Single-cell RNA sequencing identifies distinct transcriptomic signatures between PMA/ionomycin- and αCD3/αCD28-activated primary human T cells

Jung Ho Lee, Brian H Lee, Soyoung Jeong, Christine Suh-Yun Joh, Hyo Jeong Nam, Hyun Seung Choi, Henry Sserwadda, Ji Won Oh, Chung-Gyu Park, Seon-Pil Jin, Hyun Je Kim
Aims and scope

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

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

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

Published by the Korea Genome Organization
Contact information
Park, Taesung, Editor-in-Chief

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 2023, 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 Board

## Editor in Chief

Park, Taesung  
Seoul National University, Korea

## Advisory Editorial Board

- Batzer, Mark A.  
  Louisiana State University, U.S.A.
- Church, George M.  
  Harvard University, U.S.A.
- Lee, Byungkook  
  National Institute of Health, U.S.A.
- Matsuda, Fumihiko  
  Kyoto University, Japan
- Sakaki, Yoshiyuki  
  RIKEN Genomic Science Center, Japan
- Seo, Jeong-Sun  
  Seoul National University, Korea
- Cho, Soo Young  
  National Cancer Center, Korea
- Choi, Murim  
  Seoul National University, Korea
- Choi, Ik-Young  
  Kangwon National University
- Chung, Yujin  
  Kyonggi University, Korea
- Han, Kyudong  
  Dankook University, Korea
- Huh, Iksoo  
  Seoul National University, Korea
- Huh, Sun  
  Hallym University, Korea
- Kim, Hyunsoo  
  Chungnam National University
- Kim, Sangsoo  
  Soongsil University, Korea

## Associate Editors

- Cho, Soo Young  
  National Cancer Center, Korea
- Choi, Murim  
  Seoul National University, Korea
- Choi, Ik-Young  
  Kangwon National University
- Chung, Yujin  
  Kyonggi University, Korea
- Han, Kyudong  
  Dankook University, Korea
- Huh, Iksoo  
  Seoul National University, Korea
- Huh, Sun  
  Hallym University, Korea
- Kim, Hyunsoo  
  Chungnam National University
- Kim, Sangsoo  
  Soongsil University, Korea
- Nam, Seungyoon  
  Gachon University, Korea
- Oh, Bermseok  
  Kyung Hee University, Korea
- Oh, S. June  
  Inje University, Korea
- Park, Hyun Seok  
  Ewha Womans University, Korea
- Song, Minsun  
  Sookmyung Women University, Korea
- Yoon, Kyong-Ah  
  Konkuk University, Korea
- Won, Sungho  
  Seoul National University, Korea
- Woo, Hyun Goo  
  Ajou University, Korea

## Editorial Board

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

## Ethics Editor

Chung, Yeun-Jun  
The Catholic University of Korea, Korea

## Layout Editor

Jeong, Eun Mi  
M2Pi, Korea

## Statistics Editor

Won, Sungho  
Seoul National University, Korea

## Website and JATS XML File Producer

Im, Jeonghee  
M2Pi, Korea

## Manuscript Editor

Chang, Soo Hee  
Infolumi, Korea
COVID-19 progression towards ARDS: a genome wide study reveals host factors underlying critical COVID-19
Shama Mujawar, Gayatri Patil, Srushti Suthar, Tanuja Shendkar, Vaishnavi Gangadhar

e-Pharmacophore modeling and in silico study of CD147 receptor against SARS-CoV-2 drugs
Nisha Kumari Pandit, Simranjeet Singh Mann, Anee Mohanty, Sumer Singh Meena

Single-cell RNA sequencing identifies distinct transcriptomic signatures between PMA/ionomycin- and αCD3/αCD28-activated primary human T cells
Jung Ho Lee, Brian H Lee, Soyoung Jeong, Christine Suh-Yun Joh, Hyo Jeong Nam, Hyun Seung Choi, Henry Sserwadda, Ji Won Oh, Chung-Gyu Park, Seon-Pil Jin, Hyun Je Kim

Cell line-specific features of 3D chromatin organization in hepatocellular carcinoma
Yeonwoo Kim, Hyeokjun Yang, Daeyoung Lee

The genomic landscape associated with resistance to aromatase inhibitors in breast cancer
Kirithika Sadasivam, Jeevitha Priya Manoharan, Hema Palanisamy, Subramanian Vidyalkshmi

