG4</td>
<td>miR-98-5p</td>
<td>7mer-m8</td>
<td>CDK6, SALL4</td>
<td>Promoting cell proliferation, EMT, migration and invasion</td>
<td>[39]</td>
</tr>
<tr>
<td>LEF1-AS1</td>
<td>miR-489</td>
<td>7mer-m8</td>
<td>SOX4</td>
<td>Promoting cell proliferation and migration, inhibiting apoptosis</td>
<td>[40]</td>
</tr>
<tr>
<td>UCA1</td>
<td>miR-193a-3p</td>
<td>-</td>
<td>ERBB4</td>
<td>Promoting cell proliferation</td>
<td>[41]</td>
</tr>
<tr>
<td>SNHG1</td>
<td>miR-497</td>
<td>7mer-m8</td>
<td>-</td>
<td>Promoting cell proliferation, migration and invasion</td>
<td>[42]</td>
</tr>
<tr>
<td>PTAR</td>
<td>miR-101</td>
<td>-</td>
<td>-</td>
<td>Promoting cell proliferation, migration and invasion</td>
<td>[43]</td>
</tr>
</tbody>
</table>

lncRNA, long non-coding RNA; miRNA, microRNA.

*aThe number of miRNA-binding sites were provided based on the ENCORI database (http://starbase.sysu.edu.cn/index.php) [25].
that function as tumor suppressors in lung cancer growth and development. Several lncRNAs have been identified to repress mRNAs, a target of onco-miRs, which inhibit cancer cell proliferation directly to onco-miRs. Direct binding of tumor-suppressive lncRNAs increases the expressions of tumor-suppressive mRNAs.

Unlike lncRNAs that are upregulated in lung cancer cells and tissues and inhibit the development of lung cancer cells [46,47]. This function of lnc ADAMTS9-AS2 is due to direct interaction with miR-223-3p, which regulates the expression of TGFBR3 [46]. Increased lnc ADAMTS9-AS2 expression in lung cancer cells and tissues downregulated miR-223-3p [46], and as a result, TGFBR3 was upregulated and cancer progression was suppressed [46].

In addition, lncRNAs such as MT1JP, MAGI2-AS3, PLAC2, TINCR, LINC00641, FENDRR (FOXF1 adjacent non-coding developmental regulatory RNA), TRHDE-AS1, and lncRNA-p21 act as tumor suppressors in lung cancer by sponging miRNAs. For example, MJ1JP inhibited the proliferation, invasion, and migration of A549 lung cancer cells, and this inhibition was attributed to Bim upregulation due to the sponging of miR-423-3p [48]. MAGI2-AS3 is downregulated in NSCLC, and its overexpression decreased the proliferative and invasive capacities of NSCLC cells [49]. MAGI2-AS3 sponges miR-23a-3p, and miR-23a-3p directly interacts with PTEN [49]. In another example, low PLAC2 expression predicted poor survival in NSCLC patients, and the overexpression of PLAC2 downregulated miR-21 and upregulated PTEN, a direct target of miR-21 [50]. TINCR acted as a sponge for miR-544a and inhibited the proliferation and invasion of lung cancer cells, but miR-544a directly interacted with FBXW7 and reversed TINCR sponging miR-544a [51]. LINC00641 upregulated the expression of PLSCR4 by sponging miR-424-5p, and as a result, inhibited the proliferation and induced the apoptosis of NSCLC cells [52]. The lncRNA FENDRR upregulated TIMP2 (tissue inhibitor of metalloproteinase 2) by directly binding miR-761, an inhibitor of TIMP2 in NSCLC, and suppressed the aggressiveness of NSCLC cells [53]. TRHDE-AS1 inhibited the proliferation and invasion of lung cancer cells by up-regulating KLF4 (a tumor-suppressor) by inhibiting miR-103, and the overexpression of miR-103 reversed the effect of TRHDE-AS1 [54]. Also, lncRNA-p21 had a direct binding site for miR-17-5p, and binding between the two inhibited NSCLC progression [55].

Another tumor-suppressive lncRNA GAS5 (growth arrest-specific transcript 5) inhibits tumor formation in lung cancer by negatively regulating miR-205 expression, and thus, increasing PTEN expression [56]. In NSCLC, GAS5 inhibited the expression of miR-23a, cell proliferation, and invasion and promoted apoptosis [57]. In addition, GAS5 improved the radiosensitivity of NSCLC cells [58]. Radiotherapy kills cancer cells by exposing them to high-energy radiation [59], and greater radiosensitivity of cancer cells is strongly associated with positive treatment results [60].

LncRNAs such as lnc ADAMTS9-AS2, MT1JP, and GAS5 act as tumor suppressors in lung cancer through lncRNA/miRNA ceRNA networks, which regulate the expressions of well-known lncRNAs such as H19, MALAT1, and DANCR act as oncogenes in lung cancer by interacting with miRNAs. Since these lncRNAs are upregulated in lung cancer cells and tissues, and can be used as and are viewed as potential biomarkers for the early diagnosis of lung cancer. Therapies based on the use of ceRNA networks of oncogenic lncRNAs and miRNAs targeting these genes should be useful for the treatment of lung cancer.

**Tumor-suppressive lncRNAs acting as ceRNAs in lung cancer**

Unlike lncRNAs that are upregulated in lung cancer cells and tissues and function as oncogenes, tumor-suppressive lncRNAs are generally downregulated in lung cancer cells and tissues and bind directly to onco-miRs. Direct binding of tumor-suppressive lncRNAs and onco-miRs upregulates the expressions of tumor-suppressive mRNAs, a target of onco-miRs, which inhibit cancer cell growth and development. Several lncRNAs have been identified that function as tumor suppressors in lung cancer (Table 2). For example, lnc ADAMTS9-AS2 is downregulated in lung cancer cells and tissues and inhibits the development of lung cancer cells [46,47]. This function of lnc ADAMTS9-AS2 is due to direct interaction with miR-223-3p, which regulates the expression of TGFBR3 [46]. Increased lnc ADAMTS9-AS2 expression in lung cancer cells and tissues downregulated miR-223-3p [46], and as a result, TGFBR3 was upregulated and cancer progression was suppressed [46].
Table 2. Interactions of tumor-suppressive lncRNAs with oncogenic miRNAs in lung cancer

<table>
<thead>
<tr>
<th>lncRNA</th>
<th>miRNA (direct interaction with lncRNA)</th>
<th>No. of miRNA-binding sites</th>
<th>mRNA (target of miRNA)</th>
<th>Effects of lncRNA in cells</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAMTS9-AS2</td>
<td>miR-233-3p</td>
<td></td>
<td>TGFBR3</td>
<td>Inhibiting cell proliferation, migration and invasion, promoting apoptosis</td>
<td>[46]</td>
</tr>
<tr>
<td>MT1JP</td>
<td>miR-423-3p</td>
<td></td>
<td>Bim</td>
<td>Inhibiting cell proliferation, migration and invasion</td>
<td>[48]</td>
</tr>
<tr>
<td>MAGI2-AS3</td>
<td>miR-23a-3p</td>
<td>7mer-m8</td>
<td>PTEN</td>
<td>Inhibiting cell proliferation and invasion</td>
<td>[49]</td>
</tr>
<tr>
<td>PLAC2</td>
<td>miR-21</td>
<td></td>
<td>PTEN</td>
<td>Inhibiting migration and invasion</td>
<td>[50]</td>
</tr>
<tr>
<td>TINCR</td>
<td>miR-544a</td>
<td></td>
<td>FBXO7</td>
<td>Inhibiting cell proliferation, migration and invasion</td>
<td>[51]</td>
</tr>
<tr>
<td>LINC00641</td>
<td>miR-424-5p</td>
<td>7mer-m8</td>
<td>PLSCR4</td>
<td>Inhibiting cell proliferation, promoting apoptosis</td>
<td>[52]</td>
</tr>
<tr>
<td>FENDRR</td>
<td>miR-761</td>
<td>8mer</td>
<td>TIMP2</td>
<td>Inhibiting cell proliferation, migration and invasion</td>
<td>[53]</td>
</tr>
<tr>
<td>TRHDE-AS1</td>
<td>miR-103</td>
<td></td>
<td>KLF4</td>
<td>Inhibiting cell proliferation and invasion</td>
<td>[54]</td>
</tr>
<tr>
<td>lncRNA-p21</td>
<td>miR-17-5p</td>
<td></td>
<td>PTEN</td>
<td>Inhibiting cell proliferation and invasion</td>
<td>[55]</td>
</tr>
<tr>
<td>GAS5</td>
<td>miR-205</td>
<td></td>
<td>PTEN</td>
<td>Inhibiting cell proliferation and invasion</td>
<td>[56]</td>
</tr>
<tr>
<td>miR-23a</td>
<td>8mer</td>
<td></td>
<td>PTEN</td>
<td>Inhibiting cell proliferation and invasion</td>
<td>[57]</td>
</tr>
</tbody>
</table>

lncRNA, long non-coding RNA; miRNA, microRNA.

The number of miRNA-binding sites were provided based on the ENCORI database (http://starbase.sysu.edu.cn/index.php) [25].

Tumor suppressors such as PTEN and TIMP2. Like oncogenic lncRNAs, interactions between tumor-suppressive lncRNAs and onco-miRs may aid early diagnosis and provide gene-targeting therapies for lung cancer.

The ceRNA Roles of LncRNAs in Therapeutic Resistance

Non-surgical methods of treating lung cancer include radiation therapy and drug therapy. Representative drugs for the treatment of lung cancer include gefitinib and cisplatin. Gefitinib inhibits epidermal growth factor receptor (EGFR) tyrosine kinase by binding to the enzyme's ATP-binding site [61]. Gefitinib sensitivity studies in NSCLC have shown that mutations in the tyrosine kinase domain of EGFR activate the anti-apoptotic pathway [61]. On the other hand, cisplatin kills the fastest growing cancer cells by interfering with DNA replication [62]. The developments of radiation and drug resistance are major obstacles to successful non-surgical cancer treatment. Accordingly, studies are being actively conducted on genes involved in signaling pathways that improve sensitivity to radiation or drugs, and evidence is accumulating that lncRNA/miRNA networks are involved. For example, it was reported LINCO00483 silences miR-144 in lung adenocarcinoma, and thereby, increases the radiosensitivity of LTEP-A-2 cell lines [63]. Also, FAM201A lncRNA was found to be highly expressed in NSCLC patients resistant to radiation therapy and function as a ceRNA of miR-370 and increase the expressions of EGFR and HIF-1α (hypoxia-inducible factor 1 alpha) [64]. FAM201A knockdown suppressed the expressions of EGFR and HIF-1α and increased the radiosensitivity of NSCLC cells [64]. Furthermore, in NSCLC cells CYTOR (cytoskeleton regulator RNA) lncRNA sponged miR-195 and suppressed radiosensitivity of NSCLC cells in vitro [65].

In a study on drug resistance, overexpression of HOST2 (human ovarian cancer-specific transcript 2) lncRNA inhibited miR-621 and increased gefitinib resistance in NSCLC cells due to the upregulation of SYF2 (a direct target of miR-621) [66]. LINCO00460 was highly expressed in gefitinib-resistant NSCLC cells and tissues and upregulated EGFR through miR-769-5p sponging [67]. Furthermore, EGFR upregulation led to gefitinib resistance [67]. In addition, in cisplatin-resistant NSCLC cells, TATDN1 (Homo sapiens TatD DNase domain containing 1) lncRNA downregulated miR-451, which was overexpressed in these cells, and TATDN1 knockdown improved cisplatin sensitivity [68]. Also, in cisplatin-resistant NSCLC cells, TATDN1 and TRIM66 (a target of miR-451) gene expressions were positively correlated and TRIM66 was overexpressed [68]. In SCLC cells, LINCO00173
sponged miR-218 and induced cisplatin and etoposide (an SCLC therapeutic) resistance [69].

The lncRNA/miRNA ceRNA network regulates the expressions of several genes that act as oncogenes or tumor suppressors in lung cancer. In several studies, changes in gene expressions by lncRNA/miRNA induced drug and radiation resistance in lung cancer cells, which suggests that the ceRNA network has the potential to contribute to the efficient applications of traditional cancer and gene-targeting therapies.

**Conclusion**

Studies on the lncRNA/miRNA ceRNA network in lung cancer are being actively conducted. Direct binding between lncRNA and miRNA influences cancer progression by regulating the expressions of various mRNAs that act as oncogenes or tumor suppressors. In the ceRNA network, the expressions of lncRNA and miRNA are negatively correlated, lncRNA binding to onco-miRs suppresses tumor progression, whereas its binding to tumor-suppressive miRNAs promotes tumor progression. In this review, we summarize the effects of various lncRNAs that function as ceRNAs of miRNAs in lung cancer.

LncRNAs that function as oncogenes in lung cancer through ceRNA networks include H19, MALAT1, DANCR, LINC00336, MNX1-AS1, LINC00673, SNHG4, LEF1-AS1, UCA1, SNHG1, and PTAR. In contrast, lncRNAs that function as ceRNAs of miRNAs and act as tumor suppressors in lung cancer include ADA MTS9-AS2, MT1JP, MAGI2-AS3, PLAC2, TINCR, LINC00641, FENDRR, TRHDE-AS1, lncRNA-p21, and GAS5. Furthermore, these lncRNAs confer radiation and chemical resistance in lung cancer.

For example, LINC00483, FAM201A, and CYTOR induce radioresistance by directly binding miRNAs, and HOST2, LINC00460, TATDN1, and LINC00173 induce drug resistance to gefitinib and cisplatin. As such, many lncRNAs contribute to the development of lung cancer in various ways by direct binding miRNAs.

In this review, we summarize the lncRNA/miRNA ceRNA networks that impact lung cancer identified to date, and provide insight into the effects of RNAs not translated into proteins and of the various signaling pathways that act on lung cancer through downstream factors. The lncRNA/miRNA ceRNA network offers a means of discovering biomarkers that enable the early diagnosis of lung cancer and provide guidance regarding gene-specific treatments. In addition, the abilities of lncRNA and miRNA interactions to affect radiation and drug resistance suggests they can be targeted in treatment of resistant patients. We believe improved understanding of lncRNA and miRNA interactions is likely to lead to future developments in the lung cancer treatment field.

**ORCID**

Danbi Seo: https://orcid.org/0000-0002-9996-922X
Dain Kim: https://orcid.org/0000-0003-2512-2365
Yeonsoo Chae: https://orcid.org/0000-0003-2648-8238
Wanyeon Kim: https://orcid.org/0000-0002-7611-5248

**Authors’ Contribution**

Conceptualization: WK. Data curation: DS, DK, WK. Formal analysis: DS, DK, WK. Methodology: DS, YC, WK. Writing - original draft: DS, DK, YC, WK.

**Conflicts of Interest**

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

**References**

10. Saliminejad K, Khorram Khoshid HR, Soleymani Fard S, Ghaf-fari SH. An overview of microRNAs: biology, functions, thera-


Transcriptome analysis of iBET-151, a BET inhibitor alone and in combination with paclitaxel in gastric cancer cells

Sun Kyoung Kang1,2, Hyun Joo Bae1,3, Woo Sun Kwon1, Jingmin Che1,2, Tae Soo Kim1, Hyun Cheol Chung1,3,4, Sun Young Rha1,3,4*

1 Songdang Institute for Cancer Research, Yonsei University College of Medicine, Seoul 03722, Korea
2 MD Biolab Co., Ltd., Seoul 02455, Korea
3 Brain Korea 21 PLUS Project for Medical Science, Yonsei University College of Medicine, Seoul 03722, Korea
4 Division of Medical Oncology, Department of Internal Medicine, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul 03722, Korea

BET inhibitor, as an epigenetic regulator inhibitor, reduces the expression of oncogenes such as Myc and Bcl-2, which affects cancer growth and development. However, it has modest activity because of the narrow therapeutic index. Therefore, combination therapy is necessary to increase the anti-tumor effect. Paclitaxel, an anti-mitotic inhibitor, is used as second-line therapy for gastric cancer (GC) as a monotherapy or combination. In this study, we performed RNA sequencing of GC cells treated with iBET-151 and/or paclitaxel to identify the differentially expressed genes associated with possible mechanisms of synergistic effect. We also performed Gene Ontology enrichment and Kyoto Encyclopedia of Genes and Genomes pathway analyses to determine the most enriched terms and pathways of upregulated and downregulated genes. We found 460 genes in which iBET-151 and paclitaxel combination treatment changed more than single-treatment or no-treatment. Thus, additional functional studies are needed, but our results provide the first evidence of the synergistic effect between iBET-151 and paclitaxel in regulating the transcriptome of GC cells.

Keywords: BET inhibitor, combination, differentially expressed genes, gastric cancer, paclitaxel, transcriptome

Introduction

Gastric cancer (GC) is the second leading cause of cancer-related mortalities worldwide, especially in Asian countries. In South Korea, the incidence of GC was the highest in 2017 and GC ranked fourth among the causes of cancer-related deaths [1-4]. Despite advances in therapeutics, recurrence occurs in about half of the patients, and the prognosis of recurrent and metastatic GC is very poor [5-7]. Thus, new therapeutic strategies involving for tumor-specific molecular targets are critically needed to improve the outcome of recurrent and metastatic GC patients. Paclitaxel, which is one of the most commonly used standard chemotherapeutic drugs, binds to β-tubulin and stabilizes microtubules, resulting in G2/M arrest and cell death. Various targeted anticancer agents were tested in combination with paclitaxel as second-line therapy for advanced GC [8-11]. However, most combinations, except that with ramucirumab, failed to show improved efficacy [12].
Moreover, mechanisms underlying paclitaxel activity, except mitosis inhibition, have not yet been studied well. Elucidating these mechanisms is important for the selection of appropriate combination partners of paclitaxel.

Epigenetic dysregulation is a common feature of cancer, including GC. Eventually, it contributes to GC development and progression [13]. In our previous study, iBET-151, an inhibitor of the bromodomain and extraterminal (BET) protein, which is an epigenetic regulator, was found to function effectively in GC cells (unpublished data). A recent study reported that lowering the expression of B-cell lymphoma (BCL)-2 and BCL-xL, the targets of the BET protein, induces sensitivity to paclitaxel [14]. We confirmed that paclitaxel and iBET-151 exerted synergistic effects in GC cells by investigating cell viability and migration and the cell cycle (unpublished data). However, to determine the detailed synergistic mechanisms of paclitaxel and iBET-151, it is necessary to identify the overall changes in the transcriptome and associated pathways.

RNA sequencing (RNA-seq), via next-generation sequencing, is very useful for differential expression analysis involving specific conditions, such as ‘drug treatment versus no-treatment’ Therefore, researchers can comprehensively report the gene expression changes related to drug treatment by performing RNA-based genome-wide profiling [15].

In this study, we performed RNA-seq to analyze the differentially expressed genes (DEGs) observed following treatment with paclitaxel, iBET-151, and their combination in GC cells. We also identified the biological and functional pathways associated with DEGs using the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases to determine the possible action mechanisms of BET monotherapy and BET-paclitaxel combination treatment.

Methods

Cell line and cell culture
AGS GC cell line (CRL-1739) was obtained from the American Type Culture Collection (ATCC, Manassas, VA, USA). AGS cells were cultured in Eagle’s Minimum Essential Medium supplemented with 10% fetal bovine serum (Lonza, Basel, Switzerland) and 1 × penicillin/streptomycin (Lonza). All experiments were performed after AGS cells reached the exponential growth phase.

Drugs
A small-molecule inhibitor of the BET family, iBET-151, was purchased from Selleckchem (Houston, TX, USA) and prepared as a 10-mM stock solution in DMSO (Sigma-Aldrich, St. Louis, MO, USA). Paclitaxel, an anti-mitotic drug, was purchased from Sigma-Aldrich.

RNA-seq data analysis
For differential expression analysis by RNA-seq, total RNA was isolated from iBET-151- or/and paclitaxel-treated or untreated AGS cells using the RNeasy Mini Kit (Qiagen, Hilden, Germany), according to the manufacturer’s instructions. RNA quantity and integrity were assessed using both Nanodrop and Bioanalyzer, according to the manufacturer’s instructions. Library preparation and RNA-seq was performed using the Illumina TruSeq RNA Library Preparation Kit (Illumina, San Diego, CA, USA), according to the manufacturer’s protocol. Samples were sequenced on the Illumina NovaSeq 6000, 150 bp paired-end reads, to a minimum depth of 120 million reads per sample. Sequenced reads were aligned to the hg19 genome assembly using TopHat2. Differential expression analysis was processed following the Tuxedo protocol. DEGs were defined based on the cutoff values of p ≤ 0.05, false discovery rate < 0.05, and fold change (FC) > 1.5. The reference annotation based on Ensembl release 69 (ftp://ftp.ensembl.org/pub/release-69/gtf/macaca_mulatta).

Functional enrichment analysis
GO annotation enrichment analysis was performed using The Database for Annotation, Visualization, and Integrated Discovery (DAVID) (v6.8), an online tool applied for functional annotation analysis and Gene Set Enrichment Analysis using MSigDB. GO and KEGG pathway analyses provide a comprehensive set of functional annotation tools for identifying characteristic biological attributes of high-throughput genome or transcriptome data.

Western blot analysis
To study the effect of drugs on protein expression, tissues and cells were lysed in RIPA buffer (50 mM Tris [pH 8.0], 150 mM NaCl, 1% NP-40, 0.1% sodium dodecyl sulfate, 1 mM EDTA, 10% glycerol) containing protease and phosphatase inhibitors (Roche Diagnostics, Laval, QC, Canada). The cell lysates were subjected to sodium dodecyl sulfate polyacrylamide gel electrophoresis and western blotting. Membranes were then developed using an enhanced chemiluminescence (ECL) reagent (GE Healthcare, Piscataway, NJ, USA). α-tubulin was used as a loading control. Antibodies used for western blotting included anti-epidermal growth factor receptor (EGFR), anti-fibroblast growth factor receptor (FGFR)-3, and anti-insulin-like growth factor-1 receptor (IGF-1R; Cell Signaling Technology, Beverly, MA, USA); anti-human epidermal growth factor receptor (HER)-2, anti–HER-3, and anti-MET (Santa Cruz Biotechnology, Santa Cruz, CA, USA); anti-α-tubulin.
and horseradish peroxidase-conjugated secondary anti-rabbit and anti-mouse (Jackson ImmunoResearch Laboratories, West Grove, PA, USA) antibodies. Membranes were then developed using the ECL reagent (GE Healthcare), and α-tubulin was used as a loading control. Quantification was performed with a Molecular Imager ChemiDoc XRS+ Imaging System (Bio-Rad, Hercules, CA, USA).

Statistical analysis
Graphing was performed using GraphPad Prism software version 8 (Graph Pad Software Inc., San Diego, CA, USA). All statistical analyses were performed using SPSS version 25 (IBM Corp., Armonk, NY, USA). All the significance levels were set at p ≤ 0.05, and all p-values were two-sided.

Results

DEGs induced by iBET-151, paclitaxel, and combination treatments
In our previous study, we confirmed that iBET-151-paclitaxel combination treatment exerted a synergistic effect on cell viability and invasion in several cell lines (unpublished data). To understand the effects of iBET-151 and paclitaxel on the whole transcriptome better, we performed RNA-seq of AGS cells that were most affected by the synergistic effects of iBET-151 and paclitaxel to compare the global gene expression profiles of untreated and iBET-151-, paclitaxel-, and iBET-151–paclitaxel combination-treated cells.

About 16,000 genes were selected after the preprocessing step, during which genes that were not detected in both treated and untreated AGS cells were excluded. Based on our DEG-specific criteria, we observed that iBET-151 upregulated 239 genes, whereas paclitaxel upregulated 203 genes (Fig. 1A). Upregulation of 270 genes by combination treatment was mainly due to the effect of iBET-151 treatment (Fig. 1A). Of the genes upregulated by iBET-151 alone, 61% (167/270) were upregulated by the combination of iBET-151 and paclitaxel. Thus, these genes were shared between the combination- and iBET-151-treated cells.

Then, we observed that iBET-151 downregulated 395 genes, whereas paclitaxel downregulated 57 genes (Fig. 1B). In the combination-treated cells, 358 genes were downregulated; among them, 73% (263/358) overlapped with the downregulated genes in the iBET-151-treated cells (Fig. 1B). However, the genes regulated by paclitaxel alone did not overlap with those regulated by iBET-151 alone or the iBET-151-paclitaxel combination. The top 10 genes upregulated and downregulated by iBET-151, paclitaxel, and combination treatments are shown in Tables 1 and 2, respectively.

GO functional enrichment analysis
To determine the biological effects of gene expression changes, we performed GO analysis with the selected DEGs using DAVID v6.8. Volcano plots were used for the concurrent demonstration of both FC and p-value of DEGs in mono- or combination-treated cells (Figs. 2A, 3A, and 4A).

Interestingly, GO analysis showed that upregulated genes in iBET-151-treated cells were most enriched in the chromatin silencing, nucleosome assembly pathways, and DNA packing com-

---

**Fig. 1.** Overview of genes regulated by iBET-151, paclitaxel, and combination treatments. (A) Venn diagram analysis of the number of upregulated genes in treated versus untreated cells is shown. (B) Venn diagram analysis of the number of downregulated genes in treated versus untreated cells is shown.

https://doi.org/10.5808/GI.2020.18.4.e37
Table 1. Top 10 upregulated genes in AGS cells treated by iBET-151 and paclitaxel

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Fold change</th>
<th>p-value</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>iBET-151</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFR3B</td>
<td>5.880</td>
<td>0.0007</td>
<td>Membrane-anchoring component</td>
</tr>
<tr>
<td>GPR157</td>
<td>5.380</td>
<td>0.0415</td>
<td>Cell surface receptor signaling pathway, G-protein coupled receptor signaling pathway</td>
</tr>
<tr>
<td>STK31</td>
<td>4.980</td>
<td>0.0272</td>
<td>RNA catabolic process</td>
</tr>
<tr>
<td>IQFN1</td>
<td>4.440</td>
<td>0.0409</td>
<td>Cytokine-mediated signaling pathway, response to lipopolysaccharide, regulation of synapse organization</td>
</tr>
<tr>
<td>IL10RA</td>
<td>4.330</td>
<td>0.0111</td>
<td>Cytokine-mediated signaling pathway, response to lipopolysaccharide, regulation of synapse organization</td>
</tr>
<tr>
<td>CLUL1</td>
<td>4.180</td>
<td>0.0228</td>
<td>Cell death</td>
</tr>
<tr>
<td>SPOCK2</td>
<td>4.140</td>
<td>0.0484</td>
<td>Positive regulation of cell-substrate adhesion</td>
</tr>
<tr>
<td>TEX14</td>
<td>4.080</td>
<td>0.0392</td>
<td>Mitotic spindle assembly checkpoint</td>
</tr>
<tr>
<td>HIST2H4A</td>
<td>3.980</td>
<td>0.0465</td>
<td>Chromatin silencing at rDNA, negative regulation of gene expression (epigenetic)</td>
</tr>
<tr>
<td>CLU</td>
<td>3.920</td>
<td>0.0002</td>
<td>Release of cytochrome c from mitochondria, intrinsic apoptotic signaling pathway</td>
</tr>
<tr>
<td>Paclitaxel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMP7</td>
<td>86.2229</td>
<td>0.0465</td>
<td>Secreted ligand of the TGF-beta</td>
</tr>
<tr>
<td>KCNQ5</td>
<td>43.4113</td>
<td>0.04195</td>
<td>Regulating potassium channel</td>
</tr>
<tr>
<td>AMPD3</td>
<td>34.5353</td>
<td>0.00245</td>
<td>Catalyzes the deamination of AMP to IMP in red cells and plays an important role in the purine nucleotide cycle.</td>
</tr>
<tr>
<td>TSPAN32</td>
<td>29.6508</td>
<td>0.00115</td>
<td>Functional roles in cell motility, membrane fusion, proliferation, and adaptive immunity</td>
</tr>
<tr>
<td>MAGED1</td>
<td>27.0958</td>
<td>0.0111</td>
<td>Inhibits cell cycle progression, and facilitates NGFR-mediated apoptosis</td>
</tr>
<tr>
<td>MYO15A</td>
<td>25.4572</td>
<td>0.0014</td>
<td>Actin-based motor molecules with ATPase activity</td>
</tr>
<tr>
<td>ZDHHC1</td>
<td>20.3930</td>
<td>0.04455</td>
<td>Innate immune response, related to phencyclidine abuse</td>
</tr>
<tr>
<td>CATSPERE</td>
<td>18.0009</td>
<td>0.01695</td>
<td>Calcium channel pore-forming proteins</td>
</tr>
<tr>
<td>ADCY10P1</td>
<td>12.9560</td>
<td>0.0478</td>
<td>Sphingosine metabolic process</td>
</tr>
<tr>
<td>MYLK</td>
<td>11.4716</td>
<td>0.0186</td>
<td>Activated by the binding of calcium-calmodulin, interaction with actin filaments to produce contractile activity</td>
</tr>
<tr>
<td>Combination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFR3B</td>
<td>55.7152</td>
<td>0.0003</td>
<td>Acts as the membrane-anchoring component</td>
</tr>
<tr>
<td>MMP1</td>
<td>42.8137</td>
<td>0.0024</td>
<td>Calcium ion binding and metallopeptidase activity</td>
</tr>
<tr>
<td>HIST1H2BJ</td>
<td>32.0000</td>
<td>0.0362</td>
<td>Wrap and compact DNA into chromatin, limiting DNA accessibility to the cellular machineries which require DNA as a template</td>
</tr>
<tr>
<td>SELENOM</td>
<td>22.0087</td>
<td>0.02955</td>
<td>Function as a thiol-disulfide oxidoreductase that participates in disulfide bond formation</td>
</tr>
<tr>
<td>HIST2H4A</td>
<td>20.2521</td>
<td>0.0042</td>
<td>Wrap and compact DNA into chromatin, limiting DNA accessibility to the cellular machineries which require DNA as a template</td>
</tr>
<tr>
<td>TEX14</td>
<td>19.0273</td>
<td>0.035</td>
<td>Kinetochore-microtubule attachment during mitosis</td>
</tr>
<tr>
<td>HIST1H2BC</td>
<td>17.5087</td>
<td>0.04265</td>
<td>Wrap and compact DNA into chromatin, limiting DNA accessibility to the cellular machineries which require DNA as a template</td>
</tr>
<tr>
<td>HIST1H2BG</td>
<td>17.2677</td>
<td>0.0098</td>
<td>Wrap and compact DNA into chromatin, limiting DNA accessibility to the cellular machineries which require DNA as a template</td>
</tr>
</tbody>
</table>

plex (Fig. 2B, Supplementary Fig. 1A). This was because iBET-151 suppressed the epigenetic regulatory function of the BET family and affected chromatin remodeling. The genes downregulated by iBET-151 were enriched in angiogenesis; cell migration; RAS protein signal transduction; positive regulation of extracellular signal-regulated kinase (ERK)-1 and ERK cascade; positive regulation of transcription, DNA-templated; and negative regulation of apoptotic process (Fig. 2C, Supplementary Fig. 1B).