Functional annotation of uncharacterized proteins from Fusobacterium nucleatum: identification of virulence factors
Kanchan Rauthan, Saranya Joshi, Lokesh Kumar, Divya Goel, Sudhir Kumar

Bioinformatic analyses reveal the prognostic significance and potential role of ankyrin 3 (ANK3) in kidney renal clear cell carcinoma
Keerakarn Somsuan, Siripat Aluksanasuwan

Insight from sirtuins interactome: topological prominence and multifaceted roles of SIRT1 in modulating immunity, aging, and cancer
Nur Diyana Zulkifli, Nurulisa Zulkifle

Comparison of digital PCR platforms using the molecular marker
Cherl-Joon Lee, Wonseok Shin, Minsk Song, Seung-Shick Shin, Yujun Park, Kornsorn Srikulnath, Dong Hee Kim, Kyudong Han

Computational approaches for molecular characterization and structure-based functional elucidation of a hypothetical protein from Mycobacterium tuberculosis
Abu Saim Mohammad Saikat

A bioinformatic approach to identify pathogenic variants for Stevens-Johnson syndrome
Muhammad Ma’ruf, Justitia Cahyani Fadli, Muhammad Reza Mahendra, Lalu Muhammad Irham, Nanik Sulistyani, Wirawan Adikusuma, Rockie Chong, Abdi Wira Septama

Assessing the impact of recombination on the estimation of isolation-with-migration models using genomic data: a simulation study
Yujin Chung
COVID-19 progression towards ARDS: a genome wide study reveals host factors underlying critical COVID-19

Shama Mujawar*, Gayatri Patil, Srushti Suthar, Tanuja Shendkar, Vaishnavi Gangadhar

MIT School of Bioengineering Sciences and Research, MIT-Art, Design and Technology University, Loni Kalbhor, Pune 412201, India

Coronavirus disease 2019 (COVID-19) is a viral infection produced by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus epidemic, which was declared a global pandemic in March 2020. The World Health Organization has recorded around 43.3 billion cases and 59.4 million casualties to date, posing a severe threat to global health. Severe COVID-19 indicates viral pneumonia caused by the SARS-CoV-2 infections, which can induce fatal consequences, including acute respiratory distress syndrome (ARDS). The purpose of this research is to better understand the COVID-19 and ARDS pathways, as well as to find targeted single nucleotide polymorphism. To accomplish this, we retrieved over 100 patients’ samples from the Sequence Read Archive, National Center for Biotechnology Information. These sequences were processed through the Galaxy server next generation sequencing pipeline for variant analysis and then visualized in the Integrative Genomics Viewer, and performed statistical analysis using t-tests and Bonferroni correction, where six major genes were identified as DNAH7, CLUAP1, PPA2, PAPSS1, TLR4, and IFITM3. Furthermore, a complete understanding of the genomes of COVID-19-related ARDS will aid in the early identification and treatment of target proteins. Finally, the discovery of novel therapeutics based on discovered proteins can assist to slow the progression of ARDS and lower fatality rates.

Keywords: acute respiratory distress syndrome, COVID-19, JAK-STAT pathway

Introduction

The coronavirus disease 2019 (COVID-19) epidemic began in Wuhan, China, and quickly spread over the world, causing the World Health Organization (WHO) to proclaim a global pandemic on March 11, 2020 [1]. Before symptoms appear, this illness spreads through coughing and sneezing by symptomatic patients and asymptomatic carriers [1]. According to the WHO, there have been 638,175,811 cases, of which 6.6 million deaths have been reported as of November 2022 [2]. This virus infects the respiratory system, which is the most common problem associated with COVID-19, causing dyspnea and acute respiratory symptoms that can develop to refractory pulmonary failure [3]. Acute respiratory distress syndrome (ARDS), a life-threatening form of respiratory failure, is frequent among COVID-19 patients. According to weighted averages generated from several studies with COVID-19 data, ~33% of hospitalised patients had ARDS, affecting about three-fourth ~75% of the COVID-19 patients admitted to the intensive care unit and ~45% of those cases were lethal [3,4]. Berlin criteria state that COVID-19-in-
duced ARDS is significantly different from ARDS caused by other reasons, and as a result, it requires a different approach to care [3,5]. COVID-19 ARDS takes 8 to 12 days to appear, whereas the Berlin ARDS criteria restricts the onset period to one week [3,6]. Individuals with COVID-19 ARDS may have normal or even high lung compliance, which is not the case with patients who have traditional ARDS requiring [3]. COVID-19 separates the severity of ARDS into three groups based on specificity: mild, mild-moderate, and moderate-severe [3,6].