In paclitaxel-treated cells, the upregulated genes were enriched in the apoptotic signaling pathway; release of cytochrome c from mitochondria; regulation of apoptotic process; and mitochondrial electron transport, cytochrome c to oxygen (Fig. 3B, Supplementary Fig. 1C). Genes downregulated by paclitaxel were enriched in sister chromatid segregation, mitotic nuclear division, and centrosome duplication (Fig. 3C, Supplementary Fig. 1D). According to a recent study [16], paclitaxel inhibited mitosis by binding to microtubules and regulating mitosis-related pathways, including mitotic nuclear division, chromosome segregation, and G2/M transition of the mitotic cell cycle, in lung cancer cells. The results of our previous study on lung cancer cells showed that paclitaxel upregulated mitosis-related pathways, including mitotic nuclear division and chromosome segregation (unpublished data).

GO analysis showed a significant increase in the upregulation of the same genes involved in the aforementioned pathways in com-
<table>
<thead>
<tr>
<th>Gene name</th>
<th>Fold change</th>
<th>p-value</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADGRF1</td>
<td>0.006</td>
<td>0.0116</td>
<td>Cell surface receptor signaling pathway, G-protein coupled receptor signaling pathway</td>
</tr>
<tr>
<td>PMEPA1</td>
<td>0.007</td>
<td>0.0362</td>
<td>Cell proliferation, differentiation, apoptosis, motility, extracellular matrix production and immunosuppression</td>
</tr>
<tr>
<td>PSG1</td>
<td>0.013</td>
<td>0.015</td>
<td>Defense response</td>
</tr>
<tr>
<td>NTM</td>
<td>0.017</td>
<td>0.0002</td>
<td>Neural cell adhesion molecule, cell adhesion</td>
</tr>
<tr>
<td>CACNA1H</td>
<td>0.019</td>
<td>0.02405</td>
<td>Regulation of ion transmembrane transport, calcium ion transmembrane transport</td>
</tr>
<tr>
<td>SLAMF7</td>
<td>0.021</td>
<td>0.00945</td>
<td>Adaptive immune response, cell adhesion, natural killer cell activation</td>
</tr>
<tr>
<td>BMP4</td>
<td>0.025</td>
<td>0.00075</td>
<td>Activation of MAPKK activity</td>
</tr>
<tr>
<td>NRP2</td>
<td>0.044</td>
<td>0.00255</td>
<td>Angiogenesis, positive regulation of endothelial cell proliferation</td>
</tr>
<tr>
<td>TCF4</td>
<td>0.048</td>
<td>0.00005</td>
<td>Activate transcription</td>
</tr>
<tr>
<td>SRC</td>
<td>0.136</td>
<td>0.00005</td>
<td>Epidermal growth factor receptor signaling pathway, signal transduction</td>
</tr>
<tr>
<td>DOCK10</td>
<td>0.008</td>
<td>0.0147</td>
<td>Regulation of CDC42 activity</td>
</tr>
<tr>
<td>ADAP2</td>
<td>0.042</td>
<td>0.0253</td>
<td>GTPase activator activity</td>
</tr>
<tr>
<td>SYT1</td>
<td>0.067</td>
<td>0.00405</td>
<td>Calcium ion binding and transporter activity</td>
</tr>
<tr>
<td>NREP</td>
<td>0.139</td>
<td>0.0016</td>
<td>Neural function</td>
</tr>
<tr>
<td>MEGF11</td>
<td>0.14</td>
<td>0.0136</td>
<td>Retina layer formation, motopic cell-cell adhesion</td>
</tr>
<tr>
<td>FAM19A2</td>
<td>0.149</td>
<td>0.04425</td>
<td>Neuronal survival and neurobiological functions</td>
</tr>
<tr>
<td>GCNT2</td>
<td>0.168</td>
<td>0.02685</td>
<td>Acetylglucosaminyltransferase activity and N-acetyllactosaminide beta-1,6-N-acetyllactosaminyltransferase activity</td>
</tr>
<tr>
<td>TYH1</td>
<td>0.187</td>
<td>0.02065</td>
<td>Cell adhesion</td>
</tr>
<tr>
<td>RMDN2</td>
<td>0.308</td>
<td>0.0235</td>
<td>Interact with microtubules</td>
</tr>
<tr>
<td>NCKAP5</td>
<td>0.31</td>
<td>0.0394</td>
<td>Microtubule depolymerization</td>
</tr>
<tr>
<td>GBP1</td>
<td>0.003</td>
<td>0.04585</td>
<td>GTPase activity</td>
</tr>
<tr>
<td>INHBE</td>
<td>0.011</td>
<td>0.0005</td>
<td>Insulin secretion, nerve cell survival, cytokine-cytokine receptor interaction</td>
</tr>
<tr>
<td>NUPR1</td>
<td>0.012</td>
<td>0.03795</td>
<td>Positively regulates cell cycle progression</td>
</tr>
<tr>
<td>GPX5</td>
<td>0.024</td>
<td>0.0382</td>
<td>Protects cells and enzymes from oxidative damage</td>
</tr>
<tr>
<td>DOCK10</td>
<td>0.024</td>
<td>0.01775</td>
<td>Regulation of CDC42 activity</td>
</tr>
<tr>
<td>RP1</td>
<td>0.028</td>
<td>0.03635</td>
<td>Microtubule-associated protein regulating the stability and length of the microtubule-based axoneme of photoreceptors</td>
</tr>
<tr>
<td>PMEPA1</td>
<td>0.029</td>
<td>0.0249</td>
<td>A negative regulator of TGF-beta signaling</td>
</tr>
<tr>
<td>GCNT3</td>
<td>0.033</td>
<td>0.0025</td>
<td>Synthesize all known mucin beta 6 N-acetyllactosaminides</td>
</tr>
<tr>
<td>SLAMF7</td>
<td>0.034</td>
<td>0.006</td>
<td>Regulation and interconnection of both innate and adaptive immune response</td>
</tr>
<tr>
<td>PHF21B</td>
<td>0.034</td>
<td>0.0201</td>
<td>Metal ion binding</td>
</tr>
</tbody>
</table>

bination-treated cells. Moreover, an increase in the upregulation of genes involved in reactive oxygen species metabolic process and nucleotide-excision repair, DNA damage recognition was observed after combination treatment (Fig. 4B, Supplementary Fig. 1E). We observed that downregulated genes in the combination- and iBET-151-treated cells were enriched in overlapping GO terms, including angiogenesis, positive regulation of RAS protein signaling, cell migration, MAP kinase activity, and positive regulation of signaling; however, a significant decrease in gene expression was observed between the combination- and mono-treated cells (Fig. 4C, Supplementary Fig. 1F).

**Significantly altered genes in combination-treated cells, compared with that in mono-treated cells**

To understand the effects of combination treatment better, we identified genes that were more significantly upregulated or downregulated in combination-treated cells than in mono-treated cells. We first identified the genes with FC > 1.5 and p < 0.05 in combination-treated cells, compared with in control cells. We then identified the DEGs with FC > 1.5 in the respective mono-treated cells.

We identified 460 DEGs, including 238 upregulated and 222
downregulated DEGs, in AGS cells based on the above-mentioned cutoff values. As shown in Fig. 5A, the heatmap of hierarchical clustering of 460 DEGs demonstrated clear discriminating patterns. Then, GO analysis demonstrated that combination treatment upregulated genes enriched in nucleosome assembly, mitochondrial electron transport, and microtubule-based process (Fig. 5B). Subsequent GO analysis after combination treatment showed downregulation of genes involved in mitotic sister chromatid segregation, positive regulation of transcription, positive regulation of GTPase activity, and mitotic nuclear division (Fig. 5C).

Most of these results after combination treatment were similar to those obtained after iBET-151 or paclitaxel mono-treatment. However, combination treatment induced more changes in the genes than those in case of either of the mono-treatments.

Expression of BET protein target genes and major receptor tyrosine kinase

Next, we determined the mRNA expression levels of BET protein target genes. We observed that iBET-151 reduced the mRNA expression of the oncogenes, MYC and KRAS, and the anti-apoptotic gene, BCL-2. However, no change was observed in the expression of these genes on paclitaxel treatment, as expected (Fig. 6A).

In our previous study, we found that iBET-151 induced G1 cell cycle arrest (unpublished data). Consistent with this result, the expression of cyclin dependent kinase-4/6 and E2F transcription factor-2 involved in G1–S phase transition was reduced. To elucidate the underlying biology, we observed the changes in major receptor tyrosine kinase (RTK) expression. We confirmed that iBET-151 reduced the expression of RTK mRNA and proteins, including EGFR, ERBB3, MET, and IGF-1R, thereby regulating cancer cell growth and survival. Additionally, RTK expression was further reduced on iBET-151-paclitaxel combination treatment. However, in the FGFR family, FGFR-3/4 was significantly increased on iBET-151 and combination treatments (Fig. 6B and 6C).

Discussion

In this study, we performed RNA-seq to identify the transcriptome changes in transcriptome by iBET-151-paclitaxel
**Fig. 3.** DEGs induced by paclitaxel treatment. (A) Volcano plot of DEGs induced in paclitaxel-treated versus untreated cells is shown. (B) GO biological processes/KEGG pathways upregulated in response to paclitaxel treatment are shown. (C) GO biological processes/KEGG pathways downregulated in response to paclitaxel treatment are shown. DEGs, differentially expressed genes; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

**Fig. 4.** DEGs induced by iBET-151–paclitaxel combination treatment. (A) Volcano plot of DEGs induced in combination-treated versus untreated cells is shown. (B) GO biological processes/KEGG pathways upregulated in response to iBET-151–paclitaxel combination treatment are shown. (C) GO biological processes/KEGG pathways downregulated in response to iBET-151–paclitaxel combination treatment are shown. DEGs, differentially expressed genes; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.
changes induced by mono- and combined treatments with iBET-151 and paclitaxel in the AGS GC cell line. First, we confirmed that the expression of MYC and BCL-2, which is known as the target of BRD-4, was decreased on iBET-151 treatment, as expected. Interestingly, we observed that KRAS, an oncogene that plays a key role in the development and progression of various carcinomas, was downregulated in iBET-151–treated cells. Even though KRAS has been known as a significant driver in various cancers, it has long been considered undruggable. Interestingly, we confirmed that KRAS mRNA expression was reduced on iBET-151 treatment and further decreased on iBET-151–paclitaxel combination treatment. This result suggested that iBET-151 could be developed as an indirect KRAS inhibitor for potential therapeutic effects in GC patients.

Additionally, we confirmed that expression of signal transduction-related genes, such as epidermal growth factor, vascular endothelial growth factor, IGF, and SRC, was also significantly decreased. On the other hand, expression of genes related to chromatin silencing and nucleosome assembly was increased on iBET-151 treatment. This was because the BET family proteins, the targets of iBET-151, are epigenetic regulators. Particularly, expression of sirtuin (SIRT)-4 involved in histone deacetylation was significantly increased. SIRT-4 is a tumor suppressor that is reduced in various carcinomas, and reduced SIRT-4 protein expression is associated with poor prognosis. It has also been reported that SIRT-4 interferes with glutamine metabolism and inhibits cell proliferation and migration [17]. As mentioned in our previous study, the metabolism pathway is enhanced in not only the iBET-151–sensitive group, but also various carcinomas (unpublished data). Therefore, we expected that increasing SIRT-4–inhibiting glucose pyruvate dehydrogenase using iBET-151 would be a feasible therapeutic strategy for these cancers.

Paclitaxel is a drug that binds to tubulin and stabilizes microtubules to suppress mitosis. However, our RNA-seq results revealed that paclitaxel also regulates expression of genes involved in cell division in GC. In this study, expression of genes related to sister chromatid segregation, mitotic nuclear division, and centrosome duplication was decreased in paclitaxel-treated GC cells. Additionally, expression of genes involved in cytochrome c release in mitochondria and apoptotic process was increased. This suggested that

Fig. 5. DEGs induced by iBET-151–paclitaxel combination treatment. (A) Volcano plot of DEGs induced in combination-treated versus untreated cells is shown. (B) GO biological processes/KEGG pathways upregulated in response to iBET-151–paclitaxel combination treatment are shown. (C) GO biological processes/KEGG pathways downregulated in response to iBET-151–paclitaxel combination treatment are shown. DEGs, differentially expressed genes; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.
Fig. 6. Expression of BET protein target genes and major RTK. (A) mRNA expression of BET family protein target and tumor growth-related genes is shown. (B) GO biological processes upregulated in response to combination treatment are shown. (C) Validation of DEGs was performed by western blotting. BET, bromodomain and extraterminal; RTK, receptor tyrosine kinase; GO, Gene Ontology; DEGs, differentially expressed genes; FPKM, fragments per kilobase of exon model per million reads mapped; EGFR, epidermal growth factor receptor; IGF1R, insulin-like growth factor receptor 1; IGF2R, insulin-like growth factor receptor 2; FGFR, fibroblast growth factor receptor.
paclitaxel not only inhibited mitosis by binding to tubulin but also regulated mitosis inhibition and apoptosis in cell death by regulating related gene expression. Therefore, our study provides novel findings that may help in identifying appropriate combination drugs with paclitaxel for GC treatment via determination of gene expression changes induced by paclitaxel.

Additionally, we found that iBET-151 and paclitaxel separately reduced the expression of RTK, which is involved in key signaling pathways (RAS/mitogen-activated protein kinase and phosphoinositide 3-kinase/AKT pathways) that regulate cancer cell proliferation.

Trastuzumab is approved for use as HER-2-targeted therapy for various cancers, including GC, to improve survival of patients with HER-2 amplification. However, in many cases, acquired resistance is induced by re-activation due to interaction with other RTKs, and this is related to poor survival. We found that iBET-151 reduced the expression of other RTKs, such as EGFR, HER-3, and MET, which are induce signaling activation by heterodimeric binding to HER-2. This expression is further decreased on iBET-151-paclitaxel combination treatment. Therefore, the combination of iBET-151 and paclitaxel can overcome resistance induced by HER-2-targeted treatment in GC.

JQ1, a BET inhibitor, has been recently reported to induce resistance by increasing FGFR protein expression in ovarian cancer [18] and uveal melanoma [19], but the underlying mechanism remains unclear. In our results, FGFR-3 protein expression was increased in GC, and RNA-seq showed that FGFR-3 and FGFR-4 mRNA expression was increased. Further research is needed to elucidate FGFR-3 regulation, but this makes it possible to find out that FGFR3 is regulated at the transcription level by the BET inhibitor. Therefore, our results show that not only JQ1 but also iBET-151 increased FGFR-3 expression, and in particular iBET-151 regulates FGFR3 at the mRNA level. Therefore, our results add evidence for combination therapy with BET inhibitor and FGFR inhibitor.

In our previous study, iBET-151 and paclitaxel exerted synergistic effects on GC cell proliferation and migration (unpublished data). Using RNA-seq analysis, we found that combination-treated cells showed higher changes in gene expression in the above-mentioned pathways than untreated and mono-treated cells. AGS cells showed the greatest reduction in angiogenesis-related expression of vascular endothelial growth factor-A, matrix metalloproteinase-14, and X-box-binding protein (XBP)-1 in response to combination treatment, compared with either iBET-151 or paclitaxel single treatment. Particularly, XBP-1, a transcription factor, is known to promote not only angiogenesis, but also tumorigenesis, proliferation, and drug resistance.

In conclusion, we analyzed the biological functions induced by iBET-151–paclitaxel combination treatment. The results elucidated the biological mechanism underlying the synergistic effects of iBET-151 and paclitaxel. More research is needed in the future to determine the exact mechanisms underlying the action of these compounds, but we expect that the combination therapy of iBET-151 and paclitaxel will be a promising therapeutic strategy for GC.

**ORCID**

Sun Young Rha: https://orcid.org/0000-0002-2512-4531

**Authors’ Contribution**

Conceptualization: SYR, SKK, HJB, WSK, HCC. Data curation: SYR, SKK, WSK. Formal analysis: SYR, SKK, WSK. Funding acquisition: SYR. Methodology: SYR, SKK, HJB, WSK, TSK. Writing - original draft: SYR, SKK. Writing - review & editing: SYR, SKK, HJB, WSK, JC.

**Conflicts of Interest**

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

**Acknowledgments**

We thank all members of the CMRC lab for discussion. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2020R1A2B5B02001452).

**Supplementary Materials**

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

**References**

Introducing the fundamental concept of circadian rhythms, which are essential biological processes existing throughout the human body. They are characterized by a cycle of approximately 24 hours and are crucial for maintaining homeostasis and synchronizing various physiological functions. The suprachiasmatic nucleus (SCN) serves as the central pacemaker, while a variety of environmental stimuli, such as light-dark cycles, also play significant roles in regulating circadian rhythms. The sleep-wake cycle is one of the most evident circadian rhythms, with individuals showing a preferred timing for sleeping and waking, known as chronotype. This preference can range from morning types (with an advanced sleep phase) to evening types (with a delayed sleep phase). Chronotype is known to be determined by a complex interplay of factors, including age, sex, and environmental conditions such as light exposure. The current study investigates genetic associations and gene-gene interactions of clock genes for chronotype in a healthy Korean population.

**Original article**

Mira Park¹, Soon Ae Kim², Jieun Shin³, Eun-Jeong Joo⁴,⁵*

¹Department of Preventive Medicine, Eulji University School of Medicine, Daejeon 34824, Korea
²Department of Pharmacology, Eulji University School of Medicine, Daejeon 34824, Korea
³Department of Liberal Arts, Woosuk University, Wanju 55338, Korea
⁴Department of Neuropsychiatry, Eulji University School of Medicine, Daejeon 34824, Korea
⁵Department of Psychiatry, Nowon Eulji Medical Center, Eulji University, Seoul 01830, Korea

Chronotype is an important moderator of psychiatric illnesses, which seems to be controlled in part by genetic factors. Clock genes, such as *BHLHB2*, *CLOCK*, *CSNK1E*, *NR1D1*, *PER1*, *PER2*, *PER3*, and *TIMELESS*, are crucial for regulating chronotype. This study explores genetic associations and gene-gene interactions of the clock genes *BHLHB2*, *CLOCK*, *CSNK1E*, *NR1D1*, *PER1*, *PER2*, *PER3*, and *TIMELESS* for chronotype in 1,293 healthy Korean individuals. Statistical methods, including regression analysis and quantitative multifactor dimensionality reduction (QMDR), were employed to identify associations between single nucleotide polymorphism (SNP) and chronotype. The QMDR analysis identified *NR1D1* rs2314339 and *TIMELESS* rs4630333 as the best SNP pairs among two-locus interaction models associated with chronotype (cross-validation consistency [CVC] = 8/10, p = 0.041). For the three-locus interaction model, the SNP combination of *NR1D1* rs2314339, *TIMELESS* rs4630333, and *PER3* rs228669 showed the best results (CVC = 4/10, p < 0.001). However, because the mean differences between genotype combinations were minor, the clinical roles of clock gene interactions are unlikely to be critical.

**Keywords:** chronotype, circadian rhythm, clock genes, gene-gene interaction, quantitative multifactor dimensionality reduction

**Introduction**

Many biological processes have circadian rhythms with a cycle of approximately 24 hours. Such circadian rhythms appear at the cellular, molecular, tissue, and organ levels in humans. Both the suprachiasmatic nucleus, which is the internal master clock, and various environmental stimuli (e.g., the light-dark cycle) may play roles together in regulation of circadian rhythm. The sleep-wake cycle is one of the most distinct circadian rhythms. Humans have a preferred timing of sleeping and waking, the so-called chronotype. The degree of morning preference shows a continuum from morning type (with an advanced sleep phase) to evening type (with a delayed sleep phase) [1]. The chronotype is known to be determined in a complex manner by age, sex, and various environmental factors, including level of light exposure [2-4]. Developmental changes in chronotype occur; these include earlier preference during childhood, later preference in adolescence and early...
adulthood, and gradually earlier preference with advancing age [2].

Chronotype has been studied extensively in a number of ways. For convenience in many studies, chronotype has been measured subjectively and arbitrarily classified as morning, intermediate, and evening types, according to the Composite Scale of Morningness (CS) [5,6]. Chronotype shows a normal or near-normal distribution in the population; the heritability of chronotype has been shown to range between 21% and 52% [7]. Therefore, chronotype is presumably a polygenic quantitative trait [8,9]. Circadian rhythms are regulated by a set of circadian genes in mammals [9,10]. These circadian genes are also known as clock genes; they include ADCYAP1, ARNTL, BHLHB2, BHLHB3, CLOCK, PER1, PER2, PER3, NR1D1, NPAS2, NR1D1, THRA, CSNK1D, CSNK1E, CRY2, RORA, RORB, TIMELESS, and VIP. Notably, the list continues to grow due to discoveries in this field. Recently, three genome-wide association studies (GWASs) have been performed in people of European ancestry [11-13]. These GWASs identified meaningful overlap of genes that exhibited significant associations with chronotype. All three GWASs supported the association of four genes—PER2, RGS16, FBXL13, and AK5—with chronotype [14]. A meta-analysis of these three GWASs identified 351 loci associated with chronotype, of which 327 loci were novel and 24 loci had been reported in other GWASs [15].

GWASs can identify only common genetic variants, with small or modest effects for complex traits. Chronotype is a complex trait, which may be influenced by polygenes. A recent GWAS found only common genetic variants with small individual genetic effects on chronotypes. Only a subset of the significant variants found in that GWAS were known clock genes or have been suggested to function as clock genes. There were many other genetic variants without known functions as clock genes. Genetic control of chronotype must occur in multiple ways, with individual genotypic associations of polygenes, haplotypic associations of more crucial genes, and gene-gene interactions of multiple biologically related functional genes. Here, we investigated possible roles for gene-gene interactions of clock genes in chronotype in a Korean population.

Methods

Participants
The study population consisted of 1,293 unrelated healthy Korean individuals. These were the same individuals included as controls in our previous study of bipolar disorder [16]. They consisted mostly of college students, nurses, and public officials, who were recruited after a brief psychiatric interview. Potential participants were excluded if they reported a history of a psychotic disorder, mood disorder, anxiety disorder, substance use disorder, brain trauma, or intellectual disability. All participants were informed of the purpose and methods of the study and provided informed consent before enrollment. The Ethics Committee of Eulji General Hospital approved the study protocol (IRB No. 2016-08-009).

Measurement of chronotype
Chronotype was measured using a self-reported questionnaire. The CS is a 13-item questionnaire, which assesses individual differences in the time of day a person prefers to carry out various activities; it classifies people as morning, intermediate, or evening types [5]. Three items are scored on a five-point scale from 1 to 5; the other 10 items are scored on a four-point scale, from 1 to 4. Higher scores indicate morning preference. All participants completed the CS questionnaire.

Genotyping
The clock genes investigated in this study were BHLHB2, CLOCK, CSNK1E, NR1D1, PER1, PER2, PER3, and TIMELESS. These eight genes were analyzed for 19 different tag single nucleotide polymorphisms (SNPs) with minor allele frequencies exceeding 5% in Asian populations. DNA was extracted from blood and SNPs were genotyped using the TaqMan method (Applied Biosystems, Foster City, CA, USA). Table 1 presents a summary of the minor allele frequencies and chromosomal locations of the SNPs.

Statistical analysis
Individual SNPs were examined for Hardy-Weinberg equilibrium; two SNPs violating Hardy-Weinberg equilibrium were removed. Each SNP association with CS score was analyzed by simple regression analysis. Haplotype association with CS was also analyzed by PLINK if more than two SNPs for each gene were included [17].

Gene-gene interactions were analyzed using the quantitative multifactor-dimensionality reduction (QMDR) method, an extension of the multifactor-dimensionality reduction (MDR) algorithm to work with quantitative or continuous phenotypes [18]. The MDR method is one a commonly used method for detection and characterization of high-order gene-gene or gene-environment interactions in case-control studies; this comprises a nonparametric combinatorial approach that reduces the number of dimensions [19]. For each multi-locus genotype combination, QMDR calculates the mean value of phenotype and compares it to the overall mean to determine the genotype combination is high risk or low risk. By pooling all the genotypes into either high-risk or low-risk groups, a new binary attribute is created. The t-test is used to com-
pare the means between high and low risk groups using a t-test and t-statistic is used as a training score to choose the best model. In QMDR, the training and testing score are defined by t-test statistic. The training score is used to determine the best K-order interaction model. QMDR use 10-fold cross-validation and cross-validation consistencies (CVCs) of each model chosen are recorded. The best overall QMDR model is selected as that with the maximum testing score and highest cross-validation consistency. To estimate the p-values of the chosen model, empirical null distribution is used [18].

In this study, interactions of up to three loci were tested using 10-fold cross-validation in a search considering all possible SNP combinations. SNP combination with maximum CVC was considered the best model. p-values were determined empirically by 1,000-fold permutations of case and control labels.

**Results**

The study population consisted of 481 male participants (37.2%) and 812 female participants (62.8%). Mean ages were 27.5 ± 8.3 years for male participants and 23.7 ± 3.5 years for female participants. Mean CS scores were 32.3 ± 6.4 for male participants and 29.7 ± 5.7 for female participants. The classification of chronotype using total CS score revealed 305 evening type participants (23.6%), 911 intermediate type participants (70.5%), and 77 morning type participants (6.0%) in the study population. The distributions of total CS score according to age and chronotype are shown in Table 2.

Two SNPs of PER1 were excluded from further analysis because they were found to deviate from Hardy-Weinberg equilibrium. Ultimately, 17 SNPs of seven genes were analyzed. In regression analyses, no individual SNP showed a significant association with CS score (Table 3). There were no significant haplotype associations with CS score for any of the genes with more than two SNPs in this study (Table 4). On QMDR analyses, no single locus was found to be associated with chronotype, similar to the results of regression analysis. NR1D1 rs2314339 and TIMELESS rs4630333 were significantly associated with chronotype in a two-locus model (CVC = 8/10, p = 0.041). In the three-locus models, NR1D1 rs2314339, TIMELESS rs4630333, and PER3 rs228669 showed the strongest association with chronotype (CVC = 4/10, p < 0.001). A summary of QMDR results with CVC > 1/10 is presented in Table 5.

**Discussion**

We hypothesized that circadian genes play an important role in chronotype regulation and that there are gene-gene interaction effects on chronotype. We identified the best interaction models for two and three loci, as well as statistical significances of the best interaction models for chronotype in a Korean population, using the QMDR method and corresponding permutation test. However, it
Table 2. CS score by age and chronotype distribution

<table>
<thead>
<tr>
<th>Variable</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–9</td>
<td>481 (0.2)</td>
<td>0 (0.1)</td>
<td>1 (0.1)</td>
</tr>
<tr>
<td>10–19</td>
<td>1 (0.1)</td>
<td>35 (4.3)</td>
<td>44 (3.4)</td>
</tr>
<tr>
<td>20–29</td>
<td>147 (30.6)</td>
<td>361 (44.5)</td>
<td>508 (39.3)</td>
</tr>
<tr>
<td>30–39</td>
<td>261 (54.3)</td>
<td>377 (46.4)</td>
<td>638 (49.3)</td>
</tr>
<tr>
<td>40–49</td>
<td>61 (12.7)</td>
<td>39 (4.8)</td>
<td>100 (7.7)</td>
</tr>
<tr>
<td>50–59</td>
<td>2 (0.4)</td>
<td>0</td>
<td>2 (0.2)</td>
</tr>
<tr>
<td>CS score</td>
<td>32.3 ± 6.4</td>
<td>29.7 ± 5.7</td>
<td>30.7 ± 6.1</td>
</tr>
</tbody>
</table>

Chronotype
- Evening type (≤ 26): 86 (17.9) | 219 (27.0) | 305 (23.6)
- Intermediate type (27–40): 346 (71.9) | 565 (69.6) | 911 (70.5)
- Morning type (≥ 41): 49 (10.2) | 28 (3.4) | 77 (6.0)

Values are presented as number (%) or mean ± SD.
CS, Composite Scale of Morningness.

Table 3. Individual SNP association analysis using simple regression

<table>
<thead>
<tr>
<th>Gene</th>
<th>No. of SNPs</th>
<th>No. of haplotypes</th>
<th>F(df)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHLHB2 rs6442925CT</td>
<td>2</td>
<td>4</td>
<td>0.124(3)</td>
<td>0.946</td>
</tr>
<tr>
<td>BHLHB2 rs2137947CT</td>
<td>2</td>
<td>2</td>
<td>0.052</td>
<td>0.838</td>
</tr>
<tr>
<td>CLOCK rs1801260CT</td>
<td>2</td>
<td>4</td>
<td>0.096</td>
<td>0.465</td>
</tr>
<tr>
<td>CLOCK rs3805148AC</td>
<td>2</td>
<td>4</td>
<td>0.181</td>
<td>0.465</td>
</tr>
<tr>
<td>CLOCK rs1250430CG</td>
<td>2</td>
<td>2</td>
<td>0.160</td>
<td>0.465</td>
</tr>
<tr>
<td>CSNK1E rs135745CG</td>
<td>2</td>
<td>2</td>
<td>-0.262</td>
<td>-0.822</td>
</tr>
<tr>
<td>CSNK1E rs1534891CT</td>
<td>2</td>
<td>2</td>
<td>0.550</td>
<td>0.173</td>
</tr>
<tr>
<td>CSNK1E rs12649507AG</td>
<td>2</td>
<td>2</td>
<td>0.121</td>
<td>0.623</td>
</tr>
<tr>
<td>NR1D1 rs2314339CT</td>
<td>2</td>
<td>2</td>
<td>-0.229</td>
<td>0.335</td>
</tr>
<tr>
<td>NR1D1 rs2269457AG</td>
<td>2</td>
<td>2</td>
<td>0.025</td>
<td>0.916</td>
</tr>
<tr>
<td>PER2 rs2304672CG</td>
<td>2</td>
<td>2</td>
<td>-0.311</td>
<td>0.403</td>
</tr>
<tr>
<td>PER2 rs2304669AG</td>
<td>2</td>
<td>2</td>
<td>-0.372</td>
<td>-0.859</td>
</tr>
<tr>
<td>TIMELESS rs4630333AG</td>
<td>2</td>
<td>2</td>
<td>-0.127</td>
<td>-0.598</td>
</tr>
<tr>
<td>TIMELESS rs1082214AG</td>
<td>2</td>
<td>2</td>
<td>-0.372</td>
<td>0.390</td>
</tr>
</tbody>
</table>

SNP, single nucleotide polymorphism; SE, standard error; Model 1, model without adjustment for age and sex; Model 2, with adjustment for age and sex.