Due to the obvious heterogeneity of ARDS, identifying effective treatments becomes difficult [7,8]. Patients with various types of ARDS have undergone several therapies. In cases with typical ARDS, recruitment exercises, high-dose corticosteroids, and ongoing neuromuscular blocking medications are the most often employed adjuvant therapy [3]. Numerous studies have been conducted, and they all refer to ventilator management using mechanical ventilation as the most effective treatments [9]. Mechanical ventilation works on limiting the tidal volume and pressure and maintaining positive end expiratory pressure [7-9]. The use of appropriate therapies combined with a pharmacological approach was another method for treating ARDS. Due to its anti-inflammatory properties from both a preventive and a therapeutic perspective, corticosteroids are currently the subject of numerous clinical trials testing their efficacy in treating ARDS [3,9,10]. Other clinical trials using therapies like statins, heparin, and aspirin are also ongoing [10]. Utilizing stem cells to restore damaged tissue and creating 3D models of the lungs, alveoli, and bronchial tree to test treatments are two further modern scientific approaches [10].

Given the limited number of effective treatments for established ARDS, ARDS research has shifted its strategic focus to identifying patients with or at high risk of ARDS early in the course of the underlying illness, when strategies to prevent ARDS and associated organ failures can be systematically evaluated. Patients with severe and critical COVID-19 were highly advised to get systemic corticosteroid therapy, however those with non-severe COVID-19 were not [2,3]. Although there are presently no proven effective treatments for COVID-19, a number of medications are being tested in clinical research, including lopinavir-ritonavir, remdesivir, ruxolitinib, and tocilizumab [3,11]. Despite significant advances in our knowledge and treatment of ARDS patients, it still has a high morbidity and mortality rate [9,10,12].

Other than the mechanical and therapeutic approaches, genomic studies are going on for better prognosis and treatment of ARDS. The identification of the virus’s genome for its various strains has been made possible by the wide use of whole genome sequencing. Genome analysis studies have aided in the identification of COVID-19 transmission and progression to ARDS with the details of the genes involved in the process. Biomarkers are used to understand the molecular pathways and pathology of the disease by analyzing the variants in the genetic variations [13]. Later, with the use of sequence analysis, after numerous researches, we were able to learn about the phylogenetic links between the virus strains as well as other genomic variation and mutation that occurred in it [14]. Next-generation sequencing (NGS) was used to determine the expression characteristics of genes in different organ systems and particular cells after infection with the virus [15]. Eight genes were found to be linked to the molecular pathway in a study of 1,778 infected cases that included quality control and the variants were identified and mapped to their appropriate genes, chromosomes, and associated medical information [16]. In this study, 26 single nucleotide polymorphisms (SNPs) in genes and 12 unique candidates with biological function or strong supporting evidence were discovered, pointing to a high likelihood of participation in altering the clinical phenotype of COVID-19 [17].

The goal of this paper is to understand the progression of COVID-19 to ARDS and to identify probable genes associated in the pathway for which SNP variation analysis is performed on sample sequences to locate SNPs. We worked with 100 samples that were obtained from National Centre for Biotechnology Information (NCBI)’s Sequence Read Archive (SRA) sequences. The data consisted of RNA sequences obtained from patient reverse transcription polymerase chain reaction tests in the Texas, USA region. These sequences were further processed using the Galaxy server pipeline, which offers a wide range of tools for quality control, SNP analysis, and annotation in one location. The pipeline is described further in the paper, along with the tools utilised and the findings. The SNPs were then mapped in Integrative Genomics Viewer (IGV) to identify probable genes and chromosomes implicated in the pathophysiology of COVID-19 and the ARDS pathway. In our study, we also aim to perform statistical analysis. The null hypothesis we formulated stated that these genes do not contribute to the progression of the disease. Our analysis involved conducting t-tests on the data obtained from various samples [18].

Methods

Pathway analysis

In a severe acute respiratory syndrome coronavirus (SARS-CoV) animal model, it has been demonstrated that type I interferon (IFN) dysregulation and inflammatory monocyte macrophages result in fatal pneumonia. Therefore, elevated type I IFN secretion and infiltrating myeloid cells have a detrimental effect on the
course of the infection and are important contributors to lung damage and function. Dendritic cells deliver pathogen antigens to T cells in lymph nodes, facilitating the cooperation of the innate and adaptive immune systems (Fig. 1).

The cytotoxic T cells recognise the infection and kill the pathogen-infected cells as a result. On the other side, helper T cells trigger IFN-II synthesis, also develops a pathogen memory and stimulates B cells, enabling them to specialise and make pathogen antibodies [19]. Production of interleukin (IL)-10, transforming growth factor, and Treg (regulatory T cell) cells reduces inflammation and re-establishes homeostasis. According to study, CD4+ T cells that produce IL-10 are more prevalent in ARDS patients. In a small proportion of infected individuals, these immune reactions can completely stop viral replication or eradicate virus infection. Others describe a partial viral suppression, a decline in the number of B and T cells in circulation, and then an unknown mechanism. Some people’s ARDS caused by COVID-19 has been connected to cytokine storm, a risky complication brought on by protracted viral replication. Some of the cytokines and chemokines that were overexpressed during the cytokine storm were IL-1, IL-2, IL-6, IL-10, tumor necrosis factor, IFN-γ, IFN-gamma-inducible protein 10, and macrophage inflammatory protein 1. Serum IL-6 levels have been associated with the severity of illness and mortality, suggesting a crucial role for IL-6. As IL-6 circulates, it binds to soluble IL-6 receptors and creates a complex with a gp130 dimer on the surface of certain cells. The complex stimulates JAK-STAT3 in a variety of cell types, including endothelial cells, in a subgroup of hospitalised COVID-19 patients, resulting in a cytokine storm and potentially fatal symptoms like ARDS (Fig. 2) [9,20].

Data collection
The paired Illumina RNA data (Illumina, San Diego, CA, USA) was collected in FASTQ format from the NCBI [21]. Considering that SARS-COVID-19 data is produced using NGS technology, the NCBI SRA includes integrated information [22]. We retrieved a total of 100 SRA sequences in fastq.gz format from Texas, USA, for this objective in preventing variability from various locations, which assisted in limiting within the memory constraint of a large dataset (Supplementary Table 1).

Below filters were applied to collect data:
- Source: RNA
- Type: Genome
- Library layout: Paired
- Platform: Illumina

![Fig. 1. Immunopathology of cells. IFN, interferon; IL-10, interleukin 10; TGF-β, transforming growth factor β.](https://doi.org/10.5808/gi.22080)
Since the data was large in number, we made 5 datasets of 20 SRR runs in fastq.gz format for Variant analysis. Additionally, genetic changes (variants) between healthy and diseased tissue, individuals in a population, or strains of an organism can provide mechanistic insight into disease processes and the natural function of impacted genes (Fig. 3) [23].

**Workflow for SNP analysis**

SNP analysis was done using Galaxy server [24]. It is a web-based platform to perform analysis of NGS data. Galaxy allows for the characterization, analysis, and computational visualization of genetic data with comparatively minimum resources. Galaxy platform is widely used for developing pipeline for NGS data analysis because of its comparatively high level of accessibility, reproducibility, transparency, and scalability. Galaxy analysis interface has numerous potent tools to perform bioinformatics analysis [25]. Therefore, Galaxy server is used for variant analysis. Data downloaded from NCBI, firstly uploaded on Galaxy server. A pipeline for SNP analysis was developed using various tools in the Galaxy server. Different tools used for SNP analysis are FastQC for quality check and analysis, Trimmomatic tool used for trimming reads, SRA sequence aligned with reference human hg38 sequence using HISAT2 tool. SNPs were identified using the FreeBayes tool. Variants annotated by snpEff tool (Fig. 4).