Table 4. Haplotype association analysis

<table>
<thead>
<tr>
<th>Gene</th>
<th>No. of SNPs</th>
<th>No. of haplotypes</th>
<th>f(df)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHLHB2</td>
<td>2</td>
<td>4</td>
<td>0.124(3)</td>
<td>0.946</td>
</tr>
<tr>
<td>CLOCK</td>
<td>5</td>
<td>3</td>
<td>0.196(2)</td>
<td>0.822</td>
</tr>
<tr>
<td>CSNK1E</td>
<td>3</td>
<td>5</td>
<td>2.080(4)</td>
<td>0.082</td>
</tr>
<tr>
<td>NR1D1</td>
<td>2</td>
<td>4</td>
<td>0.513(3)</td>
<td>0.674</td>
</tr>
<tr>
<td>PER2</td>
<td>2</td>
<td>3</td>
<td>0.423(2)</td>
<td>0.655</td>
</tr>
<tr>
<td>TIMELESS</td>
<td>2</td>
<td>3</td>
<td>0.378(2)</td>
<td>0.685</td>
</tr>
</tbody>
</table>

No. of haplotypes indicates the number of common haplotypes (minor haplotype frequency ≥ 0.01). SNP, single nucleotide polymorphism.

was difficult to conclude that there were clinically significant gene-gene interaction effects on chronotype based on our results, since the mean differences in total CS score between genotype combinations were minor.

Human chronotype is a heritable polygenic trait, and many groups have searched for the genes involved in chronotype. Many clock genes have been considered as strong candidate genes for chronotype because of their biological functions in circadian networks. There have been many single-gene association studies to...
identify genetic factors for chronotype. The first report of an association between chronotype and clock genes was related to the 3′-untranslated region of the CLOCK gene (rs1801260) [20]. This finding was not replicated in other populations [21-23], but was replicated in a Japanese population [24]. Multiple other clock genes have also been studied. The PER3 gene was shown to be associated with delayed sleep phase syndrome. The studies thus far have focused on a variable number of tandem repeats region located in exon 18 of PER3, but the results of association analysis differed among studies. Specifically, a shorter allele was associated with delayed sleep phase syndrome [25]; a longer allele was also reportedly associated with delayed sleep phase syndrome [26] and was reported to have predictive value in response to sleep loss [27-29]. PER1 and PER2 were reported to be associated with advanced sleep phase syndrome in a British population [27,28], but not in a Japanese family [30]. Overall, the results were inconsistent and sometimes contradictory. Table 6 presents a summary of genetic association findings between clock genes and chronotype.

Table 5. Summary of QMDR results having more than 1/10 of CVC

<table>
<thead>
<tr>
<th>Gene</th>
<th>CVC</th>
<th>Score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NR1D1 rs2314339CT</td>
<td>3/10</td>
<td>0.753</td>
<td>0.503</td>
</tr>
<tr>
<td>TIMELESS rs4630333AG</td>
<td>3/10</td>
<td>0.874</td>
<td>0.274</td>
</tr>
<tr>
<td>BHLHB2 rs2137947CT</td>
<td>2/10</td>
<td>0.760</td>
<td>0.492</td>
</tr>
<tr>
<td>PER3 rs228669AG</td>
<td>2/10</td>
<td>0.920</td>
<td>0.222</td>
</tr>
<tr>
<td>NR1D1 rs2314339, TIMELESS rs4630333</td>
<td>8/10</td>
<td>1.193</td>
<td>0.041</td>
</tr>
<tr>
<td>BHLHB2 rs2137947CT, TIMELESS rs4630333</td>
<td>1/10</td>
<td>0.919</td>
<td>0.272</td>
</tr>
<tr>
<td>NR1D1 rs2314339, PER3 rs228669AG</td>
<td>1/10</td>
<td>1.054</td>
<td>0.117</td>
</tr>
<tr>
<td>NR1D1 rs2314339, TIMELESS rs4630333, PER3 rs228669</td>
<td>4/10</td>
<td>1.741</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>CLOCK rs12504300CG, NR1D1 rs2314339CT, TIMELESS rs4630333AG</td>
<td>3/10</td>
<td>0.897</td>
<td>0.303</td>
</tr>
<tr>
<td>CSNK1E rs2075984AC, NR1D1 rs2314339CT, TIMELESS rs4630333AG</td>
<td>2/10</td>
<td>0.915</td>
<td>0.277</td>
</tr>
<tr>
<td>CSNK1E rs2075984AC, NR1D1 rs2314339CT, PER3 rs228669AG</td>
<td>1/10</td>
<td>0.792</td>
<td>0.495</td>
</tr>
</tbody>
</table>

QMDR, quantitative multifactor dimensionality reduction; CVC, cross-validation consistency.

*Average testing score.

*Empirical p-value.

The proteins encoded by various clock genes cooperate physically with each other and act as transcription factors. Combinations of polymorphisms in these genes may affect phenotype. Therefore, combined analysis of the effects of different clock genes may be more accurate and more revealing, compared to analyses of single genes. A few previous studies examined the effects of gene-gene interactions on chronotype. One study in Korean college students reported a significant interaction for CS score between CLOCK gene 3111 C/T and GNB3 825 C/T, according to regression analysis [32]. Later, the same group reported a genetic interaction for eveningness among ARNTL, PER2, and GNB3, according to MDR analysis [33]. Another study by Pedrazzoli et al. reported that a specific combination of polymorphisms in four clock genes was associated with diurnal preferences in a Brazilian population [34]. They chose four polymorphisms in four clock genes: PER2, PER3, CLOCK, and BMAL1. To the best of our knowledge, other than these studies, there have been no further analyses of the epistatic effects among clock genes for chronotype. Therefore, we attempted to identify gene-gene interactions among clock genes for chronotype in this study. We used QMDR to investigate gene-gene interactions to improve statistical power. We avoided subgrouping based on total CS score because grouping into morning, intermediate, or evening types could be arbitrary; moreover, we could address the quantitative score directly in our analysis. Multifactor dimensionality reduction is a common approach for identification of gene-gene interactions in case-control studies [19]. QMDR is an extension of MDR to handle quantitative phenotypes. Instead of comparing the case-control ratio of each multi-locus genotype to a fixed threshold as in MDR, QMDR compares the mean value of each multi-locus genotype to the overall mean [18].
Table 6. Summary of SNP association findings between clock genes and chronotype

<table>
<thead>
<tr>
<th>Gene</th>
<th>Polymorphism</th>
<th>Genetic region</th>
<th>Results</th>
<th>Population</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLOCK</td>
<td>rs1801260</td>
<td>3’-UTR</td>
<td>Positive association</td>
<td>American</td>
<td>[20]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>British</td>
<td>[21]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Positive association</td>
<td>Japanese</td>
<td>[24]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>Brazilian</td>
<td>[23]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>Korean</td>
<td>[32]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>British</td>
<td>[39]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>American</td>
<td>[44]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>Italian</td>
<td>[46]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>Korean</td>
<td>[33]</td>
</tr>
<tr>
<td>ARNTL</td>
<td>rs2278749</td>
<td>Intron</td>
<td>No association</td>
<td>Korean</td>
<td>[33]</td>
</tr>
<tr>
<td></td>
<td>rs7107287</td>
<td>Intron</td>
<td>No association</td>
<td>Polish</td>
<td>[38]</td>
</tr>
<tr>
<td>ARNTL2</td>
<td>rs922270</td>
<td>Intron</td>
<td>Positive association</td>
<td>British</td>
<td>[46]</td>
</tr>
<tr>
<td>PER1</td>
<td>rs2735611</td>
<td>Coding, synonymous</td>
<td>Positive association</td>
<td>British</td>
<td>[28]</td>
</tr>
<tr>
<td>PER2</td>
<td>rs2304672</td>
<td>5’-UTR</td>
<td>Positive association</td>
<td>Japanese</td>
<td>[30]</td>
</tr>
<tr>
<td></td>
<td>rs934945</td>
<td>Missense</td>
<td>Positive association</td>
<td>Korean</td>
<td>[47]</td>
</tr>
<tr>
<td></td>
<td>rs2304671</td>
<td>Coding, synonymous</td>
<td>Positive association</td>
<td>Korean</td>
<td>[33]</td>
</tr>
<tr>
<td></td>
<td>rs3533999</td>
<td>Missense</td>
<td>Positive association</td>
<td>British &amp; American</td>
<td>[49]</td>
</tr>
<tr>
<td>PER3</td>
<td>rs57875989</td>
<td>Deletion/insertion (VNTR)</td>
<td>Positive association</td>
<td>British</td>
<td>[25]</td>
</tr>
<tr>
<td></td>
<td>rs2314339</td>
<td>Promoter</td>
<td>Positive association</td>
<td>European</td>
<td>[50]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Positive association</td>
<td>British</td>
<td>[51]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>British</td>
<td>[39]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>Colombian</td>
<td>[41]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>Norwegian</td>
<td>[40]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>Han Chinese</td>
<td>[42]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>Japanese</td>
<td>[43]</td>
</tr>
<tr>
<td>TIMELESS</td>
<td>rs2291738</td>
<td>Intron</td>
<td>Positive association</td>
<td>Polish</td>
<td>[38]</td>
</tr>
<tr>
<td>NR1D1</td>
<td>rs12941497</td>
<td>Intron</td>
<td>Positive association</td>
<td>Korean</td>
<td>[37]</td>
</tr>
<tr>
<td>MTNR1B</td>
<td>rs4753426</td>
<td>Promoter</td>
<td>Positive association</td>
<td>Brazilian</td>
<td>[52]</td>
</tr>
<tr>
<td>GNB3</td>
<td>rs5443</td>
<td>Coding, synonymous</td>
<td>Positive association</td>
<td>European</td>
<td>[53]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>Korean</td>
<td>[32]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No association</td>
<td>Korean</td>
<td>[33]</td>
</tr>
</tbody>
</table>

SNP, single nucleotide polymorphism; UTR, untranslated region.

*Positive interaction with rs5443 of GNB3.

*European population in South Africa.

*European population in Sweden.

In particular, gene-gene interactions among NR1D1 rs2314339, TIMELESS rs4630333, and PER3 rs228669 were significantly associated with chronotype in QMDR analyses in the present study. These SNPs did not show any associations as individual SNPs or haplotypes. NR1D1 (nuclear receptor subfamily 1, group D, member 1) is the gene encoding REV-ERBa, located on chromosome 17q21.3. REV-ERBa suppresses the transcription of BMAL1 mRNA [35,36], while BMAL1 activates REV-ERBa; this comprises a feedback loop of the mammalian circadian oscillator. NR1D1 has been reported to show an association with chronotype in healthy Korean young adults [37]. Kang et al. [37] reported a significant association with rs12941497 of NR1D1, but no association with rs2314339, which showed significant gene-gene interactions with other SNPs in the present study. The TIMELESS gene was reportedly associated with morningness-eveningness in healthy university students [38]. The PER3 gene showed conflicting results. Several positive associations for chronotype in European populations have been reported. In addition, negative findings...
were also reported for European populations [39-41] and Asian populations [42,43]. The circadian clock system consists mainly of transcription-translation feedback loops in the internal timekeeping clock. Heterodimers of BMAL (brain and muscle Arnt-like protein-1) and CLOCK (circadian locomotor output cycles kaput) activate the transcription of PER (period) and CRY (cryptochrome) genes. CRY and PER suppress transcriptional activity of BMAL1/CLOCK [54,55]. Because NR1D1 is related to BMAL1 and PER3 was reported to show a strong genetic interaction with BMAL1, our main finding supports this hypothesis. Unfortunately, no biological mechanism has been reported for chronotype that includes positive gene-gene interactions of all three genes identified in this study. Further studies are needed to understand the complicated biological molecular network of circadian clocks in humans.

Consistent and clear phenotyping is critical when performing genetic studies. We used CS score as a proxy for actual human diurnal preference. However, it evaluates both diurnal preference and sleep homeostasis. There are many ways to determine the timing of the circadian system (e.g., core body temperature monitoring, dim light melatonin onset, and actigraphy- or diary-based midpoint of sleep); self-reported diurnal preference is only a proxy method. Most GWASs have assessed chronotype using only a single question, such as “Are you naturally a night person or a morning person?” Non-precise phenotyping can produce unreliable significant findings that are unlikely to be replicated in subsequent studies [56]. Therefore, future phenotyping should include standardized self-reporting, clinical interviews, or objective assessment of sleep-wake periodicity, such as actigraphy. It is important to validate commonly used self-reported items. Comparisons between self-reported items and biological markers of circadian rhythms are needed to determine which questions are most closely associated with endogenous processes [14].

This study had some limitations that must be considered when interpreting our results. First, this study included only ethnically Korean individuals. Therefore, caution is needed when generalizing our results to other populations. As suggested in previous studies, chronotype and genes for chronotype are likely to differ according to ethnicity and/or between populations. Therefore, this study in a Korean population was necessary. Second, our participants were relatively young, with a mean age of 25 years. Because chronotype is affected by age, our results cannot be applied to other age groups. Third, because of resource limitations, we could include only 17 SNPs of seven circadian genes. Therefore, our results represent only a subset of the real-world epistatic interactions among clock genes. There are likely to be more complicated gene-gene interactions among circadian genes, as well as epistatic interactions between circadian genes and genes of other biological systems, which are directly and indirectly related to the circadian system.

We could not conclude that clock genes play a critical role in determining chronotype, although QMDR suggested significant gene-gene interactions. Further studies are needed to investigate gene-gene interactions of additional clock genes, especially with respect to SNPs that repeatedly show significant associations in multiple GWASs and studies in populations other than those of European ancestry.

ORCID

Mira Park: https://orcid.org/0000-0003-3827-9089
Soon Ae Kim: https://orcid.org/0000-0002-9831-0511
Jieun Shin: https://orcid.org/0000-0002-6192-2610
Eun-Jeong Joo: https://orcid.org/0000-0001-8766-8713

Authors’ Contribution

Conceptualization: EJJ. Data curation: JS, SAK. Formal analysis: JS. Funding acquisition: EJJ. Methodology: MP. Writing – original draft: MP, EJJ. Writing – review & editing: EJ, SAK.

Conflicts of Interest

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

Acknowledgments

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) & funded by the Korean government (MSIP&MOHW) (No. 2016 M3A9B6904241, 2016M3A9B6904244).

References


https://doi.org/10.5808/GI.2020.18.4.e38


44. Chang AM, Buch AM, Bradstreet DS, Klements DJ, Duffy JF. Human diurnal preference and circadian rhythmicity are not associated with the CLOCK 3111C/T gene polymorphism. J Biol Rhythms 2011;26:276-279.
Prediction of hub genes of Alzheimer’s disease using a protein interaction network and functional enrichment analysis

Jia Jin Wee, Suresh Kumar*

Faculty of Health and Life Sciences, Management and Science University, 40100 Shah Alam, Malaysia

Alzheimer's disease (AD) is a chronic, progressive brain disorder that slowly destroys affected individuals' memory and reasoning faculties, and consequently, their ability to perform the simplest tasks. This study investigated the hub genes of AD. Proteins interact with other proteins and non-protein molecules, and these interactions play an important role in understanding protein function. Computational methods are useful for understanding biological problems, in particular, network analyses of protein-protein interactions. Through a protein network analysis, we identified the following top 10 hub genes associated with AD: PTGER3, C3AR1, NPY, ADCY2, CXCL12, CCR5, MTNR1A, CNR2, GRM2, and CXCL8. Through gene enrichment, it was identified that most gene functions could be classified as integral to the plasma membrane, G-protein coupled receptor activity, and cell communication under gene ontology, as well as involvement in signal transduction pathways. Based on the convergent functional genomics ranking, the prioritized genes were NPY, CXCL12, CCR5, and CNR2.

Keywords: Alzheimer's disease, functional enrichment, hub genes, network analysis, protein-protein interaction

Introduction

As people become older, many parts of the body, including the brain, change. It is also normal for people to become forgetful, and age-associated memory impairment is considered to be part of the aging process. However, Alzheimer’s disease (AD) is distinct from age-associated memory impairment, and instead is a type of dementia that causes problems with memory, thinking, and behavior. Hence, it is understandable that people, especially the elderly, are concerned about memory loss, as it is a symptom of AD. Dementia is known to be progressive, meaning that the condition becomes worse gradually. It is well known that AD has complicated and diverse pathogenic causes, including genetic, environmental, and immunological factors, as well as head trauma, depression, and hypertension. Moreover, genetic analyses have shown that human variations in AD can originate from several genes and their variants, which exert different biological functions in coordination to increase disease risk. AD typically occurs in elderly people (aged 65 years and above), while an uncommon variant known as early-onset AD comprises about 5% of AD cases [1]. As the name suggests, people with early-onset AD develop symptoms during their 40s and 50s, although the symptoms of both variants of AD are mostly the same.
According to the National Institute on Aging (https://www.nia.nih.gov/), some main characteristics of a brain with AD include amyloid plaques, neurofibrillary tangles, and chronic inflammation [2]. The amyloid plaques refer to beta-amyloid peptide (Aβ) which are the key components of amyloid plaques in brains affected by AD [3,4]. Abnormal levels and accumulation of Aβ form plaques that disrupt cell function. A similar process of unusual accumulations accounts for neurofibrillary tangles, which are driven by the intraneuronal accumulation of tau protein, which otherwise functions to stabilize microtubules, and cause AD [5,6]. Chronic inflammation is also linked to AD through the dysfunction of microglia in the central nervous system (CNS), which maintain homeostasis in the brain. The inability of microglia to function causes chronic inflammation.

Previous research has determined that carriers of the APOE-e4 risk gene have a higher risk of AD [7], and it is estimated that 40%–65% of people with AD have that gene variant. Previous studies have mostly focused on gene regulatory networks in the late onset of AD [8] and the identification of active transcription factors by analyzing miRNA regulatory pathways [9]. To explore the molecular changes underlying AD, several genome-wide expression profiling experiments have been performed on the post-mortem brain tissues of AD patients. However, the precise pathogenesis of AD remains unknown, and no effective treatment and prevention approaches are feasible. Apart from determining the pathways involved in AD pathogenesis, detailed analyses of possible candidate genes might lead to the identification of new strategies for predictive or diagnostic AD testing. In this study, we used a comprehensive database, DisGeNET, which includes information on all the genes related to AD to identify the hub genes involved in the disease. This study aimed to identify the hub genes involved in AD via protein-protein interactions.

Methods

Protein-protein interaction data collection
DisGeNET (https://www.disgenet.org/) [10] is a database containing information about human genes and variants. The data in DisGeNET are drawn from sources such as the scientific literature, animal models, and expertly curated repositories. Using the search tool in DisGeNET, the name of the disease (AD) was entered as a search query while choosing the disease search button. Then, the option “Summary of Gene-Disease Associations” was selected. A summary of all genes’ information, such as the HGNC gene symbol, UniProt ID, and protein class, was displayed. All of this information was downloaded in a Microsoft Excel file.

Network construction and analysis
For this study, Cytoscape (https://cytoscape.org/) [11], an open-source software project, was used to construct the network by entering all the UniProt IDs [12] as search queries, while using the Search Tool for the Retrieval of Interacting Genes (STRING) for protein queries. STRING (https://string-db.org/) [13] is a database that contains information on known and predicted protein-protein interactions. Ambiguous terms were resolved by setting the confidence (score) cutoff to 0.4 and their maximum additional interactors to 0 to import the network.

Hub gene identification
To determine the potential hub genes of AD, CytoHubba [14] was used to calculate the score for each node. This tool uses the maximal clique centrality (MCC) algorithm to show the top 10 ranked nodes, which could be the potential top 10 hub genes of AD. The top 10 genes were further analyzed for gene enrichment.

Functional enrichment of hub genes
FunRich (http://www.funrich.org/) [15] was used for a functional enrichment and interaction network analysis of genes and proteins, using the gene ID of the top 10 genes. When applying the search, the analysis tab was selected and the bar graphs for four aspects (cellular component, molecular function, biological process, and biological pathway) were viewed, with the first six items of each category being shown on the chart.

Gene prioritization using the AlzData database
AlzData is an integrated AD database that uses high-throughput omics data such as the results of genome-wide association studies (GWAS), whole-exome sequencing, transcriptome analysis, and proteomics to generate a prioritized gene list. We used this database to prioritize the top 10 hub genes identified.

Results and Discussion

In the data downloaded from DisGeNET (accessed December 2019), there were 1981 genes involved in AD (Supplementary Table 1). A protein-protein interaction network was created by querying the STRING database for gene symbols with a confidence score of 0.4 to avoid false positives. The network of these genes had 1922 nodes and 57,617 edges, as shown in Fig. 1. Proteins with higher degrees in the network (hub genes) are more likely to be essential proteins. CytoHubba uses 11 methods to retrieve the top-ranked nodes. In CytoHubba, the MCC method captures more essential proteins in the top-ranked list in both high-degree and low-degree proteins. The top 10 nodes selected by MCC were
all highly essential, meaning that these genes could be potential hub genes. These hub genes, by gene ID, were prostaglandin E receptor 3 (PTGER3), C3AR1, NPY, ADCY2, CXCL12, CCR5, MTNR1A, CNR2, GRM2, and CXCL8. The results of the gene enrichment analysis for cellular components (Fig. 2), molecular function (Fig. 3), biological processes (Fig. 4), and biological pathways (Fig. 5) were displayed in the bar graphs.

Fig. 2 shows that the highest percentage of these genes was found at the plasma membrane, and including genes integral to the plasma membrane. The extracellular component accounted for the lowest percentage of AD hub genes.

Regarding the molecular function of the AD hub genes (Fig. 3), most of the genes were related to G-protein coupled receptor (GPCR) activity, and were rarely involved in other activities such as cytokine receptor, chemokine, and adenylyl cyclase activity.

In terms of biological processes (Fig. 4), most of these hub genes were involved in signal transduction and cell communication processes, rather than other processes such as the immune response or other unknown processes.

Fig. 5 shows the biological pathways of the hub genes. Most of these genes are involved in signal transduction and signaling by GPCR, as well as the class A/1 (rhodopsin-like receptors) pathway, peptide ligand-binding receptors, and chemokine receptors.

PTGER3 and prostaglandin E2 are derived from the metabolism of arachidonic acid by cyclooxygenases in the cyclooxygenase pathway. This protein is the main neuroinflammatory molecule [16], and its receptor EP3 subtype is highly expressed in the brain. Studies have shown that activation of the EP3 receptor can reduce or suppress cyclic adenosine monophosphate (cAMP) formation [17]. This affects the microglia, thereby causing many brain diseases including AD.

C3AR1, or complement C3a receptor 1, is another gene involved in AD. C3a is an anaphylatoxin released during activation of the complement system. The C3AR1 gene encodes the orphan GPCR for C3a. It also has essential functions in the immune re-
response and host defense. A study has shown the importance of activation of the C3-C3aR network in mediating neuroinflammation and tau pathology [6]. The tau protein is a key protein that has been implicated in many neurodegenerative diseases such as AD and Parkinson disease.

A widely expressed gene in the CNS, NPY encodes neuropeptide Y. Neuropeptides are signaling molecules that influence brain activity in specific ways. They are involved in the pathophysiology of AD, and those with AD were found to have notably lower plasma levels of neuropeptide Y than healthy individuals [18]. ADCY2 encodes the protein adenylate cyclase 2. This, according to the NCBI, is a membrane-associated enzyme that catalyzes the formation of cAMP, which is a messenger for intracellular signal induction. A study reported that in patients with AD, cAMP activity was higher in cerebral microvessels than in healthy individuals [19].

The CXCL12 gene encodes the chemokine protein named C-X-C motif chemokine 12, which is a member of the intracrine family of stromal cell-derived chemokines and is involved in the CXCL12-CXCL4 pathway. This molecule regulates neuronal excitability and synaptic transmission [20]. A test conducted using a mouse model [21], proved that patients with AD had reduced amounts of CXCL12. This, in turn, is linked to the fact that these patients have impaired learning and memory, which is also a symptom of AD. The C-C chemokine receptor family is a main inflammatory receptor family that has been found to be involved in AD [22]. CCR5 or C-C motif chemokine receptor 5 is one such receptor. Many studies have demonstrated upregulation of this chemokine receptor in patients with AD, and it has been reported

\[ \text{Fig. 3. Molecular functions of Alzheimer's disease hub genes.} \]

\[ \text{Fig. 4. Biological processes of Alzheimer's disease hub genes.} \]
to recruit microglia and to cause accumulations of microglia in senile plaques [23], thereby accelerating AD development [22].

MTNR1A encodes melatonin receptor 1A. Melatonin is a hormone that is released in the pineal gland and regulates the sleep-wake cycle. Furthermore, melatonin has been reported to have neuroprotective and anti-amyloidogenic effects, as it reduced Aβ production in multiple neuronal cell lines [3]. As people age, the secretion of this hormone decreases, and low levels of melatonin contribute to aging.

The endocannabinoid system (ECS) comprises endocannabinoids, which are endogenous lipid-based retrograde neurotransmitters that bind to cannabinoid receptors. The ECS is involved in regulating physiological and cognitive processes, including include memory. One of the key receptors, cannabinoid receptor 2, is encoded by CNR2. Imbalances in the ECS, including elevated expression of glial cannabinoid receptor 2, in AD models suggest its potential role in inflammatory and neuroprotective processes [24].

GRM2, which stands for glutamate metabotropic receptor 2, encodes a protein named metabotropic glutamate receptor 2 (mGluR2). Glutamate is the main excitatory neurotransmitter in the CNS, and mGluR2 modulates rapid synaptic transmission in the CNS via controlled release of the excitatory amino acid glutamate [25]. Altered glutamatergic synaptic transmission is a key event in the development of AD.

CXCL8 encodes C-X-C motif chemokine ligand 8, which is a member of the CXC chemokine family that is also known as interleukin 8 (IL-8). Inflammatory processes have been found to be involved in neurodegenerative disorders such as AD, and the involvement of chemokines such as IL-8 has been reported to be involved in these inflammatory processes [26].

We used the convergent functional genomics (CFG) ranking for target genes available in the AlzData database. AlzData (http://www.alzdata.org/) integrates five lines of evidence associated with AD. One CFG point is assigned for each piece of evidence (e.g., expression of the target gene is regulated by AD genetic variants in GWAS; the target gene has significant physical interactions with APP, PSEN1, PSEN2, APOE, or MAPT; the target gene is differentially expressed in AD mouse models before AD pathology emergence; the target gene expression is correlated with AD pathology in Aβ-line AD mouse models and tau line AD mouse models). The number of CFG points ranges from 0 to 5. According to the CFG ranking, the top four genes (with 4 points each) were NPY, CXCL12, CCR5, and CNR2, and the genes in second place (with 2 points each) were PTGER3, MTNR1A, and GRM2.
Several approaches have been used in previous research to identify potential target genes, such as ZFHX3, ERBB2, ERBB4, OCT3, MIF, CDK13, and GPI [27-33]. The most recent gene expression analysis conducted by Yan et al. [34] identified the following hub genes: CDC42, VEGFA, BDNF, PDYN, CALB, TH, CACNA1A, OXT, CD44, and TAC1. The genes identified by Wu et al. [35] were ITGB5, RPH3A, GNAS, and THY1. Thus, the present study found a few previously unreported novel hub genes, as follows: PTGER3, C3AR1, NPY, ADCY2, CXCL12, CCR5, MTNR1A, CNR2, GRM2, and CXCL8. The priority genes were identified as NPY, CXCL12, CCR5, and CNR2 based on the CFG ranking.

Protein-protein interactions are important for understanding protein function and behavior. In this study, we identified 10 hub genes of AD. The results show that most of these genes encode receptor proteins that are involved in biological pathways in the plasma membrane. The hub genes identified through a network analysis can be used as targets to suppress AD in patients. Our analysis can shed some light on a deeper understanding of the fundamental molecular pathways and key molecular players of AD and offers a new point of view for researchers studying the causes of AD.

### Table 1. CFG ranking of top the 10 hub genes of AD

<table>
<thead>
<tr>
<th>Gene</th>
<th>eQTL</th>
<th>GWAS</th>
<th>PPI</th>
<th>Early_DEG</th>
<th>Pathology cor (Aβ)</th>
<th>Pathology cor (tau)</th>
<th>CFG</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTGER3</td>
<td>1</td>
<td>0</td>
<td>APP, PSEN1, APOE</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>2</td>
</tr>
<tr>
<td>C3AR1</td>
<td>0</td>
<td>0</td>
<td>APP, APOE</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>NPY</td>
<td>NA</td>
<td>1</td>
<td>APP</td>
<td>Yes</td>
<td>-0.374*</td>
<td>0.166 ns</td>
<td>4</td>
</tr>
<tr>
<td>ADCY2</td>
<td>NA</td>
<td>0</td>
<td>APP</td>
<td>NA</td>
<td>-0.082 ns</td>
<td>-0.481 ns</td>
<td>1</td>
</tr>
<tr>
<td>CXCL12</td>
<td>1</td>
<td>0</td>
<td>APP, PSEN1, MAPT, APOE</td>
<td>Yes</td>
<td>0.432**</td>
<td>-0.069 ns</td>
<td>4</td>
</tr>
<tr>
<td>CCR5</td>
<td>1</td>
<td>0</td>
<td>APP</td>
<td>Yes</td>
<td>0.769***</td>
<td>0.616*</td>
<td>4</td>
</tr>
<tr>
<td>MTNR1A</td>
<td>1</td>
<td>0</td>
<td>APP, PSEN1, APOE</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>2</td>
</tr>
<tr>
<td>CNR2</td>
<td>1</td>
<td>0</td>
<td>APP, PSEN1, APOE</td>
<td>Yes</td>
<td>0.854***</td>
<td>0.750**</td>
<td>4</td>
</tr>
<tr>
<td>GRM2</td>
<td>1</td>
<td>0</td>
<td>APP</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>2</td>
</tr>
<tr>
<td>CXCL8</td>
<td>NA</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0</td>
</tr>
</tbody>
</table>

CFG, convergent functional genomics; AD, Alzheimer disease; GWAS, genome-wide association studies; PPI, protein protein interaction; DEG, differentially expressed gene; Aβ, beta-amyloids; NA, not available.

eQTL: expression of target gene is regulated by AD genetic variants (genetic variants: IGAP GWAS P < 1E-3; eQTL: p < 1E-3); GWAS: IGAP p < 1E-3; PPI: target gene has significant physical interaction with APP, PSEN1, PSEN2, APOE, or MAPT (p < 0.05); Early_DEG: target gene is differentially expressed in AD mouse models before AD pathology emergence; Pathology cor (Aβ): correlation of target gene expression with AD pathology in Aβ-line AD mouse models (r, p; ns, p > 0.05; *p < 0.05, **p < 0.01, ***p < 0.001); Pathology cor (tau): correlation of target gene expression with AD pathology in tau line AD mouse models (r, p; ns, p > 0.05; *p < 0.05, **p < 0.01, ***p < 0.001); CFG: total CFG score of a target gene, 1 CFG point is assigned if any of the above evidence is significant, the total of CFG points ranges from 0 to 5.

### Authors’ Contribution

Conceptualization: SK. Formal analysis: JJW. Writing - original draft: JJW, SK. Writing - review & editing: SK.

### 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


### ORCID

Jia Jin Wee: https://orcid.org/0000-0002-2295-0467
Suresh Kumar: https://orcid.org/0000-0001-5682-0938

https://doi.org/10.5808/GI.2020.18.4.e39


Development of reverse transcription loop-mediated isothermal amplification assays for point-of-care testing of avian influenza virus subtype H5 and H9

Songzi Zhang1,2,3,4, Juyoun Shin5, Sun Shin2,3,4*, Yeun-Jun Chung1,2,3,4**

1Department of Biomedicine & Health Sciences, Graduate School, The Catholic University of Korea, Seoul 06591, Korea
2Department of Microbiology, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea
3Precision Medicine Research Center, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea
4Integrated Research Center for Genome Polymorphism, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea
5ConnectaGen, Hanam 12918, Korea

Avian influenza (AIV) outbreaks can induce fatal human pulmonary infections in addition to economic losses to the poultry industry. In this study, we aimed to develop a rapid and sensitive point-of-care AIV test using loop-mediated isothermal amplification (LAMP) technology. We designed three sets of reverse transcription LAMP (RT-LAMP) primers targeting the matrix (M) and hemagglutinin (HA) genes of the H5 and H9 subtypes. RT-LAMP targeting the universal M gene was designed to screen for the presence of AIV and RT-LAMP assays targeting H5-HA and H9-HA were designed to discriminate between the H5 and H9 subtypes. All three RT-LAMP assays showed specific amplification results without nonspecific reactions. In terms of sensitivity, the detection limits of our RT-LAMP assays were 100 to 1,000 RNA copies per reaction, which were 10 times more sensitive than the detection limits of the reference reverse-transcription polymerase chain reaction (RT-PCR) (1,000 to 10,000 RNA copies per reaction). The reaction time of our RT-LAMP assays was less than 30 min, which was approximately four times quicker than that of conventional RT-PCR. Altogether, these assays successfully detected the existence of AIV and discriminated between the H5 or H9 subtypes with higher sensitivity and less time than the conventional RT-PCR assay.

Keywords: avian influenza virus, H5 subtype, H9 subtype, RT-LAMP
(HPAIV) in poultry occurred across diverse Asian countries, mainly caused by influenza H5N1 virus [2,3]. The outbreaks spread widely throughout the globe [4], and caused several human infections, including fatal cases [5]. In addition to HPAIVs, low-pathogenic avian influenza viruses (LPAIVs), which usually produce mild or no symptoms in birds, can induce pathogenic avian influenza by antigenic drift or shift, and can also induce human infection [2]. Indeed, the H9 subtype AIV, an LPAIV, was reported to be capable of human infection [6], and the World Health Organization warned that the H9N2 subtype could trigger a global influenza outbreak in humans, albeit with relatively lower pathogenicity than H5N1 or H7N9. Therefore, identification of HPAIVs and LPAIVs would be important to control avian influenza and potential human pandemic infections.

To identify AIVs at the early stage of disease, a rapid, specific, and sensitive detection method is required. Polymerase chain reaction (PCR) and reverse transcription PCR (RT-PCR) are the most commonly used tools, serving as a gold standard for the diagnosis of viral infections including influenza. However, these methods require specialized equipment and trained personnel for the entire experimental process [7], which makes these assays hard to use for point-of-care testing (POCT) in the field. In addition, ordinary PCR or RT-PCR requires about 2 hours to process and well-purified nucleic acids because the reaction is sensitive to PCR inhibitors.

Notomi et al. [8] developed a new nucleic acid amplification technology, loop-mediated isothermal amplification (LAMP), which can amplify target nucleic acids at a consistent temperature without changing the temperature for denaturation, annealing, and extension. LAMP only needs strand displacement by using the Bst DNA polymerase enzyme for nucleic acid amplification, so it can be amplified under isothermal conditions with a high amplification efficiency [8]. LAMP technology has a number of advantages for POCT [9-11]. For example, the time required for LAMP assays is much shorter than that for PCR, and LAMP assays can detect the target nucleic acid with high sensitivity without the equipment needed for thermal cycling. Furthermore, diverse detection methods can be applied, such as turbidity, colorimetric detection, and fluorescent dye incorporation.

Due to those advantages, LAMP technology has been widely applied to detect various pathogens and its applications have been advanced to include several different forms of assays, such as reverse transcription LAMP (RT-LAMP) and multiplex LAMP [12,13]. In particular, RT-LAMP assays have been found to be an efficient tool to detect RNA viruses such as the severe acute respiratory syndrome coronavirus 2, influenza, dengue, and West Nile viruses [14-22]. In RT-LAMP, reverse transcription of viral RNA and cDNA amplification can be processed in a single tube as one step under isothermal conditions. The straightforwardness of RT-LAMP is ideal for on-site field testing or diagnostics in a POCT setting where laboratory equipment is limited [13]. Together, RT-LAMP can be a simple and effective tool to detect and distinguish diverse AIV subtypes in the poultry field or at clinical quick-check desks [19].

In this study, we developed RT-LAMP assays that can detect a universal target for AIV and can discriminate between the H5 and H9 subtypes of AIV. We validated the sensitivity, specificity, and processing time of our assay.

Methods

Primer design for RT-LAMP assays

The nucleotide sequences of the matrix (M) gene and the HA genes from the H5 and H9 subtypes (H5-HA and H9-HA, respectively) were retrieved from the Influenza Research Database (IRD) and Global Initiative on Sharing All Influenza Data (GISAID), and aligned using DNAStar (Lasergene, Madison, WI, USA). The RT-LAMP assay primers were designed based on M, H5-HA and H9-HA sequence alignments by using the software PrimerExplorer version 5 (http://primerexplorer.jp/lampv5e/). The feasibility of all sets of primers was then subsequently validated using the BLAST program.

Preparation of target RNAs

Full-length fragments of the AIV M, H5-HA, and H9-HA genes with the T7 promoter sequence (TAATACGACTCCTATAGGGAGA) were chemically synthesized and cloned into a pTwist Amp High Copy plasmid (Twist Bioscience, South San Francisco, CA, USA). The RNA was transcribed by the universal M13F and M13R-pUC primers using the T7 RibomAX Express Large Scale RNA Production System (Promega, Madison, WI, USA). The transcribed RNA was 10-fold serially diluted from 10^5 copies/μL to 10 copies/μL, and used as a template for RT-LAMP and RT-PCR.

RT-LAMP

The RT-LAMP reaction was carried out as described elsewhere [18-20]. In total, 25 μL of a mixture containing 0.2 μM of each outer primer (F3 and B3), 1.6 μM of each inner primer (FIP and BIP), 0.8 μM of each loop primer (LF and LB), 8 U of the Bst 2.0 DNA polymerase (New England Biolabs, Hitchin, UK), 2 U of AMV reverse transcriptase (New England Biolabs), 8 mM of MgSO4 (New England Biolabs), 1.4 mM of each dNTPs (Thermo Fisher Scientific, Waltham, MA, USA), 1 × isothermal amplification buffer (New England Biolabs), 1 × PrimeSTAR® GXL DNA polymerase (Takara), 1 mM of each target primer, and 1 μL of the initial template RNA.
England Biolabs) and 0.4 M N-methylformamide (NMF) and isobutylamide (IBA) was prepared. For the real-time assay, 1 U of SYTO9 stain (Thermo Fisher Scientific) was added. One microliter of transcribed RNA was added to the respective tube. The reaction was carried out at 60°C for the H5-HA gene or at 68°C for the M gene and the H9-HA gene in a thermal cycler or CFX96 Touch Real-Time PCR Detection System (Bio-Rad Laboratories, Hercules, CA, USA). The reaction was terminated by heating at 80°C for 10 min. The fluorescence curve was captured in real time, and the result was illustrated as a graph on the monitor of the real-time system, verifying the amplification. RT-LAMP products were then evaluated by electrophoresis using 1.5% agarose gels to ensure the products of amplification reaction.

RT-PCR
One-step RT-PCR was performed with the AIV Multi-tube RT-PCR Kit (iNtRON Biotechnology, Seongnam, Korea), which was approved by the Animal and Plant Quarantine Agency of Republic of Korea, according to the manufacturer’s instructions. The reaction cycling conditions were as follows: 30 min of RT at 45°C, RT inactivation and polymerase activation for 5 min at 94°C, 40 cycles at 94°C for 30 s, 55°C for 60 s and 72°C for 60 s, and final extension for 5 min at 72°C. The RT-PCR products were then evaluated by electrophoresis using 1.5% agarose gel. The expected product sizes were 378 bp for the M gene, 311 bp for the H5-HA gene, and 252 bp for the H9-HA gene, respectively.

Results

Primer design and optimization of RT-LAMP assays
For RT-LAMP, primers were designed for the M, H5-HA, and H9-HA genes. The six primer sets for each target gene comprised two outer primers (F3 and B3), two inner primers (FIP and BIP), and two loop primers (LF and LB). In this study, in vitro–transcribed full-length RNA fragments of the M, H5-HA, and H9-HA genes were used as templates. Two to three different primer sets for each gene were tested and the sets that demonstrated the best amplification performance were selected for further analysis (Supplementary Fig. 1). The sequence information of the final RT-LAMP primer sets is available in Table 1. All selected primers in this study were designed within a conserved region of AIVs isolated from diverse poultry (Fig. 1). To minimize nonspecific amplification, we tested four different combinations of NMF and IBA in the reaction mixture (0.2 M NMF + IBA, 0.4 M NMF + IBA, 0.6 M NMF + IBA, 0.8 M NMF + IBA) using the H9-HA RT-LAMP set. When we applied a specific target template (H9-HA), all four reaction conditions showed significant amplification signals within 20 min regardless of the combinations of the additives (Fig. 2). When we applied a negative control template, the reaction mixtures with 0.4, 0.6, or 0.8 M NMF + IBA did not show any nonspe-

<table>
<thead>
<tr>
<th>Gene</th>
<th>Primer</th>
<th>Length (bp)</th>
<th>Sequence (5’-3’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>F3</td>
<td>19</td>
<td>GCATCGGTCTCACAGACAG</td>
</tr>
<tr>
<td></td>
<td>B3</td>
<td>19</td>
<td>ACTGGAGCTAGGGTGAGTT</td>
</tr>
<tr>
<td></td>
<td>FIP&lt;sup&gt;4&lt;/sup&gt;</td>
<td>46</td>
<td>CAGCGCTCGTCCATAGCCTTTTTTCCACAAAAACCACTAATCAGG</td>
</tr>
<tr>
<td></td>
<td>BIP&lt;sup&gt;4&lt;/sup&gt;</td>
<td>42</td>
<td>CAGCGGAAGCCTTGAGGAGGTITTTTTCCACAAAAATGGGTCTCTACGCC</td>
</tr>
<tr>
<td></td>
<td>LF</td>
<td>19</td>
<td>GGGCGACCCACTTCTGTCTTT</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>18</td>
<td>AGGCTAGGCGAGATGTTG</td>
</tr>
<tr>
<td>H5-HA</td>
<td>F3</td>
<td>23</td>
<td>GCATAAAAATTGTCAAGAAG</td>
</tr>
<tr>
<td></td>
<td>B3</td>
<td>19</td>
<td>ACTATTCTGAGTGCCAGTC</td>
</tr>
<tr>
<td></td>
<td>FIP&lt;sup&gt;4&lt;/sup&gt;</td>
<td>52</td>
<td>CGCMCTATTGAGGTTTGGACAATTTTTTTCACAAATATGAAAAATGGA</td>
</tr>
<tr>
<td></td>
<td>BIP&lt;sup&gt;4&lt;/sup&gt;</td>
<td>52</td>
<td>TCTAGATGACTTCTACCCCAATAATATTTGCAAGAGACTTTTTGTGATT</td>
</tr>
<tr>
<td></td>
<td>LF</td>
<td>20</td>
<td>GTTGGCAAGTGCCATACTC</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>21</td>
<td>ACCCTCTACATCGGGGAAT</td>
</tr>
<tr>
<td>H9-HA</td>
<td>F3</td>
<td>19</td>
<td>GAGAACTCTGAAGACCGAT</td>
</tr>
<tr>
<td></td>
<td>B3</td>
<td>19</td>
<td>CACTCTACCTGAATGAGCA</td>
</tr>
<tr>
<td></td>
<td>FIP&lt;sup&gt;4&lt;/sup&gt;</td>
<td>48</td>
<td>GTGGAAATGCGATGTCTGATTTCCAATTTTTCAATAGTGCGGCAACTGCG</td>
</tr>
<tr>
<td></td>
<td>BIP&lt;sup&gt;4&lt;/sup&gt;</td>
<td>46</td>
<td>AGCAGATGCTTGGGAACTTTTGGCACAATTCTCTAGAGCGACACT</td>
</tr>
<tr>
<td></td>
<td>LF</td>
<td>20</td>
<td>CCTCTCTACGTGGAGAG</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>25</td>
<td>GTTGGAGTCAGAGCTTCCAAACCTGG</td>
</tr>
</tbody>
</table>

M, matrix; HA, hemagglutinin; AIV, avian influenza virus; RT-LAMP, reverse transcription loop-mediated isothermal amplification.  
<sup>4</sup>Each inner primer (FIP and BIP) of RT-LAMP had two binding regions (F1c + F2 and B1c + B2, respectively) connected by a TTTT spacer.

https://doi.org/10.5808/GI.2020.18.4.e40

3 / 8

Genomics & Informatics 2020;18(4):e40
cific signals during the whole process (60 min); however, the reaction containing 0.2 M NMF + IBA showed a subtle nonspecific signal after 50 min (Fig. 2). Therefore, we decided to add 0.4 M NMF + IBA for all three RT-LAMP reactions.

### Specificity of the RT-LAMP assays

To evaluate the specificity of the three RT-LAMP assays, we applied matched and non-matched template RNA samples including a negative control (no template). The RT-LAMP assay for the universal M target showed an amplification product with the M RNA template, but no amplification product was detected with the HA RNA templates (H5 and H9) or the negative control template (Fig. 3A). The RT-LAMP assays for H5-HA and H9-HA showed target-specific amplification products without any
Fig. 2. Optimization of the combinations of additives (NMF and IBA) in the reverse transcription loop-mediated isothermal amplification reaction mixture. (A) 0.2 M NMF + IBA. (B) 0.4 M NMF + IBA. (C) 0.6 M NMF + IBA. (D) 0.8 M NMF + IBA. The x-axis represents the time for RT-LAMP reaction; the y-axis represents the relative fluorescence signal. NMF, N-methylformamide; IBA, isobutylamide; NTC, negative control; RT-LAMP, reverse transcription loop-mediated isothermal amplification.

Fig. 3. Specificity of the RT-LAMP assays. (A) The RT-LAMP products of the RT-LAMP assay for M gene were electrophoresed with 1.5% agarose gel: lane 1, negative control; lane 2, synthesized M RNA template; lane 3, synthesized H5-HA RNA template; lane 4 synthesized H9-HA RNA template; lane M, 100-bp DNA marker. (B, C) RT-LAMP products for H5-specific and H9-specific HA genes were electrophoresed with 1.5% agarose gel, respectively: lane 1, negative control; lane 2, synthesized H5-HA RNA template; lane 3, synthesized H9-HA RNA template; lane M, 100-bp DNA marker. RT-LAMP, reverse transcription loop-mediated isothermal amplification; M, matrix; HA, hemagglutinin.
cross-reaction with non-matched templates or the negative control (Fig. 3B and 3C).

Sensitivity of the RT-LAMP assays
To evaluate the sensitivity of the assays, we examined the detection limits of the three RT-LAMP assays. To do so, the template RNAs were 10-fold serially diluted (ranging from 10 to $10^5$ RNA copies) and applied for each real-time RT-LAMP assay. The detection limit of the M gene and the H5-HA RT-LAMP assays was found to be 100 copies/reaction, while the detection limit of the H9-HA RT-LAMP assay was 1,000 copies/reaction (Fig. 4A-4C). For objective validation of the sensitivity of our RT-LAMP assay, we performed conventional RT-PCR with a certified commercial AIV detection kit (AIV Multi-tube RT-PCR Kit, iNtRON Biotechnology) using the same template RNAs used for the RT-LAMP assays and compared the results. Our RT-LAMP assay was 10 times more sensitive than the reference RT-PCR kit. The detection limit of the RT-PCR assays for the M and H5-HA genes was 1,000 copies/reaction, and that for the H9-HA gene was 10,000 copies/reaction (Fig. 4D). Regarding the reaction time, the fluorescence signal of target-specific amplification appeared within 20 min in all three RT-LAMP assays when $10^5$ RNA copies were used. In the RT-LAMP assay for the M gene, even with 100 copies of the RNA template (the limit of detection), the amplification signal appeared within 20 min.

Discussion
Recently emerging viral infectious diseases, including AIV, are increasing, posing a major threat to both public health and poultry farming. POCT is a new concept of laboratory testing that enables testing to be performed where an infection occurs without transporting the samples to central clinical laboratories [7]. LAMP technology is ideal for POCT due to its ease of performance without the need for sophisticated equipment or experts to operate it [13]. In particular, RT-LAMP can be useful to identify RNA viruses at remote locations where laboratory equipment is limited [12,13]. In this study, we aimed to develop rapid and highly sensitive assays for identifying AIVs through POCT in the context of monitoring AIV epidemic outbreaks. Using LAMP technology, we developed RT-LAMP assays that can detect the presence of AIV and distinguish the H5 and H9 subtypes within 20-30 minutes. Furthermore, all three RT-LAMP assays (M, H5, and H9) were approximately 10-fold more sensitive than the approved ref-

![Fig. 4. Sensitivity of RT-LAMP and RT-PCR. The template RNAs, M (A), H5-HA (B), and H9-HA (C), were 10-fold serially diluted (ranging from 10 to $10^5$ RNA copies) and applied for each real-time RT-LAMP assay. The x-axis represents the time for the RT-LAMP reaction; the y-axis represents the relative fluorescent signal. (D) The same template RNAs were applied for conventional RT-PCR using a certified commercial AIV detection kit (AIV Multi-tube RT-PCR Kit, iNtRON Biotechnology) and the products were electrophoresed with 1.5% agarose gel. RT-LAMP, reverse transcription loop-mediated isothermal amplification; RT-PCR, reverse transcription polymerase chain reaction; M, matrix; HA, hemagglutinin; AIV, avian influenza virus; NTC, negative control.](https://doi.org/10.5808/GI.2020.18.4.e40)
We first designed an AIV universal RT-LAMP assay by targeting the M gene, which has a common sequence across AIV subtypes. Therefore, a positive M RT-LAMP assay result indicated the presence of AIV in the sample. We also designed RT-LAMP primer sets specifically targeting H5-HA and H9-HA, as these are typical subtypes of HPAIVs and LPAIVs, respectively.

All RT-LAMP primers were designed within a conserved region of AIVs isolated from diverse poultry across the world, suggesting that our RT-LAMP assays can be suitable for determining the presence of H5-HA and H9-HA regardless of the poultry or location.

After designing the RT-LAMP primers, we used in vitro transcribed full-length RNA fragments of the M, H5-HA, and H9-HA genes for experimental validation, because field-isolated AIV samples were not available due to their high pathogenicity. Due to the very high level of amplification efficiency of the LAMP, false detection is a major concern with this technology. In this study, we used additives (NMF and IBA) to minimize nonspecific amplification [23] and checked the real-time amplification signal by adding a fluorescent reagent. Finally, we selected 0.4 M NMF + IBA because this condition showed an efficient target-specific amplification signal without any nonspecific amplification signal. Moreover, none of the three RT-LAMP assays showed any nonspecific reactions with non-matched template or negative controls. These results suggest that our RT-LAMP assays are specific enough to identify AIV subtypes.

When we checked the sensitivity, the detection limits of our RT-LAMP assays were 100 to 1,000 RNA copies per reaction, which were 10 times more sensitive than the approved reference RT-PCR assays (1000 to 10,000 RNA copies per reaction). These results are consistent with previous studies reporting that LAMP assays had higher sensitivity than PCR-based assays [18,19]. In particular, the reaction time of our RT-LAMP assay was less than 30 minutes, which was approximately four times quicker than that of the conventional RT-PCR assay. These results suggest that our RT-LAMP assays can identify AIV subtypes 10 times more sensitively and four times more quickly than the conventional RT-PCR assays.

In summary, the RT-LAMP assays targeting the M, H5-HA, and H9-HA genes developed in this study demonstrated a high level of sensitivity and specificity. The assays could successfully detect the existence of AIV and discriminate between the H5 and H9 subtypes with higher sensitivity and less time than the conventional RT-PCR assays. This method could be a useful POCT tool for the rapid identification of AIV infections in the field.

ORCID

Songzi Zhang: https://orcid.org/0000-0003-4641-499X
Juyoun Shin: https://orcid.org/0000-0001-9207-7053
Sun Shin: https://orcid.org/0000-0002-2146-4336
Yeun-Jun Chung: https://orcid.org/0000-0002-6943-5948

Authors’ Contribution

Conceptualization: SS, JS, YJC. Data curation: SZ, SS, YJC. Formal analysis: SZ, SS. Funding acquisition: YJC. Methodology: SZ, JS, SS. Writing - original draft: SZ, SS. Writing - review & editing: YJC.

Conflicts of Interest

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

Acknowledgments

This work was supported by a grant from the National Research Foundation of Korea (2017R1E1A1A01074913 and 2017M3C9A6047615). We thank KREONET (Korea Research Environment Open NETwork) and KISTI (Korea Institute of Science and Technology Information) for allowing us to use their network infrastructure.

Supplementary Materials

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

References


Introduction

The problem of analyzing symbolic sequences appears in many areas of research, such as “big data” [1] and “dynamic systems” [2]. The most significant example of a symbolic sequence is nucleotide sequence. Moreover, a nucleotide sequence is an interesting and important mathematical object. Of special importance is the task of clustering nucleotide sequences [3-6]. A nucleotide sequence is hereby referred to as a sequence whose elements assume the values A, C, G, T. First the mathematical analysis of nucleotide sequences was suggested by the physicist Gamow in 1954 [7]. The problem of symbols relation of nucleotide sequences was first discussed by the physicist Yockey in the 1950s [8]. About 50 years later, in 2003, the mathematician Gelfand noted that “the use of mathematics in studying gene sequences is an adequate language” [9]. This implied the finding of formal (mathematical) properties of gene nucleotide sequences. Yet, insufficient attention has been paid to this subject.

The main method of investigating numeric sequences (or discrete numerical time series) is the construction and analysis of autocorrelation functions. However, the principal difference of numeric sequences from nucleotide sequences is that the nucleotides in the sequence take the symbolic values A, C, G, T. This means that statistical apparatus cannot be used for the analysis of such sequences, insofar as statistics does not have theoretically justified measures of correlation between symbolic (discrete) random variables. The impossibility of utilizing theoretically justified statistical methods in genetics has been noted earlier [10]. Therefore, information theory, having a solid theoretical justification, has

An information–theoretical analysis of gene nucleotide sequence structuredness for a selection of aging and cancer-related genes

David Blokh¹, Joseph Gitarts², Ilia Stambler³*

¹C.D. Technologies Ltd., Beer Sheba 8445914, Israel
²Efi Arazi School of Computer Science, Interdisciplinary Center, Herzliya 4673304, Israel
³Department of Science, Technology and Society, Bar Ilan University, Ramat Gan 5290002, Israel

We provide an algorithm for the construction and analysis of autocorrelation (information) functions of gene nucleotide sequences. As a measure of correlation between discrete random variables, we use normalized mutual information. The information functions are indicative of the degree of structuredness of gene sequences. We construct the information functions for selected gene sequences. We find a significant difference between information functions of genes of different types. We hypothesize that the features of information functions of gene nucleotide sequences are related to phenotypes of these genes.

Keywords: gene sequence, gene structuredness, information function, information theory, normalized mutual information
been increasingly used in the study of biological data. Earlier we have applied information theory to analyze data on aging-related diseases [11,12], including cancer [13-15]. The approach described in [13] is presented in the monograph [16].

An overview of the use of information theory for the analysis of biological sequences, in particular DNA sequences, has been presented earlier [17-19]. In the work by Li (1990) [20], mutual information was first used as a measure of correlation for autocorrelation symbolic sequence function. However, mutual information is a non-normalized value, and therefore it does not allow the researchers, in the general case, to compare different mutual information functions for different symbolic sequences.

The present work, for the first time, uses normalized mutual information as a measure of correlation to construct an autocorrelation function for the symbolic (nucleotide) sequence. Hence, we will refer to this function as information function. The use of normalized mutual information allows us to compare information functions of any symbolic sequences. The present article presents an algorithm for distributing sets of genes according to their information functions, that is, according to the interconnection between elements in the nucleotide sequences of these genes. Each value of the information function estimates the interconnection between elements of a nucleotide sequence with a corresponding lag. The set of all the values of the information function provides an estimate for the interconnection of the elements in a nucleotide sequence with all the lags, that is to say, it provides an estimate of the degree of structuredness of that sequence.

It may be hypothesized that genes with “close” information functions may produce similar phenotypes, and the proposed approach may help reveal unknown phenotypic properties of genes according to their nucleotide sequences.

Methods

Gene sequences

To illustrate the algorithm of distribution, we consider the nucleotide sequences of 14 genes. Table 1 lists the genes and their sizes as the number of nucleotides. The data on the genes’ sequences were obtained from NCBI Gene database (https://www.ncbi.nlm.nih.gov/gene).

The genes used in this study—BCL2, mTOR, FOXO3, FOXO1, IGF1, BRCA2, BRCA1, Klotho, Sirtuin 1, p16, BECN1, CCND1, Sirtuin 6, APOE—were selected for the most part insofar as these genes are often recognized as being involved in aging processes and often constitute networks in aging-associated pathways [21,22]. Hence both their phenotypic properties and their possible mutual relation could be suggested.

Mathematical analysis

Let X be a symbolic random variable with a distribution function as follows:

\[ X: x_1, x_2, \ldots, x_n \]
\[ P: p_1, p_2, \ldots, p_n \]

Entropy of random variable X is as follows:

\[ H(X) = -\sum_{i=1}^{n} p_i \log p_i. \]

Let X and Y be symbolic random variables. The mutual information between the variables X and Y is as follows:

\[ I(X; Y) = H(X) + H(Y) - H(XY), \]

where \( H(XY) \) is the entropy of the product of the random variables X and Y.

Let X and Y be symbolic random variables. The normalized mutual information (also termed “uncertainty coefficient”) is as follows:

\[ C(X; Y) = \frac{I(X; Y)}{H(Y)} = \frac{H(X) + H(Y) - H(XY)}{H(Y)}. \]

The normalized mutual information has the following properties.

1. \( 0 \leq C(X;Y) \leq 1 \).
2. \( C(X;Y) = 0 \) if and only if the random variables X and Y are independent (no correlation between the variables).
3. \( C(X;Y) = 1 \) if and only if there is a functional relation (correlation or influence) between X and Y.