**Quality analysis**

Millions of sequences are produced in a single run by modern high-throughput sequencing technologies like Illumina. Therefore, a quality control check is required as the initial step before using raw data for analysis. It is done to determine whether the data is good or if there are any issues with the data. Galaxy has a FastQC (Galaxy version 0.73+galaxy0) tool for quality check of Fastq files. It is a highly effective and popular tool for quality analysis. FastQC analyses fastq files and produces HTML file as an output which can be viewed in Galaxy workspace. The quality report generated gives information about per base sequence quality, sequence content, N and GC content number of reads and their length [26]. Therefore, the FastQC tool was used to perform quality analysis of data. The downloaded SRA runs (fastq.gz) were added to history in Galaxy. FastQC Read Quality reports (Galaxy version 0.73+galaxy0) this tool was selected from the tool menu on the left side. Then the datasets were uploaded in it and executed. The output of FastQC tool contains basic text and HTML file contain-
Fig. 3. Exploring and monitoring coronavirus disease 2019 (COVID-19) variants with Galaxy.

ing information about Basic Statistics, Per base sequence quality, Per sequence quality scores, Per base sequence content, Per base GC content, Per sequence GC content, Per base N content, Sequence Length Distribution, Sequence Duplication Levels, Overrepresented sequences, Kmer Content.

**Data cleaning**

FastQC report generates per base sequence quality graph, it shows quality of reads and it is found that in some dataset samples failed in quality checks. A number of the quality measures utilised by FastQC were not met by several of the samples. However, this does not imply that samples should be discarded. Some quality indicators frequently fail, and this may or may not cause issues for your downstream application. To lower false-positive rate owing to sequencing errors, it is required to eliminate some of the low-quality sequences from variant calling method.

To remove low-quality reads and bases from samples, Galaxy has a tool called Trimmomatic (Galaxy version 0.38.0). Trimmomatic is a read trimming tool for Illumina NGS data. It is a flexible tool providing several functions to be operated on reads. These functions include trailing, leading, and several other quality control operations. After performing Trimmomatic on poor quality data, made a dataset collection of good quality SRR reads and trimmomatic output of poor-quality reads.
Alignment
Alignment of sequences is done to identify the origin of reads. The most time-consuming process in analysis of RNA sequences is alignment of reads against reference genome. So, for fast alignment with less usage of memory of the system, HISAT2 A fast and sensitive alignment program (Galaxy version 2.2.1+galaxy0) tool is used. HISAT2 is hierarchical indexing for spliced alignment of transcripts. A whole-genome Ferragina-Manzini (FM) index is used by HISAT to anchor each alignment, and several local FM indexes are used to quickly extend these alignments. This indexing method is based on the Burrows-Wheeler transform and the FM index \[27\]. The reference genome used is the Homo sapiens (human) genome assembly GRCh38 (hg38).

Identification of SNPs
After alignment, the next step is to find out SNPs and MNPs i.e., single nucleotide polymorphism and multi-nucleotide polymorphisms, respectively. The efficient tool available at Galaxy server for SNP identification is Freebayes (version 1.3.1, Galaxy version 1.3.1). It can also determine insertion and deletions i.e., indels and substitutions in the dataset.

Freebayes (version 1.3.1, Galaxy version 1.3.1) will generate a VCF dataset representing SNPs, indels, and complex variations in samples in the input alignments given some BAM dataset(s) and a reference sequence. From the left side tool menu bar, Freebayes is selected and Bam dataset uploaded. The reference genome used is the Homo sapiens (human) genome assembly GRCh38 (hg38) and executed. The variants were identified in .VCF files. The output of Freebayes was given to the next step.

SNP analysis
These analyzed variants need to be annotated. Variants are changes in the sequence to be analyzed when compared with the reference genome. For annotation Galaxy has SnpEff eff: annotate variants (Galaxy version 4.3+T,galaxy) tool. Variant Call Format (VCF) file is given as input to this tool which has predicted variants. SnpEff eff tool performs annotation of variants and also determines the effects of these variants. Annotation is the process of describing variant information like whether these variants have effect on protein coding or they are present in gene, exon, etc.

VCF file obtained from FreeBayes tool uploaded in SnpEff eff tool. SnpEff Genome version Name used was hg38. Annotation
output selected was Use gene ID instead of gene name (VCF output).

Genome visualization
The IGV (version 2.12.3) is practical, quick, and easy to use software. The IGV is frequently employed for the visualization of genomic data. The real-time visualization of various genomic data sets is made possible by IGV. IGV has the benefit of operating efficiently on a regular desktop computer while consuming the least number of resources. IGV is particularly suited for genome-wide examination of NGS data sets due to its scalable architecture. This high-performance viewer provides a seamless and simple user experience while handling large heterogeneous datasets efficiently. A significant aspect of IGV is its emphasis on the integrative aspect of genomic research, with support for data from both array-based and NGS technologies as well as the blending of clinical and phenotypic information [28,29].

Steps followed for visualization are firstly, IGV was downloaded on desktop. 6 ‘.vcf’ files were uploaded in the IGV. The reference genome that we used is the Homo sapiens (human) genome assembly GRCh38 (hg38). Then SNPs were visualized by varying resolution scales.

Statistical analysis
We applied the Bonferroni correction method to adjust the p-values obtained from the t-tests. This correction is a rigorous statistical technique that reduces the likelihood of type 1 errors by accounting for multiple comparisons. By implementing the Bonferroni correction, we aimed to ensure the reliability and validity of our results by minimizing the possibility of falsely identifying genes as significant contributors to the progression of ARDS.

The t-test was chosen as our statistical tool for several reasons. Firstly, our data met the assumptions of a parametric test, meaning that the distribution of the data after appropriate cleaning and preprocessing was approximately normal. Additionally, the t-test accommodates situations where the variance of the population is unknown, making it a suitable choice for our analysis.

Results and Discussion
The SNPs, which involve modifications to a single DNA building block, are the most prevalent form of genetic variation in people. They can be applied to study drug response variability and locate disease-causing genes. This has significant implications for the development of genome-based diets and safer drugs as well as personalised treatment. SNPs can also aid in the study of sequence evolution and the underlying biological processes, particularly the involvement of selective forces in human disease.

Quality analysis
FastQC performed on each read individually and generated output reports were analyzed. Some reads had a good quality score (above 20) lying in the green zone while some had poor quality (less than 20) which has outliers or some part of the sequence in the red zone. These files contained Sanger/Illumina 1.9 encoding of quality values. The Per base sequence quality were analyzed using following criteria:
1. For each position a Box-Whisker type plot is drawn.
2. The median value is represented by the central red line.
3. The inter-quartile range (25%–75%) is represented by the yellow box.
4. The upper and lower whiskers correspond to the 10% and 90% points, respectively.

The y-axis displays the Phred quality scores. The better the base call the greater the score. The graphical background divides the y axis into calls of extremely high quality (green), calls of reasonable quality (orange), and calls of bad quality (red). Most platforms deteriorate call quality as the run proceeds, so it is normal to see base calls fall into the orange area near the end of a read [26].

In the dataset, 50%–60% of the sequences were of acceptable quality, while the remainder were of poor quality. The overall percentage of GC of all bases in all sequences was in between 39 to 44. Along with per base sequence quality, FastQC also gives series of analysis modules. These modules were analyzed for each sequence respectively (Figs. 5, 6).

Data cleaning
Reads which have poor quality are required to clean. Trimmomatic cleans the data by trimming and performing quality control operations. Poor quality reads can lead to false positive results and can generate errors while running the pipeline. As it can be seen in below images, Trimmomatic improved base quality score.

After trimming low-quality sequences with the Trimmomatic Tool, the sequences quality was verified again with FastQC. The increased FRED quality score generates more precise findings (Fig. 7).

Alignment
For each single end read file, HISAT2 output generated one bam file. These files are BAM files (short for Binary Alignment Map) and like the name suggests, is a binary file. Galaxy automatically converts these to a plain-text equivalent (SAM) file to view when
Identification of SNPs

The BAM file datasets were created for the variant calling. With the correct human genome (Hg 38) reference genome and run-in batch mode with merged output VCFs created the VCFs files for the datasets respectively. Examined the number of lines in the datasets, which is listed in the green box in the History for variants that are exactly in list [30].

SNP analysis

The default input and output formats for SnpEff are VCF. Since VCF is a common format that may be utilised by other tools and software programmes, it is strongly advised to use it as both an input and output format. Thus, integrating pipelines for analyzing genomic data is much facilitated by VCF.

The identified variations were annotated with information from well-known reference source i.e., Human genome (Hg38). Annotation data is added by SnpEff to the VCF file’s INFO field. The eighth column of a VCF file is the INFO field. A summary of the annotations made to the observed variants is included in an HTML document that SnpEff also created. The summaries of the various variant types, their effects, functional zones, etc., were analyzed. Examined tables and graphs and comprehended their meaning [31,32]. The 5 VCFs files were downloaded and viewed in IGV (Fig. 8).
Genes with SNPs related to COVID-19