Table 1. Genes used for the construction of information functions and their sizes as the number of nucleotides

<table>
<thead>
<tr>
<th>No.</th>
<th>Gene</th>
<th>Gene size (No. of nucleotides)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BCL2</td>
<td>196,935</td>
</tr>
<tr>
<td>2</td>
<td>mTOR</td>
<td>166,967</td>
</tr>
<tr>
<td>3</td>
<td>FOXO3</td>
<td>124,947</td>
</tr>
<tr>
<td>4</td>
<td>FOXO1</td>
<td>110,934</td>
</tr>
<tr>
<td>5</td>
<td>IGF1</td>
<td>85,980</td>
</tr>
<tr>
<td>6</td>
<td>BRCA2</td>
<td>84,193</td>
</tr>
<tr>
<td>7</td>
<td>BRCA1</td>
<td>81,189</td>
</tr>
<tr>
<td>8</td>
<td>Klotho</td>
<td>50,083</td>
</tr>
<tr>
<td>9</td>
<td>SirT1</td>
<td>33,722</td>
</tr>
<tr>
<td>10</td>
<td>p16</td>
<td>27,292</td>
</tr>
<tr>
<td>11</td>
<td>BECN1</td>
<td>14,185</td>
</tr>
<tr>
<td>12</td>
<td>CCND1</td>
<td>13,370</td>
</tr>
<tr>
<td>13</td>
<td>SirT6</td>
<td>8,496</td>
</tr>
<tr>
<td>14</td>
<td>APOE</td>
<td>3,647</td>
</tr>
</tbody>
</table>

Note: In the first four genes, the numbers of nucleotides exceed 100,000.
Let \( x(n) = (x(1), x(2), \ldots, x(n), \ldots) \) represent discrete time series having symbolic values.
Let \( x(n+j) = (x(1+j), x(2+j), \ldots, x(n+j), \ldots) \) be a time series \( x(n) \) with a lag.
The auto-mutual information of the time series \( x(n) \) with a lag equals:
\[
I(x(n); x(n+j)) = H(x(n)) + H(x(n+j)) - H(x(n), x(n+j)).
\]
The normalized auto-mutual information of the time series \( x(n) \) with a lag equals:
\[
C(x(n); x(n+j)) = \frac{I(x(n); x(n+j))}{H(x(n+j))} = \frac{H(x(n)) + H(x(n+j)) - H(x(n), x(n+j))}{H(x(n+j))}.
\]

The normalized mutual information \( C(x(n); x(n+j)) \) is then calculated as a function of the lag.
We shall refer to function \( F(j) = C(x(n); x(n+j)) \) as the information function of the discrete time series \( x(n) \).
Properties of the information function \( F(j) \) are as follows.
(1) \( 0 \leq F(j) \leq 1 \).
(2) \( F(j) = 0 \) if and only if \( x(n) \) and \( x(n+j) \) are mutually independent.
(3) \( F(j) = 1 \) if and only if there exists a functional relationship between \( x(n) \) and \( x(n+1) \).

Let \( \{x1(n), x2(n), \ldots, xk(n)\} \) be a set of discrete time series, whose elements are symbols, e.g. gene nucleotide sequences, \( n = 1, 2, 3, \ldots \), and the maximum value \( n \) for a sequence \( x(n) \) equals the number of elements in this nucleotide sequence.
The algorithm of distributing a set of time series \( \{x1(n), x2(n), \ldots, xk(n)\} \) consists of three procedures: (1) construction of an information function matrix; (2) ranking of columns of the information function matrix; and (3) application of a multiple comparisons method.

**Construction of an information function matrix**
For each time series \( x_(i) (n) \) \( 1 \leq i \leq k \), we construct the information function as follows:
\[
F(i)(j) 1 \leq i \leq k, 1 \leq j \leq m, \text{where} m \text{is the number of lags in the information function.}
\]
We obtain the \( k \times m \) \( [F(i)(j)] \) matrix of values of the information functions, i.e., a matrix where each row is an information function of the corresponding time series.

**Ranking of columns of the information function matrix**
Each row of \( [F(i)(j)] \) matrix is an information function of time series, and each column contains the values of information functions corresponding to the same lag.
For each column of \( [F(i)(j)] \) matrix, we rank its entries and assign the rank 1 to the smallest entry of the column. We obtain \( k \times m \) matrix of ranks \( [r_i(j)] \), with each column of the matrix containing ranks from 1 to \( k \).
We estimate the element interconnection of the \( i \)-th time series as compared to the element interconnection of other time series by the sum of all the elements of \( i \)-th row of the matrix \( [r_i(j)] \). Such an estimation allows us to use multiple comparisons of rank statistics for the comparison of time series interconnection.

**Application of a multiple comparisons method**
We compare rank sums using the Newman-Keuls test [23]. This test provides adequate results in the analysis of biomedical data, including aging-related multimorbidity [11,12], and is appropriate for the present problem.

**Results**

**The values and clustering of gene information functions**
Following the above algorithm for distributing a set of time series, we calculate and cluster the values of gene information functions, as follows.
(1) For each gene, out of the 14 genes under consideration, we calculate the information function with 12 lags. We obtain the information functions matrix \( [F(i)(j)] 1 \leq i \leq 14, 1 \leq j \leq 12 \) (Table 2).
(2) We rank the entries of each column of the information function matrix, while attributing to the least values the rank 1. We obtain the rank matrix \( [r_i(j)] 1 \leq i \leq 14, 1 \leq j \leq 12 \) (Table 3).
Let us consider Table 3 as the Friedman statistical model [24] and examine the row effect of this table.

**Hypotheses:**
H0: There is no row effect (“null hypothesis”).
H1: The null hypothesis is invalid.

**Critical range:**
The sample is “large”, therefore, the critical range is the upper 1%-range of \( \chi^2_{11} \) distribution.
Let us calculate the \( \chi^2 \)-criterion. This gives us \( \chi^2 = 91.65 \). The critical range is \( \chi^2_{11} > 27.69 \). Since \( 91.65 > 27.69 \), the null hypothesis with respect to Table 3 is rejected. Thus, according to the Friedman test, the row effect has been found. Hence, there is a difference between the rows under consideration.

For multiple comparisons, we use the Newman-Keuls test. We obtain \( /R_i - R_{i+1}/ > 8.93 \), where \( R_i \) and \( R_{i+1} \) are elements of the column “Sum of ranks” in the \( i \)-th and \( (i+1) \)-th rows of Table 3, respectively. By multiple comparisons, we construct the clustering shown in Table 4.
### Table 2. Normalized mutual information

<table>
<thead>
<tr>
<th>Gene</th>
<th>Lag 1</th>
<th>Lag 2</th>
<th>Lag 3</th>
<th>Lag 4</th>
<th>Lag 5</th>
<th>Lag 6</th>
<th>Lag 7</th>
<th>Lag 8</th>
<th>Lag 9</th>
<th>Lag 10</th>
<th>Lag 11</th>
<th>Lag 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>APOE</td>
<td>0.027149</td>
<td>0.002646</td>
<td>0.00367</td>
<td>0.002773</td>
<td>0.003221</td>
<td>0.010847</td>
<td>0.003922</td>
<td>0.00447</td>
<td>0.004731</td>
<td>0.004429</td>
<td>0.004188</td>
<td>0.003196</td>
</tr>
<tr>
<td>BRCA1</td>
<td>0.022034</td>
<td>0.004159</td>
<td>0.003791</td>
<td>0.002832</td>
<td>0.002345</td>
<td>0.004025</td>
<td>0.00123</td>
<td>0.00346</td>
<td>0.002544</td>
<td>0.002233</td>
<td>0.002244</td>
<td>0.002206</td>
</tr>
<tr>
<td>p16</td>
<td>0.011972</td>
<td>0.007661</td>
<td>0.003112</td>
<td>0.003958</td>
<td>0.001811</td>
<td>0.004812</td>
<td>0.002174</td>
<td>0.003145</td>
<td>0.00315</td>
<td>0.003304</td>
<td>0.001438</td>
<td>0.002992</td>
</tr>
<tr>
<td>Sirtuin 6</td>
<td>0.033637</td>
<td>0.00474</td>
<td>0.002757</td>
<td>0.001858</td>
<td>0.002086</td>
<td>0.006216</td>
<td>0.001017</td>
<td>0.003024</td>
<td>0.002727</td>
<td>0.002613</td>
<td>0.000491</td>
<td>0.002431</td>
</tr>
<tr>
<td>Sirtuin 1</td>
<td>0.016229</td>
<td>0.004515</td>
<td>0.003718</td>
<td>0.002581</td>
<td>0.002479</td>
<td>0.003866</td>
<td>0.001941</td>
<td>0.002843</td>
<td>0.002343</td>
<td>0.001934</td>
<td>0.002365</td>
<td>0.002714</td>
</tr>
<tr>
<td>mTOR</td>
<td>0.025724</td>
<td>0.003696</td>
<td>0.002441</td>
<td>0.001687</td>
<td>0.001643</td>
<td>0.002398</td>
<td>0.001071</td>
<td>0.002092</td>
<td>0.001761</td>
<td>0.001208</td>
<td>0.000958</td>
<td>0.001191</td>
</tr>
<tr>
<td>BCL2</td>
<td>0.024063</td>
<td>0.004531</td>
<td>0.001164</td>
<td>0.002247</td>
<td>0.00083</td>
<td>0.002089</td>
<td>0.00097</td>
<td>0.001539</td>
<td>0.000749</td>
<td>0.00125</td>
<td>0.000493</td>
<td>0.001013</td>
</tr>
<tr>
<td>BECN1</td>
<td>0.025643</td>
<td>0.003635</td>
<td>0.00203</td>
<td>0.001389</td>
<td>0.001729</td>
<td>0.002412</td>
<td>0.001329</td>
<td>0.001668</td>
<td>0.002501</td>
<td>0.002151</td>
<td>0.001305</td>
<td>0.001334</td>
</tr>
<tr>
<td>BRCA2</td>
<td>0.019161</td>
<td>0.003686</td>
<td>0.001769</td>
<td>0.001204</td>
<td>0.00096</td>
<td>0.001903</td>
<td>0.001236</td>
<td>0.001632</td>
<td>0.0019</td>
<td>0.001326</td>
<td>0.00085</td>
<td>0.001334</td>
</tr>
<tr>
<td>CCND1</td>
<td>0.021874</td>
<td>0.006757</td>
<td>0.00247</td>
<td>0.001743</td>
<td>0.00188</td>
<td>0.004093</td>
<td>0.002921</td>
<td>0.001539</td>
<td>0.002893</td>
<td>0.00269</td>
<td>0.000207</td>
<td>0.000329</td>
</tr>
<tr>
<td>FOXO1</td>
<td>0.021496</td>
<td>0.003354</td>
<td>0.001571</td>
<td>0.001381</td>
<td>0.001337</td>
<td>0.001946</td>
<td>0.00097</td>
<td>0.001452</td>
<td>0.001385</td>
<td>0.001303</td>
<td>0.00074</td>
<td>0.001029</td>
</tr>
<tr>
<td>FOXO3</td>
<td>0.023727</td>
<td>0.004032</td>
<td>0.00149</td>
<td>0.001242</td>
<td>0.001189</td>
<td>0.001845</td>
<td>0.001215</td>
<td>0.001161</td>
<td>0.00133</td>
<td>0.001025</td>
<td>0.000623</td>
<td>0.00093</td>
</tr>
<tr>
<td>IGF1</td>
<td>0.026729</td>
<td>0.003</td>
<td>0.00673</td>
<td>0.00128</td>
<td>0.000587</td>
<td>0.000979</td>
<td>0.00054</td>
<td>0.001135</td>
<td>0.000639</td>
<td>0.00051</td>
<td>0.000284</td>
<td>0.00048</td>
</tr>
<tr>
<td>Klotho</td>
<td>0.020122</td>
<td>0.00376</td>
<td>0.001905</td>
<td>0.001739</td>
<td>0.001148</td>
<td>0.001732</td>
<td>0.001069</td>
<td>0.001482</td>
<td>0.001822</td>
<td>0.001409</td>
<td>0.001414</td>
<td>0.001559</td>
</tr>
</tbody>
</table>

### Table 3. Table of ranks: ranking by columns

<table>
<thead>
<tr>
<th>Gene</th>
<th>Lag 1</th>
<th>Lag 2</th>
<th>Lag 3</th>
<th>Lag 4</th>
<th>Lag 5</th>
<th>Lag 6</th>
<th>Lag 7</th>
<th>Lag 8</th>
<th>Lag 9</th>
<th>Lag 10</th>
<th>Lag 11</th>
<th>Lag 12</th>
<th>Sum of ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>APOE</td>
<td>13</td>
<td>1</td>
<td>11</td>
<td>11</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>148</td>
</tr>
<tr>
<td>BRCA1</td>
<td>7</td>
<td>9</td>
<td>13</td>
<td>12</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td>13</td>
<td>10</td>
<td>12</td>
<td>9</td>
<td>124</td>
<td></td>
</tr>
<tr>
<td>p16</td>
<td>1</td>
<td>14</td>
<td>10</td>
<td>14</td>
<td>9</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>13</td>
<td>10</td>
<td>121</td>
<td></td>
</tr>
<tr>
<td>Sirtuin 6</td>
<td>14</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>13</td>
<td>12</td>
<td>2</td>
<td>10</td>
<td>116</td>
<td></td>
</tr>
<tr>
<td>Sirtuin 1</td>
<td>2</td>
<td>10</td>
<td>12</td>
<td>10</td>
<td>13</td>
<td>11</td>
<td>11</td>
<td>13</td>
<td>11</td>
<td>8</td>
<td>13</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>mTOR</td>
<td>11</td>
<td>6</td>
<td>8</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>BCL2</td>
<td>9</td>
<td>11</td>
<td>2</td>
<td>9</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>5.5</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>59.5</td>
<td></td>
</tr>
<tr>
<td>BECN1</td>
<td>10</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>BRCA2</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>CCND1</td>
<td>6</td>
<td>13</td>
<td>14</td>
<td>13</td>
<td>10</td>
<td>11</td>
<td>13</td>
<td>5.5</td>
<td>11</td>
<td>11</td>
<td>14</td>
<td>132.5</td>
<td></td>
</tr>
<tr>
<td>FOXO1</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>FOXO3</td>
<td>8</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>IGF1</td>
<td>12</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Klotho</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td>89</td>
<td></td>
</tr>
</tbody>
</table>
The obtained clustering possesses the following properties: (1) For two neighboring sets of Table 4, the smallest element of one set and the greatest element of another set located nearby are significantly different (\(a_T = 0.01\)); (2) Elements belonging to the same set do not differ from each other (\(a_T = 0.01\)).

Note that the differences between cluster 1 (APOE gene) and all the other elements (genes) are statistically significant (\(a_T = 0.01\)). The same holds true for cluster 3 (BECN1 gene), cluster 4 (mTOR gene), and cluster 7 (IGF1 gene).

### The significance of gene information functions

The domain of the information functions under consideration is the set \{Lag 1, Lag 2, Lag 3, …, Lag 12\}, and the values are the set of real numbers 0 to 1. We perform the comparative analysis of the values of information functions on the domain of those functions.

In Table 2, each row represents the values of the information function of a corresponding gene. We rank the values of each row of Table 2, attributing rank 1 to the least value. We obtain Table 5.

We evaluate the values of the information functions in Lag \(j\) as the sum of elements of the column Lag \(j\) of Table 5. Let us consider Table 5 as the Friedman statistical model, and examine the column effect of this table.

**Hypotheses:**

H0: There is no column effect (“null hypothesis”).

H1: The null hypothesis is invalid.

**Critical range:**

The sample is “large”, therefore, the critical range is the upper 1%-range of \(\chi^2\) distribution.

Let us calculate the \(\chi^2\)-criterion. This gives us \(\chi^2 = 121.5\). The critical range is \(\chi^2 > 24.73\). Since 121.5 > 24.73, the null hypothesis with respect to Table 4 is rejected. Thus, according to the Friedman test, the column effect has been found. Hence, there is a difference between the columns under consideration.

For multiple comparisons, we use the Newman-Keuls test. We obtain \(|R_i - R_{i+1}| > 9.64\), where \(R_i\) and \(R_{i+1}\) are elements of the column “Sum of ranks” in the \(i\)-th and \((I+1)\)-th rows of Table 5, respectively. By multiple comparisons, we construct the clustering shown in Table 6.

The obtained clustering possesses the following properties: (1) For two neighboring sets of Table 5, the smallest element of one set and the greatest element of another set located nearby are significantly different (\(a_T = 0.01\)); (2) Elements belonging to the same set do not differ from each other (\(a_T = 0.01\)).

### Table 4. Gene distribution according to sums of ranks

<table>
<thead>
<tr>
<th>No.</th>
<th>Cluster</th>
<th>Sub-cluster</th>
<th>Gene</th>
<th>Sum of ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cluster 1</td>
<td></td>
<td>APOE</td>
<td>147</td>
</tr>
<tr>
<td>2</td>
<td>Cluster 2</td>
<td>2.1</td>
<td>CCND1</td>
<td>132.5</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>2.2</td>
<td>p16</td>
<td>131</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>2.3</td>
<td>BRCA1</td>
<td>124</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>2.4</td>
<td>Sirtuin 1</td>
<td>122</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>2.5</td>
<td>Sirtuin 6</td>
<td>116</td>
</tr>
<tr>
<td>7</td>
<td>Cluster 3</td>
<td></td>
<td>BECN1</td>
<td>91.5</td>
</tr>
<tr>
<td>8</td>
<td>Cluster 4</td>
<td></td>
<td>mTOR</td>
<td>80</td>
</tr>
<tr>
<td>9</td>
<td>Cluster 5</td>
<td>5.1</td>
<td>Klotho</td>
<td>69</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>5.2</td>
<td>BRCA2</td>
<td>62.5</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>5.3</td>
<td>BCL2</td>
<td>59.5</td>
</tr>
<tr>
<td>12</td>
<td>Cluster 6</td>
<td>6.1</td>
<td>FOXO1</td>
<td>50</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>6.2</td>
<td>FOXO3</td>
<td>50</td>
</tr>
<tr>
<td>14</td>
<td>Cluster 7</td>
<td></td>
<td>IGF1</td>
<td>25</td>
</tr>
</tbody>
</table>

### Table 5. Table of ranks: ranking by rows

<table>
<thead>
<tr>
<th>Gene</th>
<th>Lag 1</th>
<th>Lag 2</th>
<th>Lag 3</th>
<th>Lag 4</th>
<th>Lag 5</th>
<th>Lag 6</th>
<th>Lag 7</th>
<th>Lag 8</th>
<th>Lag 9</th>
<th>Lag 10</th>
<th>Lag 11</th>
<th>Lag 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>APOE</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>5</td>
<td>10</td>
<td>1</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>BRCA1</td>
<td>12</td>
<td>11</td>
<td>5</td>
<td>9</td>
<td>2</td>
<td>10</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>p16</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td>3</td>
<td>4</td>
<td>11</td>
<td>2</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Sirtuin 6</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>10</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Sirtuin 1</td>
<td>12</td>
<td>11</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>9</td>
<td>2</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>mTOR</td>
<td>12</td>
<td>11</td>
<td>6</td>
<td>10</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>BCL2</td>
<td>12</td>
<td>11</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>BECN1</td>
<td>12</td>
<td>11</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>4</td>
<td>7</td>
<td>9</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>BRCA2</td>
<td>12</td>
<td>11</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>9</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>CCND1</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>10</td>
<td>2</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>FOXO1</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>5</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>FOXO3</td>
<td>12</td>
<td>11</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>IGF1</td>
<td>12</td>
<td>11</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>5</td>
<td>9</td>
<td>3</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Klotho</td>
<td>12</td>
<td>11</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>5</td>
<td>9</td>
<td>3</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Sum of ranks</td>
<td>168</td>
<td>153</td>
<td>117</td>
<td>94</td>
<td>52</td>
<td>129</td>
<td>39</td>
<td>90</td>
<td>101</td>
<td>62</td>
<td>27</td>
<td>60</td>
</tr>
</tbody>
</table>
Note that the differences between cluster 1 (Lag 1) and all the elements are statistically significant (\( \alpha_T = 0.01 \)). The same holds true for cluster 2 (Lag 2), cluster 3 (Lag 6), and cluster 4 (Lag 3).

The values of the information functions in Lag 6 are greater than the values of the information functions in Lag 3, Lag 4, and Lag 5. This signifies that, for the group of genes under consideration, the interconnection between elements distanced five elements from each other is greater than the interconnection between elements located closer together, namely distanced 2, 3, and 4 elements from each other.

### Discussion

In this work we established a novel information theory based method for the evaluation of the level of structuredness of gene sequences (information function) by the sequences’ normalized mutual information. This new method may serve as an additional structural evaluation tool for genomic analysis, and for omics biomarkers analysis generally. In the future, it may be possible to associate between the gene structuredness as evaluated by the present method and the expression and phenotype of particular genes under consideration. Here we, for the first time, describe the methodology to calculate the gene structuredness, while the association of the gene structuredness with gene expression and phenotypic function will be the task of future work.

Even though the present work only describes the methodology, some hypotheses may be advanced considering the possible association of the value of gene structuredness as shown here by the clustering (Table 4) with some known phenotypic properties of the selected genes considered in this study. Thus the genes APOE, BECN1, mTOR, and IGF1 each form a separate cluster according to their level of structuredness. This may indicate that each of these genes possesses properties not common for the other genes. The genes FOXO1 and FOXO3 are in the same cluster, which may be expected for the genes of the same group. Interestingly, the genes BRCA1 and BRCA2 are found in different clusters. As it has been demonstrated, the BRCA1 and BRCA2 genes are associated with different types of tumors, and this distinction may have been reflected in the information function (structuredness) of these genes [25,26].

Of special interest are clusters 2 and 5. Cluster 2 includes the genes Sirtuin 1 and Sirtuin 6, together with the genes CCND1, p16, and BRCA1. A special characteristic of all these five genes in cluster 2 is that under conditions of overexpression, these genes are associated with oncological diseases, though not necessarily under conditions of normal expression or under-expression [27-32]. On the other hand, a characteristic feature of cluster 5 is that all the three genes in this cluster—Klotho, BRCA2, and BCL2—under conditions of under-expression are associated with oncological diseases [33-35]. Yet, under normal expression or overexpression, such an association is not observed. Thus it may be hypothesized that the level of gene sequence structuredness, at least in the present gene selection, may be somehow associated with effects of extreme gene expression, either overexpression or under-expression. Yet, a clarification of such a hypothesis, as well as positing and testing additional hypotheses for a potential association of gene structure and function, will require further extensive investigation.

### ORCID

David Blokh: https://orcid.org/0000-0003-0403-063X
Joseph Gitarts: https://orcid.org/0000-0002-9793-7128
Ilia Stambler: https://orcid.org/0000-0003-4636-4742

### Authors’ Contribution

Conceptualization: DB, IS. Data curation: IS, JG. Formal analysis: DB, JG. Methodology: DB. Writing - original draft: DB, IS. Writing - review & editing: DB, IS.

### Conflicts of Interest

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

### References


https://doi.org/10.5808/GI.2020.18.4.e41
Introduction

Protein-binding regions in the context of chromatin have been detected by the chromatin immunoprecipitation (ChIP) method. Since the first ChIP coupled with high-throughput DNA sequencing (ChIP-Seq) technology for histone modification mapping was introduced with the combination of ChIP and next-generation sequencing, a large amount of ChIP-Seq data has been produced at the genome level, and the development of data analysis tools should thus be emphasized [1-3].

The basic building block of chromatin, the nucleosome, consists of 146 base pairs (bp) of DNA and a histone octamer composed of four core histones: H2A, H2B, H3, and H4. Post-translational modifications of histone tails play an important role in the epigenetic regulation of genome activity. These modifications include acetylation, methylation,
phosphorylation, and ubiquitination. Depending on the types of histone modifications and binding sites, different enrichment patterns and related biological effects are expected. For example, acetylated histones provide a chromatin environment easily accessible to the transcriptional machinery by changing the chromatin conformation. Some histone methylations, such as H3K4me2 and H3K4me3, are mostly located on promoters, whereas H3K36me3 is predominantly found on the gene bodies of transcriptionally active genes [4,5].

The Encyclopedia of DNA Elements (ENCODE) Consortium, aiming at the identification of all functional elements in the human genome, proposed a guideline for categorizing protein-bound regions occupied by point source factors, broad source factors, and mixed source factors [6].

The distribution patterns of ChIP-Seq data on the genome have been analyzed using many different software programs with specific algorithms, which use different strategies for searching potential binding regions, judging the peaks, and calculating significance [7-10]. Most previous studies have focused on detecting the enriched peaks, and several groups have already evaluated peak calling programs [11-16]. Although most previous studies compared the performance of each program for analyzing transcription factor binding patterns, some tested histone modifications, including H3K4me3, H3K9me3, H3K27me3, and H3K36me3 [11,12,14]. However, the performance evaluation of ChIP-Seq analysis programs needs to be more extensively examined to understand the nature of enrichment of various types of histone modifications. Herein, we tested ChIP-Seq data from 12 histone modifications covering three source types with five peak calling programs (CisGenome, MACS1, MACS2, PeakSeq, and SISSRs).

Methods

Data filtering and cross-correlation analysis
The ChIP-Seq datasets of 12 histone modification types, input, and RNA-sequencing of human embryonic stem cell line (H1) were downloaded from the NIH Roadmap Epigenomics Project Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/) (Supplementary Table 1). The downloaded SRA format files were converted to the FASTQ format via fastq-dump in SRA Toolkit (version 2.4.5). Raw sequencing reads were filtered by fastq_quality_filter (FASTX-Toolkit version 0.0.13.2) with the following options (-p 80, -q 20, and -Q33). High-quality reads were mapped to the human genome (hg19) using Bowtie (version 1.1.1) with the default options (-n 2, -e 70, -l 28, -i 0, -X 250, and -maxbts 250) [17].

To evaluate the signal-to-noise ratio of a ChIP-Seq experiment, strand cross-correlation analysis was performed using the SPP program with the default options (-s -100:5:600, and -x 10), considering two metrics: (1) the normalized strand coefficient, which quantifies the fragment length cross-correlation over the background cross-correlation rate, and (2) the relative strand correlation, which calculates the ratio of cross-correlation observed at the predicted fragment size against the artificial cross-correlation observed at the read length [18].

Identification of regions enriched with specific histone modifications
To detect peaks, CisGenome (version 2.0), MACS1 (version 1.4.2), MACS2 (version 2.1.0), PeakSeq (version 1.31), and SISSRs (version 1.4), were used with the default options and recommended parameters for a direct comparison without any optimization (Supplementary Table 2). For CisGenome, the Bowtie-format output files were converted into the aln format and the s证券 command was used. For MACS1, the options of --p 1e-5, -m 10:30, and --keep-dup 1 were used and for MACS2, the default options (-q 0.01, -m 5:50, and --keep dup 1) were applied. In MACS2, the broad options (-q 0.1, -m 5:50, and --keep-dup 1) were also used for the broad source peaks. The signal map was prepared from the Bowtie output using the PeakSeq -preprocess command. During the step of PeakSeq -peak_selection, the default options were used, such as Enrichment_mapped_fragment_length 200, target_FDR 0.05, N_Simulations 10, Minimum_interpeak_distance 200, and max_Qvalue 0.05. SISSRs detected peaks with the recommend options (-F 0.001, -e 10, -p 0.001, -m 0.8, -w 20, -E 2, and -L 500). All peaks in each set were ranked by the following guidelines: CisGenome and PeakSeq, pre-sorted peak lists; MACS1 and MACS2, sorted by the significance level (10 × 2log10(p-value)) and then by the fold enrichment; SISSRs, ranked by the fold enrichment and by the significance level (p-value). Frequently detected false positive peaks, regardless of cell line or experiment (called the ENCODE blacklist) were removed for quality control of peaks [19,20].

Comparison of peak calling performance
The coincidence of peak positions obtained by the individual programs was examined using the intersectBed and multiIntersectBed functions (BEDTools version 2.23.0) with a minimum overlapping size of 1 bp [21]. Pearson correlation coefficients based on peak ranks between overlapped peaks were calculated, because the peak rank represents the order of importance according to algorithm characteristics. For the multiple comparison analyses of each histone mark, we used multiIntersectBed in BEDTools. The multi-IntersectBed function provided a comparison among the multiple
files.

The Jaccard similarity coefficients (or index J) were calculated for the measurement of variability: \( J(A, B) = \frac{|A \cap B|}{|A \cup B|} \) where A and B are sets of enriched regions in base pairs identified by peak calling programs. Irreproducibility discovery rate (IDR) analysis with all replicates was performed using the recommended parameters (peak.half.width = -1, min.overlap.ratio = 0, is.broadpeak = F, and ranking.measure = p.value for MACS1 and MACS2; q.value for CisGenome and PeakSeq; signal.value for SISSRs) [22]. For the specificity test, the control sequence reads were mixed with the original ChIP-Seq data and then the performance was computed. At a different sequencing read depth, the genomic coverage of the enriched regions was calculated by genomeCoverageBed in BEDTools.

The genomic coverage of the regions was calculated by genomeCoverageBed in BEDTools by considering randomly selected reads (0.5, 0.75, 1.0, 2.5, 5.0, 7.5, 10, 15, 20, and 30 million). To detect enriched regions in subsampled data, the algorithms with the same parameters as in the above analysis were used. The specificity of the immunoprecipitated signals to nonspecific noise was examined by mixing the ChIP-Seq data with different noise levels (50% 100%, and 150% of control reads).

**Results**

**Overview of ChIP-Seq data analysis**

For the comparative analysis of ChIP-Seq peak calling programs, data on 12 types of histone modifications were initially filtered and only high-quality mappable reads were used for further analysis (Supplementary Table 3). The histone modification marks were grouped into narrow (4 histone modifications), broad (5), and mixed (3) sources according to the ENCODE guideline. Peaks were called by five commonly used programs [7-10] and their number, position, coverage, and specificity were compared individually. An overview of this study is summarized in Fig. 1.

**Concordance of peak regions**

The peaks representing the enrichment patterns of each histone modification were more affected by histone types than by peak calling programs. The peak counts of H3K4me3, H3K9ac, H3K27me3, and H3K56ac were similar in most peak calling programs except SISSRs. Peak lengths were strongly affected by the program used, with the average length varying from 57.7 to 1941.8 bp (Supplementary Table 4). Peaks from MACS2 with the broad option and PeakSeq covered a longer genomic region, while CisGenome, MACS1, MACS2 with the default, and SISSRs suggested relatively short regions as peaks. Notably, SISSRs identified the shortest peaks. The concordance or co-occupancy of peaks regions identified from two different callers were calculated at the same genomic loci. The peaks from H3K4me2, H3K4me3, H3K-9ac, H3K27me3, and H3K36me3 varied in length. As a representative example, the number of peaks enriched with H3K4me3, a typical narrow source mark, ranged from 24,000 to 37,000 and its enrichment profile was very similar at promoters of actively transcribed genes with all peak callers (Fig. 2A). The peak positional variability was highly dependent on the histone mark type. Histone marks such as H3K4me2, H3K4me3, H3K27ac, and H3K-9ac, which are associated with transcriptional activation, showed a high level of concordance. The overlapping ratio of H3K4ac and H3K79me1 was below 60% on average. Our results indicated that histone marks that covered narrow regions with high enrichment could be identified by any of the peak callers used in this study, but peak positions from broad source marks differed according to the peak calling algorithm.