The potential drug targets are listed with the help of IGV and were found significant based on the statistical analysis are listed in Table 1. We viewed listed genes related to COVID-19 from the literature and checked for variants in IGV. There are several SNPs found in chromosome 2 encodes DNAH7 (dynein axonemal heavy chain 7) gene. The DNAH7 gene, which codes for dynein axonemal heavy chain 7, a part of the inner dynein arm of ciliary axonemes. After SARS-CoV-2 infection of human bronchial epithelial cells, the gene DNAH7 is reported to be the most downregulated gene, suggesting a potential role in respiratory function. Reduced respiratory cilia performance may result from DNAH7 downregulation. COVID-19 patients who have mutations in DNAH7 may be more likely to pass away from the disease. The variant is composed of the single SNP, which is found in the promotor region of the gene CLUAP1 (clusterin associated protein 1)’s 5’ untranslated region. CLUAP1 at chr16_4 also has cilia-related functions. CLUAP1, which is produced by the gene CLUAP1, is an evolutionarily conserved protein necessary for ciliogenesis [16]. The severity of COVID-19 was linked to a genetic variation of the interferon-induced transmembrane protein 3 (IFITM3) gene, specifically the single-nucleotide polymorphism rs12252. It has been discovered as a potential factor in the susceptibility and severity of respiratory viral infections, including COVID-19, and it may play a role in the pathogenesis of COVID-19 ARDS. Asian groups frequently carry this genetic variation, and influenza severity has been linked to homozygosity for the C allele [33]. There are other two significant SNPs associated with genes related to COVID-19: rs35258888 for the PPA2 (pyrophosphatase 2) gene and rs70947091 for the PAPSSI (3’-phosphoadenosine 5’-phosphosulfate synthase 1) gene.
on the other hand, has been linked to an elevated risk of ARDS. A strong correlation between the PPA2 gene and sudden cardiac death (SCD) has been discovered. Some COVID-19 patients, according to recent reports, had SCD, which ultimately caused their tragic deaths. Reverse transcriptase–polymerase chain reaction analysis of nasopharyngeal swabs and radiological examinations

Fig. 7. Per base sequence quality before and after trimmomatic.
**Table 1. Variants in identified genes related to COVID-19**

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Gene name</th>
<th>Description</th>
<th>Variant Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DNAH7</td>
<td>DNAH7 is involved in cilia dysfunction. Primary ciliary dyskinesia is usually an autosomal recessive genetic condition in which the microscopic organelles (cilia) in the respiratory system have defective function. Ciliary dysfunction prevents the clearance of mucous from the lungs.</td>
<td>Chr 2: 195762733 SNP; A:T</td>
</tr>
<tr>
<td>2</td>
<td>CLUAP1</td>
<td>Involved in cilia dysfunction</td>
<td>Chr 16: 3503463 SNP; A:C, 3509507-3509510 MNP: AGGG*:TGGA</td>
</tr>
<tr>
<td>3</td>
<td>PPA2</td>
<td>Causes myocardial damage in most COVID patients</td>
<td>Chr 4: 105381563 SNP; T*:C</td>
</tr>
<tr>
<td>4</td>
<td>PAPSS1</td>
<td>Increases risk of ARDS</td>
<td>Chr 4: 107643163 SNP; G*:A</td>
</tr>
<tr>
<td>5</td>
<td>TLR4</td>
<td>Innate immune receptor on the cell surface that recognizes pathogen-associated molecular patterns including viral proteins</td>
<td>Chr 9: 117716883 SNP; T*:A</td>
</tr>
<tr>
<td>6</td>
<td>IFITM3</td>
<td>IFITM3 plays a role in adaptive and innate immune response.</td>
<td>Chr 11: 320115 SNP; G*:A</td>
</tr>
</tbody>
</table>

COVID-19, coronavirus disease 2019; SNP, single nucleotide polymorphism; MNP, multi-nucleotide polymorphism; ARDS, acute respiratory distress syndrome; DNAH7, dynein axonemal heavy chain 7; CLUAP1, clusterin associated protein 1; PPA2, pyrophosphatase 2; PAPSS1, 3'-phosphoadenosine 5'-phosphosulfate synthase 1; TLR4, toll-like receptor 4; IFITM3, interferon-induced transmembrane protein 3.
were used in research to identify three COVID-19 patients in July 2020. Ultimately, SCD took these victims' lives. A fair association between SCD and COVID-19 is suggested by the examination of the most recent data, notwithstanding the lack of a direct causal link between the two [34]. TLR4 is a part of a toll-like receptor family, which is associated with the pattern recognition family. TLR4 is a membrane protein, which identifies pathogenic proteins and produces interferons and proinflammatory cytokines in order to fight the infection. In the study done by [35], it was proposed that binding of TLR4 with the viral spike protein led to expression of angiotensin converting enzyme 2 resulting in decrease of air-tissue surface tension and blocking TLR4 thus promoting ARDS. Although the recent studies mention to have a very small or negligible role of TLR4 SNPs resulting in ARDS [36].