![Fig. 1. Overview of the analysis.](https://doi.org/10.5808/GI.2020.18.4.e42)
The significance of the identified peaks was examined by Pearson correlation coefficients. To explore peak coherence among algorithms, the correlation coefficients of histone types were categorized (Fig. 2B). The highest correlations were obtained from the peaks of H3K4me2, H3K4me3, H3K9ac, and H3K27ac. The lowest group included H3K4ac, H3K4me1, H3K9me3, H3K36me3, and H3K79me1.

**Peak consistency between replicates**

The reproducibility of the peak calling algorithm across biological replicates was measured by considering the Jaccard similarity and the IDR [22]. The Jaccard similarity coefficients between replicates at a single base level were computed except for H3K27ac, for which the duplicated data set was not provided. H3K4me2, H3K4me3, H3K9ac, and H3K27me3 had high similarity between replicates in all peak callers except SISSRs (Fig. 3A). The mean values of the Jaccard similarity coefficients between H3K4me3 replicates were above 0.5. Interestingly, the similarity between H3K36me3 replicates was higher only in MACS2 with the broad option than in any other callers, which means that H3K36me3

---

**Fig. 2.** Pairwise comparison of shared regions. (A) Percentage of peaks recaptured by programs shown pairwise. Each panel shows the percentage of total peaks from one method (column) that was recaptured by another peak caller (row) after filtering blacklist peaks. (B) The concordance rate of peak regions derived from two peak callers. The ranked coincidence was calculated and the values of percentage and correlation coefficients were denoted after filtering blacklist peaks.
mark clearly belongs to the group of broad source marks. The consistency of other histone modifications was fairly low. Some significant histone type-dependent consistency was detected, such as for H3K4me1, H3K9me3, and H3K79me1 with MACS2 with the broad option; H3K4ac and H3K79me2 with PeakSeq; and H3K56ac with MACS2 with the broad option and PeakSeq.

According to the ENCODE guidelines [6], the IDR should be used for narrow peaks such as transcription factors, as well as for punctate chromatin marks such as H3K4me1, H3K4me3, H3K9ac, and H3K27ac. Considering the average number of peaks reducible in replicate pairs with an IDR threshold of 0.01%, the reproducibility of different peak callers was dependent on the histone type (Fig. 3B). H3K4me2 and H3K4me3 showed a relatively large number of reproducible peaks. The MACS1 program gave the most reproducible results across replicates in these histone modifications. The peaks identified from H3K4ac, H3K56ac, H3K79me1, and H3K79me2 seemed not to be reproducible.

**Peak coverage with different sequencing depths**

The importance of sequencing depth has been emphasized for measuring the experimental validity of ChIP-Seq. To assess the number of peaks at the level of sequencing read saturation by different peak calling algorithms, the peak calling procedure was repeatedly applied with different numbers of subsampled reads from the total number of sequencing reads (Fig. 4). The peak coverage of point source marks in all peak callers except SISSRs dramatically increased at a lower depth (≤ 2.5 million reads for H3K4me2 and ≤ 1 million reads for H3K4me3). Broad source marks like H3K9ac, H3K27ac, H3K27me3, and H3K36me3 needed more reads to reach the level of saturation and their coverage of enriched genomic regions was consistently increased at > 10 million reads. The size of enriched regions derived from MACS2 with a broad option generally covered larger loci than any other algorithms.

**Specificity of peak calling against the noise signal**

The peak specificity called by different algorithms was compared by mixing the ChIP-Seq reads with randomly-selected input control reads (50%, 100%, and 150% of the corresponding ChIP-Seq reads). The percentage of enriched regions recaptured by CisGenome, MACS1, and PeakSeq was not substantially affected by the noise level (Fig. 5). CisGenome and PeakSeq recaptured over 80% of the enriched regions even with noise reads for H3K4me1, H3K4me3, H3K9ac, H3K27me3, and H3K36me3. In particular, MACS2 was very responsive to the noise and the recaptured peak ratio fell down to the minimum level when it was tested with the H3K4ac, H3K56ac, and H3K79me1 marks. SISSRs had the lowest performance for peak recapturing. Interestingly, the recaptured peak ratio of the H3K9me3 mark was dramatically decreased with all peak callers, which implied that the sequencing depth and the number of peaks for this modification might not reach the saturation level.

**Discussion**

The identification of exact protein-binding sites on chromatin is the most important step for ChIP-Seq analysis. Many ChIP-Seq peak calling programs and algorithms have been published. Some of them compared individual performance for transcription factor binding profiles. In this study, to obtain relevant information for the practical usage of peak callers, we analyzed the enrichment of 12 histone marks at specific genomic regions with respect to different sequencing depths, consistency between replicates, specific-
Fig. 4. Peak coverage with different sequencing depths. The genomic coverage of the regions was shown by sampling with different sequence reads (0.5, 0.75, 1.0, 2.5, 5.0, 7.5, 10, 15, 20, and 30 million).
ity, and correlation. Generally, narrow source histone modifications like H3K4me2, H3K4me3, H3K9ac, and H3K27ac showed relatively consistent peaks across the peak callers. However, the peaks identified from H3K4ac, H3K56ac, H3K79me1, and H3K79me2 ChIP-Seq data varied depending on the peak caller, which means that the proper choice of a peak caller is critical.

For the evaluation of reproducibility of peak detection, the Jaccard similarity coefficient and the IDR were considered and both gave fairly good results with point source marks. The broad source marks had lower Jaccard correlation coefficients and low reproducibility.

The sequencing depth, or the count of sequencing reads, is an important factor for identifying the region occupied by a specific protein factor in the genome. A recent study suggested that a sufficient sequencing depth for human ChIP-Seq is 40–50 million reads [9]. However, most published ChIP-Seq data did not reach this read count, probably due to the sequencing cost. Considering sequencing depth, we also analyzed the effect of peak calling performance with different numbers of sequencing reads. The size distribution of enriched regions occupied by peaks was saturated under 1 million reads for H3K4me3 and 2.5 million reads for H3K4me2, but most broad source marks like H3K4me1, H3K9me3, H3K27me3, H3K36me3, and H3K79me2 did not show a distinct saturation profile due to the low sequencing depth or a histone modification type-specific feature.

The validity of ChIP-Seq data can be assessed by the specificity

**Fig. 5.** Specificity of peak calling against the noise signal. The specificity of each program was calculated by sampling with different noise levels. Fifty percent (A), 100% (B), and 150% (C) of control reads added.
of peaks distinguishing true peaks from noise. Randomly selected noise reads from ChIP-Seq input data were used to test the specificity. The individual peak callers showed a good performance with 50% of noise reads, but with a high amount of noise signal, only two callers (CisGenome and PeakSeq) could recapture the original peak regions consistently.

The purpose of this comparative study was to provide practical suggestions for the selection of ChIP-Seq peak calling programs, and thus the comparison of the algorithms and/or statistics used in each program was beyond our research scope. Our results indicated that a proper selection of the peak caller considering the histone modification type is a critical step for the identification of protein-enriched regions specifically. In particular, the peaks occupied by broad and mixed histone marks were dramatically affected by the performance of the peak caller. Based on this study, we constructed an optimal analysis pipeline for ChIP-Seq data and have provided a free ChIP-Seq analysis tool at the Korean Bioinformatics Center (KOBIC) (https://closha.kobic.re.kr/).

ORCID

Hyunji Lee: https://orcid.org/0000-0002-1673-1452
Byunghee Kang: https://orcid.org/0000-0002-6398-5762
Insoon Jang: https://orcid.org/0000-0001-9542-6984
Tae-Young Roh: https://orcid.org/0000-0001-5833-0844

Authors’ Contribution

Conceptualization: HJ, TYR. Formal analysis: HJ, HL, BK, IJ. Writing - original draft: HJ, HL, BK, TYR.

Conflicts of Interest

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

Acknowledgments

This work was supported by grants from the National Research Foundation of Korea (NRF-2014M3C9A3064548 and NRF-2017M3C9A6047625) and BK21 Plus program funded by the Ministry of Education, Republic of Korea (10Z20130012243).

Supplementary Materials

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

References

15. Rye MB, Saetrom P, Drablos F. A manually curated ChIP-seq benchmark demonstrates room for improvement in current

https://doi.org/10.5808/GI.2020.18.4.e42
Druggability for COVID-19: *in silico* discovery of potential drug compounds against nucleocapsid (N) protein of SARS-CoV-2

Manisha Ray¹, Saurav Sarkar¹, Surya Narayan Rath²*

¹All India Institute of Medical Sciences, Bhubaneswar, Odisha 751019, India
²Department of Bioinformatics, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha 751003, India

The coronavirus disease 2019 is a contagious disease and had caused havoc throughout the world by creating widespread mortality and morbidity. The unavailability of vaccines and proper antiviral drugs encourages the researchers to identify potential antiviral drugs to be used against the virus. The presence of RNA binding domain in the nucleocapsid (N) protein of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) could be a potential drug target, which serves multiple critical functions during the viral life cycle, especially the viral replication. Since vaccine development might take some time, the identification of a drug compound targeting viral replication might offer a solution for treatment. The study analyzed the phylogenetic relationship of N protein sequence divergence with other 49 coronavirus species and also identified the conserved regions according to protein families through conserved domain search. Good structural binding affinities of a few natural and/or synthetic phytocompounds or drugs against N protein were determined using the molecular docking approaches. The analyzed compounds presented the higher numbers of hydrogen bonds of selected chemicals supporting the drug-ability of these compounds. Among them, the established antiviral drug glycyrrhizic acid and the phytochemical theaflavin can be considered as possible drug compounds against target N protein of SARS-CoV-2 as they showed lower binding affinities. The findings of this study might lead to the development of a drug for the SARS-CoV-2 mediated disease and offer solution to treatment of SARS-CoV-2 infection.

**Keywords:** COVID-19, molecular docking, molecular modeling, nucleocapsid protein, SARS-CoV-2

---

**Introduction**

The outbreak of novel coronavirus infection has drastically affected the lives of the human population worldwide. This infection started as respiratory illness/pneumonia of unknown origin in Wuhan city of China at the end of the year 2019. The organism identified and termed as novel on 7 January 2020. The World Health Organization (WHO) declared it as a public health emergency of international concern as the disease spread to other regions of the world [1]. The official name of this infection was made as coronavirus disease 2019 (COVID-19) on 11 February 2020. The epidemic was declared a pandemic officially by WHO on 11 March 2020. The novel coronavirus is also termed, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. SARS-CoV-2 infection mainly causes pneumonia, upper and lower respiratory tract infection with fever and
cough as significant clinical symptoms. But some other symptoms include shortness of breath, muscle pain, confusion, headache, sore throat, and acute respiratory distress syndrome, leading to respiratory or multi-organ failure including renal and neurological diseases [2,3].

Coronaviruses (CoVs) are a group of large enveloped viruses with positive sense, single-stranded RNA genomes. Previously identified CoVs in human disease are the alpha CoVs (hCoV-NL63, hCoV-229E) and the beta CoVs (hCoV-OC43), severe acute respiratory syndrome CoV (SARS-CoV), and the Middle East respiratory syndrome CoV (MERS-CoV) [4]. However, among these emerging, highly pathogenic human CoVs, SARS-CoV, MERS-CoV and the newly emerged SARS-CoV-2 infection can result in life-threatening disease conditions and the potential to cause pandemic [2].

The outcome of SARS-CoV-2 sequencing (NCBI reference sequence: NC_045512.2) has proposed about the significant sequence level identity of SARS-CoV-2 with SARS-CoV (79%) rather than MERS-CoV (50%). Besides, the higher levels of transmissibility and pandemic risk of COVID-19 at an early stage has been reported in many studies [1]. In the available literatures, the size of the SARS-CoV-2 (NCBI reference sequence: NC_045512.2) genome is 30KB. The genomic virion consists of four major protein regions including matrix (M) protein, an envelope (E) protein, spike (S) protein, and a nucleocapsid (N) protein within the viral envelope [5,6]. The functional architectures of each of these viral proteins have accurately characterized. S protein primarily binds to the host cell receptor and form attachment with the host body. Alternatively, M and E proteins are involved in the formation of the viral envelope [6]. Similarly, SARS-CoV-2 protein N is a multifunctional RNA binding protein, necessary for viral RNA transcription, replication and/or assembly of virus [6]. Interestingly, a unique N-terminal RNA binding domain of SARS-CoV-2 N protein has identified as a novel antiviral drug target site [7]. The viral N protein packages the genome into long, flexible, and helical RNP complexes, called nucleocapsids which protect the SARS-CoV-2 virion structure [5]. Additionally, N protein has a significant contribution towards timely replication and reliable transmission of SARS-CoV-2 during its life cycle. Therefore, N protein (PDB ID: 6VYO) can be considered as a novel drug target of SARS-CoV-2.

The SARS-CoV-2 infection has created a dangerous pandemic situation due to its quick transmission and deadly nature. It has affected both the health and economy of human population across the globe tremendously. Many ongoing pieces of research are trying to develop vaccines to control this situation, but all are in various phases of trials. Thus, the present study has focused on in silico discovery of potent leads from several antiviral drugs and compounds of plant origin against SARS-CoV-2 infection. The present study would throw lights on the discovery of antiviral drug against SARS-CoV-2.

**Methods**

**Sequence retrieval and construction of phylogenetic tree**
Nucleocapsid protein sequences of total 49 CoV species and/or strains including SARS-CoV-2 were retrieved in FASTA format from NCBI web server (https://www.ncbi.nlm.nih.gov/) on 30 March 2020. Two N proteins of Ebola and H1N1 virus were included to study evolutionary divergence across species. Further, total 51 N protein sequences were aligned using MUSCLE algorithm of Molecular Evolutionary Genetics Analysis 7 (MEGA 7) package [8]. The resulted alignment was used to generate phylogenetic tree using neighbour joining (NJ) method of MEGA 7 for 1,000 bootstrap replicates.

**Conserved domain search**
Functional domains of SARS-CoV-2 N protein (YP_009724397.2) were identified using NCBI conserved domain database (CDD) (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) search. The CDD is a collection of domain models which imports information from Pfam, SMART, COG, and NCBI to provide a more accurate assessment of neighbor relationships between protein sequences [9].

**Prediction of structural element**
The secondary structure of SARS-CoV-2 N protein was predicted from its complete amino acid sequence (accession No. YP_0097 24397.2) using PSIPRED 4.0 algorithm [10]. Similarly, protein disorder portion and membrane helix region was predicted by using DISOPRED3 [10] and MEMSAT-SVM algorithm [10] of PSIPRED web server (http://bioinf.cs.ucl.ac.uk/psipred/).

**Retrieval and preparation of 3D structure**
Available N-terminal domain structure (PDB ID: 6VYO) of SARS-CoV-2 N protein was retrieved from Protein Data Bank (PDB) (https://www.rcsb.org/). Initially, hydrogen atoms were added to protein structure after removal of all water and other hetero molecules. Further, energy minimization was performed using Discovery Studio 3.5 suite to obtain a properly optimized structure of target protein.

**Drug-binding cavity prediction**
In absence of knowledge on exact drug-binding site, probable
binding cavity within SARS-CoV-2 N protein was predicted using metaPocket 2.0 (https://projects.biotec.tu-dresden.de/metapocket/). MetaPocket tool identifies cavities on protein surface for drug-binding site prediction using multiple computational approaches [11] such as PASS11, LIGSITE, Fpocket, SURFNET, GHECOM, and ConCavity.

Selection of ligand molecules
Different natural compounds of plant origin reported with antiviral, anti-inflammation, anti--influenza, anti--human immunodeficiency virus, anti-hepatic properties were shortlisted from different literatures. In addition, few Food and Drug Administration approved, and investigational antiviral drugs were also selected from Drug Bank (https://www.drugbank.ca/) database for further investigation.

Ligand structure retrieval and correction
Three-dimensional structures of natural ligands were retrieved from PubChem (https://pubchem.ncbi.nlm.nih.gov/) database in SDF format and converted into PDB format using Discovery Studio 3.5 suite. Similarly, PDB structures of antiviral drugs were collected from the Drug Bank (https://www.drugbank.ca/). Further, structure optimization and protonation state of all ligands were achieved using Discovery Studio 3.5 suite.

Molecular docking
Molecular docking was performed between all selected ligands (phytochemicals and antiviral drugs) and the drug target (N protein, PDB ID: 6VYO) separately in order to identify the most efficient inhibitor against SARS-CoV-2. AutoDock 4.2 (http://autodock.scripps.edu/) and AutoDock Tools 4 tool [12] were used to perform molecular docking study. The N-terminal RNA binding domain of SARS-CoV-2 N protein was observed as a homotramer structure; therefore, only chain A of the available crystal structure was employed for docking analysis. Prior to docking, Kollman charges and polar hydrogen atoms were added to the target structure. Both ligand and receptor structures were prepared using ADT tool and converted to pdbqt format before docking. A virtual grid box was set around the drug-binding cavity of the target structure with size of 74, 78, and 74 Å in x, y, and z direction in spacing of 0.375 Å. Semi flexible docking was performed by maintaining target structure as rigid and allowing flexibility to ligand molecules within the drug-binding pocket [13]. Lamarckian genetic algorithm was used with 25,000,000 energy evaluation steps for each dock run. Auto dock generated 10 conformers based on free binding energy for each protein-ligand complex. The most energetically favorable (lowest energy) binding complex was considered for analysis. Further analysis and presentation of atomic interaction between docked complexes were performed using PyMol molecular graphics tool (http://www.pymol.org).

Results

Molecular phylogeny ascertained sequential divergence of SARS-CoV-2 N protein
Total 49 N proteins different CoV species, including SARS-CoV-2 (Table 1) were retrieved to construct the phylogenetic tree.

Again, protein sequences of two distance homologues of SARS-CoV-2 such as Ebola (accession No. SCD11531.1) and H1N1 (accession No. YP_009118629.1) virus were included within the tree in order to establish sequential divergence pattern across species. The phylogenetic tree was constructed using NJ method [14] with tree evaluation step for 1,000 bootstrap replicates. The resulted rooted tree (Fig. 1) clustered into two major clades. Total 49 species were diversified within both of the clades (clade-I, 26; clade-II, 23). The target N protein sequence of SARS-CoV-2 (accession No. YP_009724397.2) was grouped with SARS-CoV (severe acute respiratory syndrome-related virus) (accession No. NP_828858.1) sequence within clade-I with branch frequency of 100% which pointed out regarding their significant evolutionary closeness. One separate clade was formed within the tree with branch frequency of 61% among the two outgroups (Ebola and H1N1) which clearly revealed their divergence from all other 49 sequences.

Functional domain identified for SARS-CoV-2 N protein
The complete sequence of SARS-CoV-2 N protein (accession No. YP_009724397.2) comprises of 419 amino acids. All functional domain regions within the N protein sequence of SARS-CoV-2 were identified from its conserved pattern among the members of beta CoV nucleocapsid protein family. The conserved domains were observed within the aligned region of SARS-CoV-2 N protein from 14-368 amino acids (Fig. 2A) with the members of the superfamily (pfam00937) (Fig. 2B). The CD search identified one. N-terminal (50-175 amino acids) and one C-terminal (258-359 amino acids) functional domain (Fig. 2C) with good bit score (424.07) and lowest e-value (7.05e-148). The nucleocapsid N-terminal domain (NTD) of SARS-CoV-2 was shown significant similarities with the conserved domain of family cd21554 whereas the C-terminal domain (CTD) found conserved within the family members of cd21559 (Fig. 2D).

Structural elements of SARS-CoV-2 N protein
In the absence of full-length structure, the secondary structural el-
<table>
<thead>
<tr>
<th>No.</th>
<th>Species name</th>
<th>NCBI accession</th>
<th>Length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Duck coronavirus (avian CoV)</td>
<td>AKF17732.1</td>
<td>414</td>
</tr>
<tr>
<td>2</td>
<td>Turkey coronavirus (avian CoV)</td>
<td>YP_001941174.1</td>
<td>409</td>
</tr>
<tr>
<td>3</td>
<td>Infectious bronchitis virus (avian CoV)</td>
<td>NP_040838.1</td>
<td>409</td>
</tr>
<tr>
<td>4</td>
<td>Infectious bronchitis virus (avian CoV)</td>
<td>AKV63212.1</td>
<td>409</td>
</tr>
<tr>
<td>5</td>
<td>Rat CoV parker (murine CoV)</td>
<td>YP_003029852.1</td>
<td>454</td>
</tr>
<tr>
<td>6</td>
<td>Murine hepatitis virus (murine CoV)</td>
<td>AAU06361.1</td>
<td>454</td>
</tr>
<tr>
<td>7</td>
<td>Murine hepatitis virus (murine CoV)</td>
<td>NP_045302.1</td>
<td>454</td>
</tr>
<tr>
<td>8</td>
<td>Bovine coronavirus (beta CoV)</td>
<td>NP_150083.1</td>
<td>448</td>
</tr>
<tr>
<td>9</td>
<td>Human coronavirus OC43 (beta CoV)</td>
<td>YP_00955245.1</td>
<td>448</td>
</tr>
<tr>
<td>10</td>
<td>Middle East respiratory syndrome-related coronavirus (MERS-CoV)</td>
<td>YP_007188585.1</td>
<td>411</td>
</tr>
<tr>
<td>11</td>
<td>Mink coronavirus 1</td>
<td>YP_009019186.1</td>
<td>376</td>
</tr>
<tr>
<td>12</td>
<td>Feline infectious peritonitis virus (alpha coronavirus 1)</td>
<td>YP_004070199.1</td>
<td>377</td>
</tr>
<tr>
<td>13</td>
<td>Transmissible gastroenteritis virus (alpha coronavirus 1)</td>
<td>NP_058428.1</td>
<td>382</td>
</tr>
<tr>
<td>14</td>
<td>Rousettus bat coronavirus HKU9</td>
<td>YP_001039975.1</td>
<td>468</td>
</tr>
<tr>
<td>15</td>
<td>Pipistrellus bat coronavirus HKU5</td>
<td>YP_001039969.1</td>
<td>427</td>
</tr>
<tr>
<td>16</td>
<td>Canada goose coronavirus</td>
<td>YP_009755908.1</td>
<td>414</td>
</tr>
<tr>
<td>17</td>
<td>Tylonycteris bat coronavirus HKU4</td>
<td>YP_001039960.1</td>
<td>423</td>
</tr>
<tr>
<td>18</td>
<td>Severe acute respiratory syndrome-related coronavirus (SARS-CoV)</td>
<td>NP_828858.1</td>
<td>422</td>
</tr>
<tr>
<td>19</td>
<td>Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2)</td>
<td>YP_009724397.2</td>
<td>419</td>
</tr>
<tr>
<td>20</td>
<td>Alpha coronavirus bat-CoV/Pkuhii/Italy/3398-19/2015</td>
<td>YP_009755894.1</td>
<td>432</td>
</tr>
<tr>
<td>21</td>
<td>Miniopterus bat coronavirus 1</td>
<td>YP_001718609.1</td>
<td>389</td>
</tr>
<tr>
<td>22</td>
<td>Wencheng Sm shrew coronavirus</td>
<td>YP_009389428.1</td>
<td>366</td>
</tr>
<tr>
<td>23</td>
<td>Coronavirus AcCoV-JC34</td>
<td>YP_009380526.1</td>
<td>389</td>
</tr>
<tr>
<td>24</td>
<td>Lucheng Rn rat coronavirus</td>
<td>YP_009336487.1</td>
<td>391</td>
</tr>
<tr>
<td>25</td>
<td>NL63-related bat coronavirus</td>
<td>APD51488.1</td>
<td>433</td>
</tr>
<tr>
<td>26</td>
<td>NL63-related bat coronavirus</td>
<td>YP_009328939.1</td>
<td>407</td>
</tr>
<tr>
<td>27</td>
<td>Rousettus bat coronavirus</td>
<td>YP_009273009.1</td>
<td>443</td>
</tr>
<tr>
<td>28</td>
<td>Ferret coronavirus</td>
<td>BAV31353.1</td>
<td>374</td>
</tr>
<tr>
<td>29</td>
<td>BtMr-AlphaCoV/SAX2011</td>
<td>YP_009198836.1</td>
<td>429</td>
</tr>
<tr>
<td>30</td>
<td>BtNv-AlphaCoV/SC2013</td>
<td>YP_009201734.1</td>
<td>431</td>
</tr>
<tr>
<td>31</td>
<td>BtRf-AlphaCoV/HuB 2013</td>
<td>YP_009198974.1</td>
<td>383</td>
</tr>
<tr>
<td>32</td>
<td>BtRf-AlphaCoV/YN2012</td>
<td>YP_009200739.1</td>
<td>375</td>
</tr>
<tr>
<td>33</td>
<td>Swine enteric coronavirus</td>
<td>YP_009199247.1</td>
<td>382</td>
</tr>
<tr>
<td>34</td>
<td>Camel alpha coronavirus</td>
<td>YP_009194643.1</td>
<td>382</td>
</tr>
<tr>
<td>35</td>
<td>Beta coronavirus HKU24</td>
<td>YP_009113031.1</td>
<td>443</td>
</tr>
<tr>
<td>36</td>
<td>Bat-Hp-Betacoronavirus/Zhejiang 2013</td>
<td>YP_009072446.1</td>
<td>418</td>
</tr>
<tr>
<td>37</td>
<td>Betacoronavirus Erinaceus/VMC/DEU/2012</td>
<td>YP_009513018.1</td>
<td>424</td>
</tr>
<tr>
<td>38</td>
<td>Bat coronavirus CDPHE15/USA/2006</td>
<td>YP_008439206.1</td>
<td>425</td>
</tr>
<tr>
<td>39</td>
<td>Rousettus bat coronavirus V HKU10</td>
<td>YP_006908646.1</td>
<td>402</td>
</tr>
<tr>
<td>40</td>
<td>Rabbit coronavirus HKU14</td>
<td>YP_005454249.1</td>
<td>444</td>
</tr>
<tr>
<td>41</td>
<td>Beluga whale coronavirus SW1</td>
<td>YP_001876448.1</td>
<td>379</td>
</tr>
<tr>
<td>42</td>
<td>Miniopterus bat coronavirus HKU8</td>
<td>YP_001718616.1</td>
<td>422</td>
</tr>
<tr>
<td>43</td>
<td>Rhinolophus bat coronavirus HKU2</td>
<td>YP_001552240.1</td>
<td>375</td>
</tr>
<tr>
<td>44</td>
<td>Scotophilus bat coronavirus 512</td>
<td>YP_001351688.1</td>
<td>394</td>
</tr>
<tr>
<td>45</td>
<td>Human coronavirus HKU1</td>
<td>YP_173242.1</td>
<td>441</td>
</tr>
<tr>
<td>46</td>
<td>Human CoV NL63</td>
<td>YP_003771.1</td>
<td>377</td>
</tr>
<tr>
<td>47</td>
<td>Bat coronavirus BM48-31/BGR/2008</td>
<td>YP_003858591.1</td>
<td>417</td>
</tr>
<tr>
<td>48</td>
<td>Human coronavirus 229E</td>
<td>NP_073556.1</td>
<td>389</td>
</tr>
<tr>
<td>49</td>
<td>Porcine epidemic diarrhea virus</td>
<td>NP_598314.1</td>
<td>441</td>
</tr>
</tbody>
</table>
lements of SARS-CoV-2 N protein were predicted from its primary sequence using PSIPRED web server. Secondary structural elements such as two long, eight medium, two short helical regions and two medium, nine short β-sheets were predicted within the complete sequence of SARS-CoV-2 N protein (Fig. 3).

Most of the NTD (50–175) regions were predicted as β-sheets and coils. On the contrary, structural elements such as helices, β-sheets, and coils were observed within CTD (258–359) regions (Fig. 3). Further, highly disordered regions of SARS-CoV-2 N protein were observed above the cut off score (0.5) from amino acid positions 1–50, 180–250, and 350–419 (Fig. 4A). However, significant disorder portions were absent within the NTD (50–175) and CTD (258–359) regions (Fig. 4A). According to MEMSAT-SVM algorithm, the sub-cellular localization of SARS-CoV-2 nucleocapsid NTD was found as cytoplasmic, whereas a small C-terminal transmembrane region was noticed from 302–317 amino acids (Fig. 4B).

Structure preparation and active site identification of N protein NTD

Homology search using BLASTP algorithm revealed the structure of N-terminal RNA binding domain occupied 30% region of SARS-CoV-2 N protein (accession No. YP_009724397.2) sequence with 100% identity. Therefore, the three-dimensional structure of SARS-CoV-2 N protein was retrieved and processed for structural correction and optimization. The possible drug-binding cavity of SARS-CoV-2 N protein was predicted in the absence of literary evidence. Algorithm of metaPocket was generated top three hits after clustering the results of PASS11, LIGSITE, Fpocket, SURFNET, GHECOM, and ConCavity. Out
Fig. 2. Conserved functional domains of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) nucleocapsid protein. (A) Sequence alignment between SARS-CoV-2 and members of super family (pfam00937). (B) The alignment between SARS-CoV-2 and consensus sequence of pfam00937 nucleocapsid protein. The conserved amino acid patterns were highlighted using boxes. (C) All functional domain regions of SARS-CoV-2 nucleocapsid protein were presented in schematic diagram. N-NTD, nucleocapsid protein N-terminal domain; N-CTD, nucleocapsid protein C-terminal domain. (D) The sequence alignment of N-NTD (50-175) and N-CTD of SARS-CoV-2 with their respective conserved domain family.

Fig. 3. Predicted secondary structural elements for full length nucleocapsid protein of severe acute respiratory syndrome coronavirus 2. Helix, pink cylinder; Sheet, yellow cylinder.
of these three, the large active pocket was considered a possible drug-binding cavity (Fig. 5).

Structure preparation natural/synthetic ligands against SARS-CoV-2 N protein
As of literature, a total of eight natural compounds of plant origin and three synthetic compounds (Table 2) were identified with antiviral properties, therefore, prepared to dock against SARS-CoV-2 N protein.

Again, seven antiviral drugs (Table 3) were also included within the study to discover potent inhibitor against N protein of SARS-CoV-2. Finally, 3D structures of a total of eighteen ligands were extracted from online databases (PubChem/Drug Bank) and prepared for docking study.

Molecular docking identified efficient ligand against SARS-CoV-2 N protein
Molecular docking is an efficient technique to identify the binding affinity of a drug compound against a drug target [15,25]. Therefore, all possible inhibitors were docked separately against SARS-CoV-2 N protein to discover effective ligand and important atomic interaction between protein-ligand complexes within the drug-binding cavity. The resulted in free binding energy, and the inhibition constant of each binding complex was reported in Table 4. According to docking energy score and inhibition constant (KI), total eight antiviral compounds such as glycyrrhizic acid (−12.61 kcal/mol; KI, 573.72 pm), theaflavin (−10.35 kcal/mol; KI, 26.03 nM), diosgenin (−10.06 kcal/mol; KI, 42.53 nM), U18666A (−9.08 kcal/mol; KI, 219.38 nM), ethyl brevifolinicar-
### Table 2. Eleven ligand molecules (natural and synthetic) and antiviral properties

<table>
<thead>
<tr>
<th>No.</th>
<th>Natural/Synthetic compounds</th>
<th>Pub Chem CID</th>
<th>Source/Plant name</th>
<th>Property</th>
<th>Virus</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Theaflavin</td>
<td>135403798</td>
<td><em>Camellia sinensis</em> (tea plant)</td>
<td>Prevents influenza by inhibiting replication using potentially directs virucidal effect</td>
<td>H1N1</td>
<td>[15]</td>
</tr>
<tr>
<td>2</td>
<td>Curcumin</td>
<td>969516</td>
<td><em>Curcuma longa</em> L. (turmeric)</td>
<td>Antiviral activity against FIPV</td>
<td>FIPV, HIV, influenza</td>
<td>[16]</td>
</tr>
<tr>
<td>3</td>
<td>Diosgenin</td>
<td>99474</td>
<td>Synthetic</td>
<td>Effectively blocks the replication of hepatitis C virus</td>
<td>Hepatitis C virus</td>
<td>[17]</td>
</tr>
<tr>
<td>4</td>
<td>Ladanein</td>
<td>3084066</td>
<td><em>Marrubium peregrinum</em> L.</td>
<td>Effectively inhibits the post attachment entry step of hepatitis C virus</td>
<td>Hepatitis C virus</td>
<td>[17]</td>
</tr>
<tr>
<td>5</td>
<td>Quercetin</td>
<td>5280343</td>
<td>Synthetic</td>
<td>Inhibits virus replication and viral nucleocapsid formation by inhibiting DNA polymerase of hepatitis B</td>
<td>Hepatitis B virus</td>
<td>[17,18]</td>
</tr>
<tr>
<td>6</td>
<td>Ethyl brevifolincarbocyste</td>
<td>5487248</td>
<td>Synthetic</td>
<td>Inhibits virus replication and viral nucleocapsid formation by inhibiting DNA polymerase of hepatitis B</td>
<td>Hepatitis B virus</td>
<td>[17]</td>
</tr>
<tr>
<td>7</td>
<td>Quercitrin</td>
<td>5280459</td>
<td><em>Phyllanthus niruri</em></td>
<td>Inhibit virus replication and viral nucleocapsid formation by inhibiting DNA polymerase of hepatitis B</td>
<td>Hepatitis B virus</td>
<td>[18]</td>
</tr>
<tr>
<td>8</td>
<td>U18666A</td>
<td>9954082</td>
<td>Synthetic</td>
<td>Inhibits the proliferation of type 1 FIPV</td>
<td>Feline infectious peritonitis virus</td>
<td>[19]</td>
</tr>
<tr>
<td>9</td>
<td>Apigenin</td>
<td>5280443</td>
<td><em>Ocimum sanctum</em> (Tulsi)</td>
<td>Prevents the early multiplication of H1N1 virus and control the viral growth</td>
<td>H1N1</td>
<td>[20]</td>
</tr>
<tr>
<td>10</td>
<td>Resveratrol</td>
<td>445154</td>
<td><em>Vitis labrusca</em></td>
<td>Effectively reduce the inflammatory cell production and pro-inflammatory cytokine accumulation</td>
<td>Inflammatory virus</td>
<td>[21]</td>
</tr>
<tr>
<td>11</td>
<td>Allicin</td>
<td>65036</td>
<td><em>Allium sativum</em> (garlic)</td>
<td>Inhibit virus penetration and proliferation (inhibit cell proliferation, protect the heart injury, liver damage, anti-inflammation)</td>
<td>Influenza</td>
<td>[22]</td>
</tr>
</tbody>
</table>

CID, compound ID; FIPV, feline infectious peritonitis virus; HIV, human immunodeficiency virus.

### Table 3. Seven antiviral drugs and medicinal value

<table>
<thead>
<tr>
<th>No.</th>
<th>Synthetic/Natural drug compound</th>
<th>DB ID</th>
<th>Status</th>
<th>Source/Plant name</th>
<th>Treatment/Property</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Glycyrrhizic acid (glycyrrhizin)</td>
<td>DB13751</td>
<td>Approved, experimental</td>
<td>Glycyrrhiza glabra</td>
<td>Inhibit viral replication of SARS-CoV</td>
<td>[23]</td>
</tr>
<tr>
<td>2</td>
<td>Ribavirin</td>
<td>DB00811</td>
<td>Approved</td>
<td>Synthetic</td>
<td>Effective against chronic hepatitis C virus, SARS-CoV</td>
<td>PMID:18565019, [17]</td>
</tr>
<tr>
<td>3</td>
<td>Tenofovir</td>
<td>DB14126</td>
<td>Experimental, investigational</td>
<td><em>Phyllanthus niruri</em></td>
<td>Hepatitis B virus</td>
<td>[18]</td>
</tr>
<tr>
<td>4</td>
<td>Berberine</td>
<td>DB04115</td>
<td>Approved, investigational</td>
<td>Berberis vulgaris</td>
<td>Prevents the HIV-PI induced inflammation</td>
<td>[22]</td>
</tr>
<tr>
<td>5</td>
<td>Emodin</td>
<td>DB07715</td>
<td>Investigational</td>
<td>Radix et Rhizoma Paei, Radix Polygoni Multiflori</td>
<td>Blocks the S protein of SARS-CoV and ACE2 interaction</td>
<td>[24]</td>
</tr>
<tr>
<td>6</td>
<td>Chloroquine</td>
<td>DB00608</td>
<td>Approved, investigational, vet approved</td>
<td>Synthetic</td>
<td>HIV, influenza A/H5N1, SARS-CoV, human coronavirus 229E</td>
<td>PMID: 23648708</td>
</tr>
<tr>
<td>7</td>
<td>Hydroxy chloroquine</td>
<td>DB01611</td>
<td>Approved</td>
<td>Synthetic</td>
<td>HIV, DENV</td>
<td>PMID:25321315</td>
</tr>
</tbody>
</table>
boxylate (−9.07 kcal/mol; KI, 226.42 nM), quercitrin (−9.04 kcal/mol; KI, 238.18 nM), curcumin (−8.68 kcal/mol; KI, 434.59 nM), and ladanein (−8.19 kcal/mol; KI, 988.63 nM) showed good binding efficiency than rest of the compounds (Table 4). Presence of an ample number of polar interactions has a significant contribution towards the stability of a specific ligand within the binding site of drug target. Therefore, h-bond interaction between the drug target and ligands were inspected. Interestingly, good binding affinity and strong h-bond interaction within distance ≤ 3.5 Å from binding cavity were identified in case of 10 suitable compounds such as glycyrrhizic acid (−12.61 kcal/mol; h-bond, 16 nos), theaflavin (−10.35 kcal/mol; h-bond, 11 nos), ethyl brevifolin carboxylate (−9.07 kcal/mol; h-bond, 6 nos), quercitrin (−9.04 kcal/mol; h-bond, 11 nos), curcumin (−8.68 kcal/mol; h-bond, 5 nos), ladanein (−8.19 kcal/mol; h-bond, 8 nos), apigenin (−7.98 kcal/mol; h-bond, 6 nos), tenofovir (−6.92 kcal/mol; h-bond, 9 nos), resveratrol (−6.91 kcal/mol; h-bond, 5 nos), ribavirin (−6.41 kcal/mol; h-bond, 12 nos), indicated about their efficacy to block the important site within the RNA binding domain of SARS-CoV-2 N protein (Tables 4 and 5, Fig. 6).

To its support, few amino acid residues such as PHE 66, PRO 67, ARG 68, GLY 69, GLN 70, TYR 123, TRP 132, and ALA 134 were found commonly interacting with all of these ligands within the binding cavity of SARS-CoV-2 N protein. However, presence of h-bond interaction with quite good binding energy and inhibition constant values were also noticed in case of rest seven antiviral compounds such as diosgenin (−10.06 kcal/mol; KI, 42.53 nM; h-bond, 3 nos), U18666A (−9.08 kcal/mol; KI, 219.38 nM; h-bond, 2 nos), berberine (−7.87 kcal/mol; KI, 1.69 µM; h-bond, 2 nos), emodin (−7.82 kcal/mol; KI, 1.86 µM; h-bond, 6 nos), quercetin (−7.47 kcal/mol; KI, 3.33 µM; h-bond, 8 nos), hydroxychloroquine (−7.35 kcal/mol; KI, 4.07 µM; h-bond, 2 nos), chloroquine (−6.86 kcal/mol; KI, 9.34 µM; h-bond, 1 nos) inbound form with SARS-CoV-2 N protein (Table 4, Fig. 7). Overall docking study confirmed the binding potential of the discussed phytochemicals and drugs, against drug target, Nucleocapsid protein of SARS-CoV-2.

**Discussion**

The SARS-CoV-2 or COVID-19 pandemic has created an alarming situation due to severe infection and death rate worldwide. Researchers all over the world are in search to identify novel drug/vaccine target as well as the development of drug/vaccine to combat the disease. Several recent studies have been reported probable synthetic drug candidates such as conivaptan, amyrin, ZINC00027115482 [26], ritonavir, lopinavir, umifenovir [27], theophylline, pyrimidine [28], simprevir and grazoprevir [29] against nucleocapsid protein of SARS-CoV-2. As, N protein has a vital role for the survival and growth of SARS-CoV-2 thus authors focused on the discovery of potential natural or synthetic compounds to block its regular mechanism. In support of the present

---

**Table 4.** Docking scores of 18 ligands against SARS-CoV-2 N protein

<table>
<thead>
<tr>
<th>No.</th>
<th>Ligands (phytochemicals/drugs)</th>
<th>Docking energy scores (kcal/mol)</th>
<th>Intermolecular energy (kcal/mol)</th>
<th>Inhibition constant (KI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Glycyrrhizic acid (glycyrrhin)</td>
<td>−12.61</td>
<td>−14.7</td>
<td>573.72 pM</td>
</tr>
<tr>
<td>2</td>
<td>Theaflavin</td>
<td>−10.35</td>
<td>−13.63</td>
<td>26.03 nM</td>
</tr>
<tr>
<td>3</td>
<td>Diosgenin</td>
<td>−10.06</td>
<td>−10.35</td>
<td>42.53 nM</td>
</tr>
<tr>
<td>4</td>
<td>U18666A</td>
<td>−9.08</td>
<td>−10.87</td>
<td>219.38 nM</td>
</tr>
<tr>
<td>5</td>
<td>Ethyl brevifolin carboxylate</td>
<td>−9.07</td>
<td>−10.86</td>
<td>226.42 nM</td>
</tr>
<tr>
<td>6</td>
<td>Quercitrin</td>
<td>−9.04</td>
<td>−12.02</td>
<td>238.18 nM</td>
</tr>
<tr>
<td>7</td>
<td>Curcumin</td>
<td>−8.68</td>
<td>−11.66</td>
<td>434.59 nM</td>
</tr>
<tr>
<td>8</td>
<td>Ladanein</td>
<td>−8.19</td>
<td>−9.68</td>
<td>988.63 nM</td>
</tr>
<tr>
<td>9</td>
<td>Apigenin</td>
<td>−7.98</td>
<td>−9.17</td>
<td>1.43 µM</td>
</tr>
<tr>
<td>10</td>
<td>Berberine</td>
<td>−7.87</td>
<td>−8.47</td>
<td>1.69 µM</td>
</tr>
<tr>
<td>11</td>
<td>Emodin</td>
<td>−7.82</td>
<td>−8.71</td>
<td>1.86 µM</td>
</tr>
<tr>
<td>12</td>
<td>Quercetin</td>
<td>−7.47</td>
<td>−9.26</td>
<td>3.33 µM</td>
</tr>
<tr>
<td>13</td>
<td>Hydroxy chloroquine</td>
<td>−7.35</td>
<td>−10.04</td>
<td>4.07 µM</td>
</tr>
<tr>
<td>14</td>
<td>Tenofovir</td>
<td>−6.92</td>
<td>−8.41</td>
<td>8.53 µM</td>
</tr>
<tr>
<td>15</td>
<td>Resveratrol</td>
<td>−6.91</td>
<td>−8.4</td>
<td>8.63 µM</td>
</tr>
<tr>
<td>16</td>
<td>Chloroquine</td>
<td>−6.86</td>
<td>−9.25</td>
<td>9.34 µM</td>
</tr>
<tr>
<td>17</td>
<td>Ribavirin</td>
<td>−6.41</td>
<td>−8.2</td>
<td>19.88 µM</td>
</tr>
<tr>
<td>18</td>
<td>Allicin</td>
<td>−4.69</td>
<td>−6.18</td>
<td>363.41 µM</td>
</tr>
</tbody>
</table>

SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

https://doi.org/10.5808/GI.2020.18.4.e43
Table 5. Polar interaction (distance ≤ 3.5 Å) between selected antiviral compounds and nucleocapsid protein of SARS-CoV-2

<table>
<thead>
<tr>
<th>No.</th>
<th>Phytochemical/Drug</th>
<th>H-bond residue</th>
<th>Bond</th>
<th>Length (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Glycyrrhizic acid (glycyrrhizin)</td>
<td>LYS 65</td>
<td>NZ…O</td>
<td>2.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PHE 66</td>
<td>N…O</td>
<td>3.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N…O</td>
<td>2.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OH…O</td>
<td>2.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PRO 67</td>
<td>OH…O</td>
<td>3.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ARG 68</td>
<td>NE…O</td>
<td>3.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLY 69</td>
<td>N…O</td>
<td>2.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLN 70</td>
<td>OH…O</td>
<td>2.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OH…O</td>
<td>3.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TYR 123</td>
<td>OH…O</td>
<td>2.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OH…O</td>
<td>3.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLY 124</td>
<td>OH…O</td>
<td>3.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ALA 134</td>
<td>N…O</td>
<td>2.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OH…O</td>
<td>3.09</td>
</tr>
<tr>
<td>2</td>
<td>Theaflavin</td>
<td>PHE 66</td>
<td>N…OH53</td>
<td>3.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLY 69</td>
<td>N…O</td>
<td>2.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLN 70</td>
<td>O…H60</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TYR 123</td>
<td>OH…O9</td>
<td>2.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OH…O9</td>
<td>3.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ILE 130</td>
<td>O…H64</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TRP 132</td>
<td>O…H53</td>
<td>2.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N…O1</td>
<td>3.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ALA 134</td>
<td>O…OH59</td>
<td>3.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N…OH59</td>
<td>3.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OH…O10</td>
<td>3.23</td>
</tr>
<tr>
<td>3</td>
<td>Ethyl brevifolin carboxylate</td>
<td>PHE 66</td>
<td>N…O5</td>
<td>3.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ARG 68</td>
<td>NE…O6</td>
<td>3.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NH1…O6</td>
<td>3.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLY 69</td>
<td>N…O8</td>
<td>2.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLN 70</td>
<td>O…H31</td>
<td>2.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TRP 132</td>
<td>O…H30</td>
<td>2.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O…O5</td>
<td>2.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N…O5</td>
<td>3.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ALA 134</td>
<td>O…H35</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N…O8</td>
<td>3.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N…O7</td>
<td>2.98</td>
</tr>
<tr>
<td>4</td>
<td>Quercitrin</td>
<td>PHE 66</td>
<td>O…H43</td>
<td>2.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O…H51</td>
<td>2.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N…O10</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PRO 67</td>
<td>O…H41</td>
<td>2.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLY 69</td>
<td>N…O4</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLN 70</td>
<td>O…H42</td>
<td>2.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O…H41</td>
<td>3.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TYR 123</td>
<td>OH…O2</td>
<td>3.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OH…O7</td>
<td>2.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLY 124</td>
<td>O…H50</td>
<td>2.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ALA 134</td>
<td>N…O4</td>
<td>2.68</td>
</tr>
</tbody>
</table>

Table 5. Continued

<table>
<thead>
<tr>
<th>No.</th>
<th>Phytochemical/Drug</th>
<th>H-bond residue</th>
<th>Bond</th>
<th>Length (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Curcumin</td>
<td>PHE 66</td>
<td>N…O5</td>
<td>3.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLY 69</td>
<td>N…O4</td>
<td>2.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ASN 126</td>
<td>O…H40</td>
<td>2.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LYS 127</td>
<td>N…O3</td>
<td>2.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ALA 134</td>
<td>N…O4</td>
<td>2.81</td>
</tr>
<tr>
<td>6</td>
<td>Ladanein</td>
<td>PRO 67</td>
<td>O…O5</td>
<td>3.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ARG 68</td>
<td>NE…O2</td>
<td>2.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLY 69</td>
<td>N…O3</td>
<td>2.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N…O4</td>
<td>3.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLN 70</td>
<td>O…H30</td>
<td>2.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O…O5</td>
<td>2.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ALA 134</td>
<td>O…H31</td>
<td>2.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N…O3</td>
<td>2.75</td>
</tr>
<tr>
<td>7</td>
<td>Apigenin</td>
<td>ARG 68</td>
<td>NE…O3</td>
<td>3.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NH1…O3</td>
<td>3.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLY 69</td>
<td>N…O1</td>
<td>3.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLN 70</td>
<td>O…H29</td>
<td>2.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THR 135</td>
<td>O…H30</td>
<td>3.00</td>
</tr>
<tr>
<td>8</td>
<td>Tenofovir</td>
<td>PRO 67</td>
<td>O…N</td>
<td>3.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ARG 68</td>
<td>NE…O</td>
<td>3.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NE…O</td>
<td>3.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NH1…O</td>
<td>2.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NH1…O</td>
<td>2.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLN 70</td>
<td>O…N</td>
<td>2.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O…N</td>
<td>3.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ALA 134</td>
<td>O…N</td>
<td>3.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O…N</td>
<td>2.77</td>
</tr>
<tr>
<td>9</td>
<td>Resveratrol</td>
<td>GLY 69</td>
<td>N…O2</td>
<td>2.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLN 70</td>
<td>O…O2H28</td>
<td>2.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TYR 123</td>
<td>OH…O1</td>
<td>2.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ALA 134</td>
<td>O…O2H28</td>
<td>3.03</td>
</tr>
<tr>
<td>10</td>
<td>Ribavirin</td>
<td>PRO 67</td>
<td>O…O5</td>
<td>3.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLY 69</td>
<td>N…O5</td>
<td>2.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLN 70</td>
<td>O…O5</td>
<td>3.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TYR 123</td>
<td>OH…N8</td>
<td>2.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TRP 132</td>
<td>O…N9</td>
<td>2.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ALA 134</td>
<td>O…N</td>
<td>3.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O…O5</td>
<td>3.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O…N7</td>
<td>2.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O…O2H24</td>
<td>2.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLU 136</td>
<td>OE2…H25</td>
<td>2.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLY 137</td>
<td>N…O4H27</td>
<td>3.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ALA 138</td>
<td>N…O4H27</td>
<td>2.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O…H27</td>
<td>2.99</td>
</tr>
</tbody>
</table>

SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.
scenario, the current study has tried to conduct some critical analyses on important drug target, i.e., nucleocapsid (N) protein of SARS-CoV-2. The present research also focuses on in silico discovery of potent natural/synthetic compounds against the virus.

The phylogenetic study among different CoV species community identified the close relation and less diversification between N proteins of SARS-CoV and SARS-CoV-2, which indicates the high similarities between those species. The protein family sequence similarity search or the conserved domain search points out the versatility of SARS-CoV-2 N protein, which is predicted by the conserved amino acid regions from different members CoV superfamilies such as SARS-CoV, murine CoV (murine hepatitis virus) and alpha CoV-1 species (Feline infectious peritonitis virus).

Primary sequence analysis resulted in two crucial functional domain regions both in N and C terminals of SARS-CoV-2. Interestingly, the NTD comprises RNA binding site, which signifies its importance towards a viral cellular mechanism. To its support, the available crystal structure of NTD SARS-CoV-2 N protein was retrieved and utilized in further study. The SARS-CoV-2 N protein had no binding site information including drug-binding sites till the end of March 2020, which influences the researchers to predict the drug-binding pocket in RNA binding domain of N protein. But recently, Kang et al. [30] reported about the crystal structure and showed the drug-binding pocket (including the amino acids Tyr 110, Tyr 112, Tyr 55, Ala56, and Arg89) of N protein with PDB ID 6M3M whereas this present study predicted the binding domain in SARS-CoV-2 N protein (PDB ID: 6VYO) with amino
The COVID-19 outbreak has caused havoc throughout the world, changing the course of human lives. Researchers are trying to design a vaccine against SARS-CoV-2 but that might take some time. This study attempts to find a drug for treating the disease condition, which will help to save human lives and mitigate the sufferings of millions of people infected by the virus worldwide. Some antivirals phytochemicals and synthetic drugs have been analyzed in this in silico study, which would target the N protein, responsible for replication of SARS-CoV-2 in the host body. Of all the compounds in this study, glycyrrhizic acid and theaflavin can be used as the antiviral drug, as they showed a higher binding affinity with the target protein. The effective drug candidates would be helpful to prevent the SARS-CoV-2 viral N protein and to reduce the risk of infection in the host body.

**ORCID**

Manisha Ray: https://orcid.org/0000-0001-7565-244X
Saurav Sarkar: https://orcid.org/0000-0001-9589-5115
Surya Narayan Rath: https://orcid.org/0000-0002-5458-8351

**Authors’ Contribution**

Conceptualization: SNR, MR. Data curation: MR, SNR, SS. Formal analysis: SNR, MR. Methodology: SNR, MR. Writing - original draft: MR. Writing - review & editing: SNR, SS, MR.

**Conflicts of Interest**

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

**Acknowledgments**

We are thankful to Dr. Pawan Kumar Agrawal, Vice chancellor, Odisha University of Agriculture and Technology for his moral support and valuable suggestions.

**References**

Spike protein D614G and RdRp P323L: the SARS–CoV–2 mutations associated with severity of COVID–19

Subrata K. Biswas†, Sonchita R. Mudi

†Department of Biochemistry and Molecular Biology, Bangabandhu Sheikh Mujib Medical University (BSMMU), Dhaka 1000, Bangladesh
‡Department of Biochemistry, Kumudini Women’s Medical College, Mirzapur 1940, Bangladesh

The severity of coronavirus disease 2019 (COVID–19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS–CoV–2), greatly varies from patient to patient. In the present study, we explored and compared mutation profiles of SARS–CoV–2 isolated from mildly affected and severely affected COVID–19 patients in order to explore any relationship between mutation profile and disease severity. Genomic sequences of SARS–CoV–2 were downloaded from Global Initiative on Sharing Avian Influenza Data (GISAID) database. With the help of Genome Detective Coronavirus Typing Tool, genomic sequences were aligned with the Wuhan seafood market pneumonia virus reference sequence and all the mutations were identified. Distribution of mutant variants was then compared between mildly and severely affected groups. Among the numerous mutations detected, 14408C>T and 23403A>G mutations resulting in RNA-dependent RNA polymerase (RdRp) P323L and spike protein D614G mutations, respectively, were found predominantly in severely affected group (>82%) compared with mildly affected group (<46%, p < 0.001). The 241C>T mutation in the non-coding region of the genome was also found predominantly in severely affected group (p < 0.001). The 3037C>T, a silent mutation, also appeared in relatively high frequency in severely affected group compared with mildly affected group, but the difference was not statistically significant (p = 0.06). We concluded that spike protein D614G and RdRp P323L mutations in SARS–CoV–2 are associated with severity of COVID–19. Further studies will be required to explore whether these mutations have any impact on the severity of disease.

Keywords: COVID–19, genome sequence, mutation, RNA-dependent RNA polymerase, SARS–CoV–2, spike protein

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS–CoV–2), the viral pathogen that causes coronavirus disease 2019 (COVID–19), has infected millions of people worldwide just in 8 months [1]. Although majority of the SARS–CoV–2 infected individuals recover after developing mild to moderate symptoms, more than 800,000 people have already been died due to severe form of COVID–19 [1]. Severity of COVID–19 has been found to greatly vary from patient to patient, but it is so far not entirely clear what is responsible for the variable severity of COVID–19 in the population [1,2]. We recently hypothesized that genetic variation in SARS–CoV–2 may explain the variable severity of COVID–19 [2]. To the best of our knowledge, it has not yet been investigated whether mutation profile of SARS–CoV–2 has any relationship with the severity of COVID–19.
However, some studies suggested that D614G mutation in the spike protein may contribute to increased infectivity or transmissibility of SARS-CoV-2 leading to increased severity of COVID-19 [3,4]. In fact, spike protein mediates viral entry into the host cell through binding of the virus with host cell receptor angiotensin-converting enzyme-2 (ACE2) [5]. The D614G mutation of the spike protein was observed sometimes in late January 2020 both in Europe and in China, but then this mutation spread first in the Europe and gradually globally [3,4]. Thus the distribution of spike protein D614G mutation has temporal and geographical variation. The RNA-dependent RNA polymerase (RdRp) enzyme of SARS-CoV-2 is the product of RdRp gene that catalyzes replication of viral RNA. Mutation in the RdRp gene was previously found to be associated with overall increase in mutation rate in the viral genome [6]. In the present study, we explored and compared mutation profiles of SARS-CoV-2 isolated from mildly affected and severely affected COVID-19 patients in order to explore relationship between mutation profile and disease severity.

**Methods**

Genomic sequences of SARS-CoV-2 were downloaded from Global Initiative on Sharing Avian Influenza Data (GISAID) database (https://www.gisaid.org). In the Browse option, we selected ‘complete’ and ‘w/Patient status’ to retrieve sequences that were complete (>29,000 bases in length) and had patient status information available. Then we explored ‘sample information’ of retrieved sequences, and included sequences in the mildly affected group if patient status was ‘mild’/‘asymptomatic’/‘not hospitalized’. On the other hand, we included sequences in the severely affected group if patient status was ‘severe’/‘ICU’/‘deceased’. Sequences with patient status described with ambiguous words like ‘released’, ‘hospitalized’, ‘alive’, ‘live’, ‘unknown’, etc. were excluded due to uncertainty whether the patients were mildly or severely affected. This sampling procedure is presented as a flow diagram in Fig. 1. Although there were 45,000 SARS-CoV-2 genomic sequences deposited in the GISAID website by June 12, 2020, only 2,443 complete sequences had patient status available. Following the above mentioned search and inclusion/exclusion criteria, we were able to include 46 sequences in the mildly affected group and 56 in the severely affected group (n = 102) in the present study.

Mutation profile was determined using the Genome Detective Coronavirus Typing Tool (available at https://www.genomedetective.com/app/typingtool/cov), a web-based bioinformatics pipeline that can accurately identify changes at nucleotides, coding regions and proteins using a novel dynamic aligner to allow tracking new viral mutations [7]. With the help of this coronavirus typing tool, each of the 102 SARS-CoV-2 sequences was aligned with the Wuhan seafood market pneumonia virus reference sequence NC_045512.3, and all the nucleotide and amino acid sequence variations were identified comparing with the reference sequence. Each of the mutations was counted for mildly and severely affected groups, and expressed in number and percentage. Distribution of selected mutant variants was compared between the mildly affected and severely affected groups by chi-squared test using SPSS version 21.0 (IBM Corp., Armonk, NY, USA). A p-value of less than 0.01 was considered statistically significant.

**Data availability**

The mutation profiles of the genomic sequences that support the findings of this study are available from the corresponding author upon reasonable request.

**Results**

The SARS-CoV-2 genomic sequences that we included in the present study were sequenced from viral isolates collected in the USA (number; mild, severe: 3, 3), Mexico (0, 3), Brazil (4, 1), Austria (0, 15), Russia (0, 13), Belgium (5, 8), Hungary (1, 0), Spain (1, 3), Turkey (1, 0), Bosnia and Herzegovina (0, 1), India (21, 6), Sri Lanka (0, 1), Japan (3, 0), Indonesia (0, 1), Lebanon (0, 1), Kuwait (1, 0), and Nigeria (6, 0). The viral isolates were collected between February 3 and May 27, 2020. There were 29 men, 13 women and 4 with gender information unavailable in the mildly affected group (n = 46), and 31 men and 25 women in the severely affected group (n = 56). Age distribution was 17 to 98 years.

![Flow diagram showing sampling procedure. GISAID, Global Initiative on Sharing Avian Influenza Data; ICU, intensive care unit; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.](https://doi.org/10.5808/GI.2020.18.4.e44)
years for mildly affected group and 17 to 93 years for severely affected group.

A comprehensive list of all mutations compared to the Wuhan reference sequence NC_045512.3 in mildly affected and severely affected groups is presented in Supplementary Table 1. In the coding region, there were 103 mutations in the mildly affected group with 37 silent and 66 missense mutations. In the severely affected group, there were 111 mutations with 40 silent and 71 missense mutations. In the non-coding region, there were 2 and 8 mutations in the mildly affected and severely affected groups, respectively, in the 5′ untranslated region (UTR); whereas, there were 9 and 15 mutations in the 3′ UTR in the mildly affected and severely affected groups, respectively. However, majority of the mutations appeared in low frequency, i.e., the mutations were found only in a few cases of mildly and severely affected groups (Supplementary Table 1), and therefore, those mutations were unlikely to be related to the severity of COVID-19.

Any mutation with a frequency of 5 or more in either mildly affected or severely affected group is presented in Table 1. In the open reading frame (ORF) 1ab of the SARS-CoV-2 genome, the most frequent mutation identified was 14408C > T at the nucleotide level. This mutation results in a missense mutation P4715L in the RNA-dependent RNA polymerase. This mutation was predominantly occurred in severely affected group (82.1%) compared with mildly affected group (45.7%, p < 0.001) (Table 1, Fig. 2). In ORF1ab, 11083G > T mutation at the nucleotide level caused L3606F mutation at the amino acid level and involved non-structural protein (nsp) 6. This mutation was found mainly in the mildly affected group (28.3%) compared with severely affected group (1.8%, p < 0.001). Another mutation 5700C > A in ORF1ab caused missense mutation A1812D at the amino acid level. This mutation in

<table>
<thead>
<tr>
<th>Gene/genomic region</th>
<th>Nucleotide variation</th>
<th>Amino acid variation</th>
<th>No. of mutations (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mild (n = 46)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Severe (n = 56)</td>
</tr>
<tr>
<td>5′ UTR</td>
<td>241C &gt; T</td>
<td>N/A</td>
<td>21 (45.7)</td>
</tr>
<tr>
<td></td>
<td>313C &gt; T</td>
<td>---</td>
<td>6 (13.0)</td>
</tr>
<tr>
<td></td>
<td>3037C &gt; T</td>
<td>---</td>
<td>21 (45.7)</td>
</tr>
<tr>
<td></td>
<td>5700C &gt; A</td>
<td>A1812D (nsp3)</td>
<td>7 (15.2)</td>
</tr>
<tr>
<td></td>
<td>8782C &gt; T</td>
<td>---</td>
<td>12 (26.1)</td>
</tr>
<tr>
<td></td>
<td>11083G &gt; T</td>
<td>L3606F (nsp6)</td>
<td>13 (28.3)</td>
</tr>
<tr>
<td></td>
<td>14408C &gt; T</td>
<td>P4715L (RdRp)</td>
<td>21 (45.7)</td>
</tr>
<tr>
<td></td>
<td>14805C &gt; T</td>
<td>---</td>
<td>5 (10.9)</td>
</tr>
<tr>
<td></td>
<td>15324C &gt; T</td>
<td>---</td>
<td>1 (2.2)</td>
</tr>
<tr>
<td></td>
<td>15957G &gt; T</td>
<td>---</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td></td>
<td>18877C &gt; T</td>
<td>---</td>
<td>1 (2.2)</td>
</tr>
<tr>
<td></td>
<td>20268A &gt; G</td>
<td>---</td>
<td>3 (6.5)</td>
</tr>
<tr>
<td>S (spike)</td>
<td>22468G &gt; T</td>
<td>---</td>
<td>8 (17.4)</td>
</tr>
<tr>
<td></td>
<td>23403A &gt; G</td>
<td>D614G</td>
<td>21 (45.7)</td>
</tr>
<tr>
<td></td>
<td>24197G &gt; T</td>
<td>A879S</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>ORF3a</td>
<td>25563G &gt; T</td>
<td>Q57H</td>
<td>5 (10.9)</td>
</tr>
<tr>
<td>M (membrane)</td>
<td>26735C &gt; T</td>
<td>---</td>
<td>1 (2.2)</td>
</tr>
<tr>
<td>ORF8</td>
<td>28144T &gt; C</td>
<td>L84S</td>
<td>12 (26.1)</td>
</tr>
<tr>
<td>N (nucleocapsid)</td>
<td>28854C &gt; T</td>
<td>S194L</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td></td>
<td>28878G &gt; A</td>
<td>S202N</td>
<td>8 (17.4)</td>
</tr>
<tr>
<td></td>
<td>28880G &gt; A</td>
<td>R203K</td>
<td>9 (19.6)</td>
</tr>
<tr>
<td></td>
<td>28883G &gt; C</td>
<td>G204R</td>
<td>9 (19.6)</td>
</tr>
<tr>
<td>3′ UTR</td>
<td>29742G &gt; A</td>
<td>N/A</td>
<td>8 (17.4)</td>
</tr>
<tr>
<td></td>
<td>29827A &gt; T</td>
<td>N/A</td>
<td>16 (34.8)</td>
</tr>
<tr>
<td></td>
<td>29830G &gt; T</td>
<td>N/A</td>
<td>20 (43.5)</td>
</tr>
</tbody>
</table>

Any mutation with a frequency of 5 or more in either mildly affected or severely affected group is included in this Table. Positions of nucleotides are numbered continuously irrespective of gene or genomic region, and positions of amino acids are numbered separately for each protein.

SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; COVID-19, coronavirus disease 2019; Mild, mildly affected group; Severe, severely affected group; UTR, untranslated region; ORF, open reading frame; N/A, not applicable; ---, silent mutation; nsp, non-structural protein; RdRp, RNA-dependent RNA polymerase.

https://doi.org/10.5808/GI.2020.18.4.e44
volved nsp3 and was found in 15.2% of mildly affected group but not found in severely affected group. Among the silent mutations present in ORF1ab, 3037C > T mutation was found more commonly in severely affected group (64.3%) compared with mildly affected group (45.7%, p = 0.06). Other silent mutations in ORF1ab appeared in relatively low frequency in mildly and severely affected groups (Table 1).

In the spike protein, 23403A > G mutation at the nucleotide level resulted in D614G mutation at the amino acid level, and it was predominantly found in severely affected group (85.7%) compared with mildly affected group (45.7%, p < 0.001) (Table 1, Fig. 2). In ORF3a, 25563G > T mutation at the nucleotide level resulted in Q57H mutation at the amino acid level, and this mutation was more prevalent in severely affected group (26.8%) compared with mildly affected group (10.9%, p = 0.08). In ORF8, 28144T > C mutation at the nucleotide level resulted in L84S mutation at the amino acid level. This mutation was found in 26.1% of mildly affected group and in 5.4% of severely affected group (p = 0.008). Other mutations affecting spike, membrane and nucleocapsid proteins of SARS-CoV-2 genome appeared in low frequency (Table 1).

Among all the mutations in the non-coding region of SARS-CoV-2 genome, the 241C > T mutation in the 5′ UTR appeared most predominantly in severely affected group (85.7%) compared with mildly affected group (45.7%, p < 0.001) (Table 1). The 29742G > A, 29827A > T, and 29830G > T mutations in the 3′ UTR appeared at a frequency of 17.4%, 34.8%, and 43.5%, respectively, in mildly affected group; but none of these mutations was...
found in the severely affected group (Table 1).

Of note, in the mildly affected group, four most common mutations (241C>T, 3037C>T, 14408C>T, and 23403A>G) coincided. In the severely affected group, however, 241C>T and 23403A>G coincided, and 3037C>T and 14408C>T occurred in subsets of them. There was temporal and geographical variation in the distribution of 23403A>G mutation that cause D614G mutation in the spike protein of SARS-CoV-2 [3,4]. In Table 2, we showed collection period of viral isolates in month and frequency of D614G mutation in mildly affected and severely affected groups. Although the percentage of D614G mutation gradually increased from February towards May in both groups, there was more D614G mutation in severely affected group compared with mildly affected group in March (68.2% vs. 45.8%) and April (96.4% vs. 60.0%) when most of the viral isolates were collected for sequencing (Table 2). This finding suggests that increased spike protein D614G mutation in severely affected group was unlikely to be due to temporal variation in the distribution of the mutation. However, in this study, majority of samples of mildly affected group were from India whereas those of severely affected group were from Europe, as described above. To explore whether this fact contributed to the increased spike protein D614G mutation in severely affected group, we showed frequency of D614G mutation in mildly affected and severely affected groups for India and Belgium in Table 3. Of note, the percentage of D614G mutation was found higher in severely affected group compared with mildly affected group in March (68.2% vs. 45.8%) and April (96.4% vs. 60.0%) when most of the viral isolates were collected for sequencing (Table 2). This finding suggests that increased spike protein D614G mutation in severely affected group was unlikely to be due to temporal variation in the distribution of the mutation.

<table>
<thead>
<tr>
<th>Month</th>
<th>Mildly affected group</th>
<th>Severely affected group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of samples</td>
<td>No. of D614G mutations (%)</td>
</tr>
<tr>
<td>March</td>
<td>24</td>
<td>11 (45.8)</td>
</tr>
<tr>
<td>April</td>
<td>15</td>
<td>9 (60.0)</td>
</tr>
<tr>
<td>May</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
<td>21 (45.7)</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Country</th>
<th>Mildly affected group</th>
<th>Severely affected group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of samples</td>
<td>No. of D614G mutation (%)</td>
</tr>
<tr>
<td>India</td>
<td>21</td>
<td>9 (42.8)</td>
</tr>
<tr>
<td>Belgium</td>
<td>5</td>
<td>3 (60.0)</td>
</tr>
</tbody>
</table>

Discussion

The severity of COVID-19 greatly varies from patient to patient. Majority of the patients either remain asymptomatic or develop mild to moderate symptoms. However, some COVID-19 patients who develop severe disease die even after hospitalization and intensive care [1]. Why the disease severity differs so much from one person to another is one of the mysteries scientists are still trying to solve [1]. The present study was designed to explore whether genetic variation in SARS-CoV-2 can explain variable severity of COVID-19. Mutation profiles of SARS-CoV-2 isolated from mildly affected and severely affected COVID-19 patients were explored and compared. Among numerous mutations observed in this study, two missense mutations, 14408C>T and 23403A>G, affecting RdRp and spike protein genes, respectively, were found most predominantly in the severely affected group compared with mildly affected group. Along with these two mutations, 241C>T in the 5′ UTR and a silent mutation 3037C>T in the ORF1ab were predominantly found in severely affected group (the later not significantly), however, these mutations do not alter amino acid sequence in a protein. Many other mutations that were found in low frequency in the present study are unlikely to exert an effect on the severity of COVID-19. Thus the ability of spike protein and RdRp mutations on the severity of COVID-19 needs to be considered.

The spike protein of SARS-CoV-2 is responsible for binding with host cell receptor ACE2, and thus it allows entry of the virus into the host cell [5]. In fact, the spike protein of SARS-CoV-2 has 10 to 20 folds higher affinity for ACE2 receptor than the corresponding spike protein of SARS-CoV [8]. Thus the spike protein is potentially related to the infectivity of SARS-CoV-2. The 23403A>G mutation in the genome of SARS-CoV-2 causes re-
placement of aspartic acid (D) with glycine (G) at position 614 (D614G) of the spike protein. This D614G spike protein mutation appeared sometimes in late January 2020 and then it has spread initially in Europe and then all over the world [3,4]. Several ways have been proposed through which spike protein D614G mutation may increase the infectivity of SARS-CoV-2 [3]. However, computer-based structural analysis of spike protein with D614G mutation suggested that the mutation is unlikely to alter its interaction with human ACE2 receptor [4]. But Korber et al. [3] found that patients infected with spike protein D614G mutant form of SARS-CoV-2 had higher viral loads since fewer PCR cycles were needed for their diagnosis. Furthermore, in cell culture experiment, viral particles with spike protein D614G mutation was found to infect ACE2 expressing cells more efficiently, and this increased infectivity was found to correlate with less shedding of S1 domain of spike protein and more incorporation of spike protein in the virion [9]. In spite of these facts, previous studies were unable to explore an association between the spike protein D614G mutation and disease severity due to relative lack of clinical data of the patients included in their studies [3,4].

As we found in the present study, previous studies also identified that the spike protein D614G mutation frequently accompanies a silent mutation 3037C>T and a missense mutation 14408C>T in ORF1ab [3]. The 14408C>T mutation in ORF1ab replaces a proline (P) with leucine (L) at position 4715 (P4715L) of ORF1ab polyprotein which actually appears as a replacement of proline with leucine at position 323 (P323L) of RdRp enzyme. The RdRp enzyme of SARS-CoV-2 catalyzes replication of viral RNA and it possesses proof-reading capability [6]. Thus a critical mutation in RdRp gene has the potential to alter viral replication capability with fidelity, and thereby a mutation in RdRp may contribute to infectivity of the virus and severity of the disease. The presence of 14408C>T mutation in SARS-CoV-2 genome that causes RdRp P323L mutation was found to be associated with overall increase in mutation rate in the viral genome [6]. Although the 14408C>T (RdRp P323L) mutation was predominantly found in severely affected patients in the present study, further studies will be required to elucidate whether this RdRp mutation has any significant impact on the viability and infectivity of SARS-CoV-2 and the severity of COVID-19.

The vast majority of genomic sequences of SARS-CoV-2 available at GISAID database do not contain patient status information. Even many of the sequences that contain patient status information use such ambiguous words to describe the information that do not reflect the severity status of the patient. That’s why we were unable to include large number of sequences to analyze in the present study. For the same reason, we were unable to include sequences in such a way that biases due to geographical, temporal, ethnic and gender variations could have been avoided. In spite of small sample size, month wise and country wise analyses of our data suggested that the temporal and geographical variation in the distribution of mutation did not influence our findings to a large extent. However, we do not know whether the patients included in the mildly affected group subsequently developed severe disease or not. In fact, GISAID database does not include any follow up information about patient status. Thus, it is another limitation of the present study, and to overcome this limitation, further studies will be required in which follow up information on patient status is available.

In spite of all these limitations, in the present study, to the best of our knowledge, we for the first time compared mutation profiles of SARS-CoV-2 between mildly affected and severely affected COVID-19 patients. Based on our findings, it may be concluded that the spike protein D614G and RdRp P323L mutations predominate in severely affected COVID-19 patients. Further studies will be required to explore whether spike protein D614G mutation or RdRp P323L mutation or the combination of both mutations can exert an impact on the severity of COVID-19.

ORCID

Subrata K. Biswas: https://orcid.org/0000-0002-2372-6133
Sonchita R. Mudi: https://orcid.org/0000-0002-0100-6506

Authors’ Contribution

Conceptualization: SKB. Data curation: SRM, SKB. Formal analysis: SKB, SRM. Funding acquisition: SKB. Methodology: SKB. Writing - original draft: SKB, SRM. Writing - review & editing: SKB.

Conflicts of Interest

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

Acknowledgments

We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GenBank and GISAID’s hCoV-19 database on which this research is based.

Supplementary Materials

Supplementary data can be found with this article online at http://
References


Introduction

In October, US President Donald J. Trump received Regeneron Pharmaceuticals’ experimental monoclonal antibody (mAb) cocktail REGN-COV2 as part of his treatment for coronavirus disease 2019 (COVID-19) when diagnosed with the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) on October 2, 2020. These and 11 other experimental mAb treatments targeting the SARS-CoV-2 S protein are undergoing human testing, with at least another 150 other antibodies in discovery research [1].

As the COVID-19 pandemic continues to unfold, interests have grown in COVID-19 antibody testing, also known as a serology test, as a way to measure how far the infection has spread in the population and to identify individuals who may be immune. Recently, many countries reported their population-based antibody titer study results. South Korea recently reported their third antibody formation rate, where it divided the study between the general population and the young male youths in their early twenties. As previously stated, these simple point estimates may be misinterpreted without proper estimation of standard error and confidence intervals. In this article, we provide an updated 95% confidence intervals for COVID-19 antibody formation rate for the Korean population using asymptotic, exact, and Bayesian statistical estimation methods. As before, we found that the Wald method gives the narrowest interval among all asymptotic methods whereas mid p-value gives the narrowest among all exact methods and Jeffrey’s method gives the narrowest from Bayesian method. The most conservative 95% confidence interval estimation shows that as of 00:00 November 23, 2020, at least 69,524 people were infected but not confirmed. It also shows that more positive cases were found among the young male in their twenties (0.22%), three times that of the general public (0.051%). This thereby calls for the quarantine authorities’ need to strengthen quarantine managements for the early twenties in order to find the hidden infected people in the population.

Keywords: confidence interval, COVID-19 antibody, retention rate
tested, the total sample size of patients, including those who passed by without knowing that they were infected with COVID-19, can be estimated [3].

These serological tests are known to be in use in other countries to figure out how many people in their population are infected with the potentially deadly virus. For example, results from Spain’s final stage of a nationwide antibody study shows that 5.2% of Spain’s population has been exposed to the new virus as of July 6, 2020 [4], 0.07% (11 out 14,000) for Taiwan (0.05% after age correction for the entire population) from the adult patients who visited the Taipei Longminzung hospital in May and July, 2020 [5], one out of four Mexican citizens were found to have antibodies to the novel coronavirus infection with 70% being asymptotic from August to November, 2020 [6], and 6% of all regions of the UK and 13% in London alone of citizens recruited from 20 June to 13 July 2020 [7], 2.5% of Georgia, USA, and 3.2% of the general residents of Wuhan, China, have antibodies [8].

South Korea also recently released its third antibody titer test results for the Korea National Health and Nutrition Examination Survey (KNHANES) and military enlistment in addition to previous results found [9]. Residents of 15 cities and provinces nationwide excluding Gwangju and Jeju participated in this third survey. In this study, a total of three showed positive antibody responses (2 of them even had neutralizing antibodies that neutralized the virus) out of 1,379 people who participated in the third round of the KNHANES from August 14 to October 31. The antibody formation rate calculated for undiagnosed confirmed cases excluding existing confirmed cases is 0.07% (1 out of 1,379). This is not significantly different from the results of the previous 1st survey (0.03%, 1 out of 3,553) and 2nd survey (0.07%, 1 out of 1,440). Also, a total of 25 people tested positive for 6,859 soldiers enlisted in September–October. Of these, 10 were confirmed patients, and the remaining 15 were infected by the local community. Therefore, the rate of formation of undiagnosed antibodies was 0.22% (15 out of 6,859), three times higher than 0.07% of the general public [8]. In total, 8,238 people were surveyed, giving a rate of antibody formation as 0.19%.

As stated previously in our article [3] on the same topic, the above-reported results in addition to being sparse, only captures sample proportion (point estimation) but does not provide its confidence interval which can be misleading to the general public, since confidence intervals give better interpretation to point estimation. Therefore, we report updated results for the point estimations along with proper interval estimations (95% confidence intervals) using asymptotic [10], exact [10,11], MidP [12], and Bayesian [13] statistical inferential methods already explained in detail in our previous article mentioned above.

Results

Table 1 presents the 95% confidence intervals for antibody results using the sum of total samples from the first and the second survey and only KNHANES samples from the third survey. It also presents the 95% confidence intervals for the total population in South Korea. The first two columns show the methods and the next two columns the 95% confidence interval for antibody retention rate in the samples. The second and third columns represent the estimated 95% confidence interval of antibody carriers in the Korean population by multiplying the total number of Korean population (51,289,593 people, as of December 19, 2020 [14]) with the antibody ratio (the proportion of samples with neutralizing antibodies provided as confidence intervals [CIs]). Note that this estimation was derived from a simple random sampling assumption, while the antibody sample does not represent the total Korean population (Data: 3 \( = 0 + 1 + 1 + 1 \) out of 5,874 \( = 1,555 + 1,500 + 1,440 + 1,379 \)).

![Table 1. 95% CIs for antibody results using the sum of total samples from the first and the second survey and only KNHANES samples from the third survey and for the total population in South Korea](https://doi.org/10.5808/GI.2020.18.4.e45)
From Table 1, Wald gives the minimum upper bound which is 55,906 whereas MidP provides the narrower confidence intervals among all types of confidence interval methods. Score CI from asymptomatic estimation methods and exact CI from exact estimation method are almost similar. Same goes for Likelihood ratio CI from asymptomatic estimation methods and Jeffrey’s CI from Bayesian estimation method.

Table 2 presents the 95% CIs of antibody results using the sum of samples of all cases (first, second, third [both KNHANES and military personnel]) and also for the total population in South Korea. The second and third columns show the methods whereas the next two columns show the 95% confidence intervals of antibody ratio for the total samples (Data: 18 = 0 + 1 + 1 + 15) out of 12,733 = 1,555 + 1,500 + 1,440 + 1,379 + 6,859).

The MidP-value method gives the narrowest interval among exact estimation method for Tables 1, 2, and Supplementary Table 1. The Bayesian method using the uniform prior gives the narrower interval than the Bayesian method using Jeffrey’s prior. For asymptotic estimation, Wald method gives the narrower interval for general and military personnel but when we consider all cases, likelihood ratio method gives the narrowest interval.

As the sample size increases, the confidence interval becomes narrower; which indicates that more accurate estimation of antibody formation rate is possible (Fig. 1). Through an actual test, the lower bound can be replaced by the number of confirmed patients in the confidence interval. Among the upper bound, the smallest value provides a conservative interpretation while the largest value provides a more aggressive interpretation.

Antibody titer testing helps in (1) discovery of neutralizing antibodies if present in the population. Neutralizing mAbs promise an adjunct to vaccines and traditional drugs in the treatment of COVID-19 [1]. (2) It measures how far the infection has spread in the population. For example, Mexico’s results show that about 31 million people in Mexico already have COVID-19, much higher than the current official count of about 1,267,000 cumulative confirmed cases as of December 17, 2020 [6]. By subtracting a day’s cumulative number of confirmed cases from the smallest upper limit, the result is interpreted as the minimum number of COVID-19 cases that were infected but not confirmed. For South Korea, as of 00:00 on November 11, 2020, at least 69,524 (= 100,528 – 31,004) [15] people were infected but not confirmed, higher than the previous value of 32,602 as of September 15, 2020. This can be interpreted as having a high probability of cumulative infection.

South Korea’s antibody formation rate of 0.0014 (0.14%) which suggests that about 71,805 (= 0.0014 × 51,289,593) people in the population have already been exposed to COVID-19 virus as of November 23, 2020, is lower than for most foreign countries. For example, recently USA, Italy, and Sweden [16], Mexico (24.8%), Spain (5.2%), and UK (6%) have higher antibody rate but Taiwan has the lowest here of 0.07%. This demonstrates the adequacy and effectiveness of Taiwan’s quarantine measures in mitigation and suppression of the virus. Although South Korea’s antibody titer result (0.14%) is lower than for most foreign countries, more positive cases were found among the young male people in their twenties (0.22%), three times that of the general public (0.051%) (Supplementary Table 1). This means that there are relatively many infected people among the younger generation but undiscovered since most younger age groups even if infected, are asymptomatic or mild. And given that they are active in social activities without receiving medical treatment or examination at a medical institution, the risk of spreading infection among the population is quite high. Therefore, this calls for the quarantine authorities’ need to strengthen quarantine managements for the early twenties in order to find hidden infected people in the population. However, the difference in these proportion between the young people and the general population is not statistically significant (p = 0.2491) by Fisher exact test.

### Table 2. 95% CIs of antibody results for all three samples and for the total population in South Korea

<table>
<thead>
<tr>
<th>Point estimation</th>
<th>Interval estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>18/12733 = 0.0014 (0.14%)</td>
<td>Asymptotic estimation</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Exact estimation</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayesian estimation</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CI, confidence interval.

https://doi.org/10.5808/GI.2020.18.4.e45
Fig. 1. Confidence intervals (CIs) for the first to third antibody titer test results. The black line indicates the coronavirus disease 2019 cumulative confirmed cases. KNHANES, Korea National Health and Nutrition Examination Survey.

References


Authors’ Contribution

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

Conflicts of Interest

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

Supplementary Materials

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


Introduction

As a means for the functional study of genes at the genome-wide level, a genome-wide screening method based on the loss of function of genes is useful. Various methods have been developed for this purpose, including random mutagenesis-based gene knockout, RNA interference-based knockdown, homologous recombination-based deletion with barcode technology, and CRISPR/Cas9-based knockout [1-6]. With a careful design of genome-wide barcodes and decoding of barcode information by next-generation sequencing, those tools and libraries have been widely used to characterize gene function and to discover novel drug targets in various organisms.

As genome-wide barcode screening tools have been widely used, many algorithms and tools have been developed to help researchers to analyze their genome-wide barcode screening data [6-8]. Among the most widely used tools are MAGeCK and MAGeCK-flute, which provide a comprehensive suite of tools from quality control to data analysis and visualization using the R statistical language system [6-8]. While useful, those tools are not easy to use for many researchers with limited bioinformatics skills, as they were developed to operate on a command line basis in the Linux operating system. Therefore, it is necessary to develop a tool that can be easily accessed by general researchers who perform functional genomics research using genome-wide pooled screening data. Here,
we present BaSDAS (Barcode-Seq Data Analysis System) as a user-friendly web service that is useful for the analysis of genome-wide pooled barcode screening data using next-generation sequencing technology.

**Pipeline Overview**

BaSDAS is an automated pipeline that analyzes CRISPR-Cas9 knockout screening data in four steps: (1) data input, (2) primary analysis using the MAGeCK algorithm [7], (3) downstream analysis and visualization, and (4) generation of output.

We used PHP for the creation of the web interface, R for the data analysis at the server side, Python for data handling, and a MySQL database for the effective management of data and analysis jobs of multiple users (Fig. 1).

The BaSDAS system receives read-count file and analysis parameters as input data and then processes the data by using MAGeCK’s robust ranking aggregation or maximum likelihood estimation algorithm [7] depending on the experimental design. The required parameters include the type of experiment design and source organism information (Supplementary Fig. 1). The counts of the sgRNA or barcode sequence are analyzed in the primary analysis steps, and the results are used for various secondary analyses including plots of negative and positive selection of genes, enrichment analysis, and visualization [8].

**Analysis of Data**

The main functions of the BaSDAS system are (1) analyzing positively or negatively selected genes with the user tag-read count file from genome-wide screening experiments as a primary analysis (Fig. 2B and 2C) and (2) conducting a secondary in-depth analysis to estimate the molecular functions or pathways to which the selected gene groups belong (Fig. 2D–2F). The analysis of the user gene knockout screening data is conducted through the following procedure (Supplementary Fig. 1): (1) clicking on the menu item “Analysis,” (2) selecting an experiment model, (3) selecting a source organism, (4) entering the user’s e-mail address, (5) selecting and uploading the user data file (read-count file), (6) selecting the samples for each condition group, (7) submitting the analysis job, (8) monitoring the analysis process, and (9) completing the analysis.

**Generation of the Analysis Results**

After the completion of the analysis, users can access to the analysis report in one of the following three ways: (1) by clicking the hyperlink to the analysis report provided on the job monitoring page, (2) by clicking the hyperlink to the analysis report provided in the results table of the job search function, or (3) by clicking the hyperlink to the analysis report in the e-mail reporting that the analysis is complete. Any of those links lead to the analysis report page shown in Fig. 2. The content of the analysis report depends on the selected experimental models. The results of the in-depth analysis and visualization can be revised or modified by repeated reanalyses with various parameters. The total analysis results, from the initial analysis or reanalysis, can be downloaded as a single compressed file through the user’s web browser. The compressed report file includes (1) primary gene selection results, (2) the results of an in-depth analysis and their visualization, and (3) a report in PDF or HTML format.

**Revision of the In-depth Analysis Results**

Various plots are given in the analysis report to help users to understand the results through an intuitive visualization. If it is neces-
sary to revise the plotting range or the content of a plot, the user can proceed to re-analyze and plot the data repeatedly by modifying the parameters of the plots or analysis (Fig. 2C). The re-analyzed results and graphs are included in the final analysis report as PDF or HTML files and finally provided as a compressed file. For re-analyzable graphs, the graph editing buttons are provided at the right end of the graph title in the analysis report page. By clicking the graph editing buttons, users can access the parameter-setting table for reanalysis and plot the graphs again. In the table, parameter values used for the current results and plots are listed and the input fields for the parameters—cutoff values, maximum number of genes or pathways, positive or negative selection, and so on—are provided to reset or change the parameters.

Future Work

Our ultimate goal is to construct a comprehensive analysis environment for the comparative analysis and interpretation of ge-
name-wide pooled screening experiments by building a database of various public pooled CRISPR-Cas9 screening data. Toward this goal, we will collect public genome-wide screening datasets from diverse organisms, construct a database, and also update the comparative analysis modules.

**Conclusion**

We have developed a tool that allows researchers with limited bioinformatics skills to easily and effectively analyze their pooled genome-wide screening data. Our system provides many useful functions such as quality control, median normalization, sgRNA mean-variance modeling, sgRNA ranking, and identification of essential genes and enriched pathways from the knockout tag read sequences obtained by genome-wide barcode screening experiments. By developing a GUI-based interface, user convenience is maximized, and various secondary statistical analyses and visualization functions are provided for an intuitive interpretation of the given results. In the future, we plan to build a comprehensive analysis environment for comparative analysis and downstream research by collecting public pooled CRISPR-Cas9 knockout screening data and analysis results and converting them into a database in this BaSDAS system. We hope that BaSDAS will provide researchers a useful tool to effectively analyze and interpret their data to support the development of novel drugs based on functional genomics information.

**ORCID**

Young-Kyu Park: https://orcid.org/0000-0001-8283-7813
Byoung-Ha Yoon: https://orcid.org/0000-0002-5663-5000
Seung-Jin Park: https://orcid.org/0000-0003-0666-605X
Byung Kwon Kim: https://orcid.org/0000-0002-1030-7730
Seon-Young Kim: https://orcid.org/0000-0002-1030-7730

**Authors’ Contribution**

Data curation: SJP. Formal analysis: BHY, SJP. Funding acquisition: SYK. Methodology: YKP, BKK. Writing - original draft: YKP, BHY. Writing - review & editing: SYK, BKK.

**Conflicts of Interest**

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

**Acknowledgments**

This research was supported by the National Research Foundation of Korea (NRF) grant (NRF-2014M3C9A3068554 and NRF-2017MBA9BS060884) funded by the Korea government (MST) and a grant from the KRIBB Research Initiative Program.

**Supplementary Materials**

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

**References**

Instructions for authors

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 allows 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
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, author's should contact the editorial office.

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

**Detailed Description of Use of Articles of Genomics & Informatics**

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

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

**Copyrights**
Publisher applies the Creative Commons Attribution Non-
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.

<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 xxx@yyy. To gain access, data requestors will need to sign a data access agreement.</td>
<td>Proposals may be submitted up to 36 months following article publication. After 36 months, the data will be available in our University's data warehouse but without investigator support other than deposited metadata. Information regarding submitting proposals and accessing data may be found at (link to be provided).</td>
<td>Not applicable</td>
</tr>
</tbody>
</table>

Table 1. Examples of data sharing statements that fulfill ICMJE requirements

ICMJE, International Committee of Medical Journal Editors.
*These examples are meant to illustrate a range of, but not all, data sharing options.

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

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

**Contact address**

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

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

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

<table>
<thead>
<tr>
<th>Title of article</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td></td>
</tr>
<tr>
<td>Author’s signature</td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td></td>
</tr>
</tbody>
</table>

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.