Statistical analysis
In hypothesis testing, a p-value less than 0.05 is conventionally considered statistically significant, leading to the rejection of the null hypothesis. Therefore, if a gene exhibits a p-value less than 0.05, we reject the null hypothesis, indicating evidence to suggest that the gene is associated with the progression of ARDS.

SNPs are variations in the DNA sequence that occur when a single nucleotide (adenine, cytosine, guanine, or thymine) in the genome is changed. Hence, we were able to analyze disease-causing genes in SARS COVID-19 through our research into SNPs, which can lead to ARDS in some situations. This research will aid in the future understanding of medication response diversity among individuals, which has important therapeutic consequences. By demonstrating a link between an individual's genetic make-up and pharmacological response, it may be possible to develop a genome-based diet and drugs that are more effective and safer for each individual. SNPs can also be used to learn about sequence evolution and its molecular mechanisms. The rate, nature, and position of nucleotide changes, as well as the selection pressure on codons, vary across a gene. Understanding these processes can provide insights into the evolution of disease-causing genes and may inform the development of new treatments. Overall, the study of SNPs and their role in disease susceptibility and drug response has the potential to significantly advance our understanding of the genetic basis of health and disease, and to improve healthcare outcomes for individuals.

ORCID
Shama Mujawar: https://orcid.org/0000-0003-3828-4662
Gayatri Patil: https://orcid.org/0000-0003-0332-9816
Srushhti Suthar: https://orcid.org/0000-0002-7793-2653
Tanuja Shendkar: https://orcid.org/0000-0002-6425-075X
Vaishnavi Gangadhar: https://orcid.org/0000-0002-7138-0480

Authors’ Contribution
Conceptualization: TS, VG. Data curation: GP, SS. Formal analysis: SM. Methodology: SM, GP, SS, TS, VG. Writing – original draft: GP, SS, TS, VG. Writing – review & editing: SM.
Conflicts of Interest

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

Acknowledgments

The authors would like to thank the Director and Head of School, MIT School of Bioengineering Sciences & Research for giving an opportunity to work independently and all there support with the best lab facilities.

Supplementary Materials

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

References


Introduction

Coronaviruses belong to Coronaviridae family which causes severe respiratory diseases in human beings. These viruses were initially isolated from a human in the year 1965 and are positive-sense, single-stranded RNA viruses. Until now, there are three major outbreaks caused due to coronavirus, namely, severe acute respiratory syndrome coronavirus (SARS-CoV) in the year 2002, Middle East respiratory syndrome coronavirus (MERS-CoV) in the year 2012, and recently, SARS-CoV-2 in the year 2019 [1]. The recent pandemic of coronavirus disease 2019 (COVID-19) that emerged in December 2019 has affected the whole nation and has put restrictions on the movement of people. It has been cautioned to every individual globally to remain at home and not to assemble, wash hands regularly, use face mask and hand sanitizer and keep physical distance [2]. SARS-CoV-2 which triggers the coronavirus disease is considered to have originated in the laboratories of Wuhan, China. It is the seventh human coronavirus that can cause severe health issues, majorly infecting the lungs that can eventually cause death. The frequently reported symptoms of...
COVID include shivering with high body temperature, cough, and difficulty breathing [3].

Currently, there are no relevant medications available for effective treatment of coronavirus. Drug repurposing has shown positive effects but they are incapable of preventing the disease. To treat coronavirus, it is very important to identify the therapeutic target or a drug that efficiently arrests the replication of SARS-CoV-2 [4]. One of the methods can be finding an antiviral drug or a molecule that can interfere with the spike protein or CD147 expression [5].

There are mainly two receptors that could be pathways for SARS-CoV-2 to get inside the host cell, namely angiotensin-converting enzyme 2 (ACE2) and CD147. The spike protein present externally on coronavirus binds with either ACE2 or CD147 receptor, thus intervening capture and spreading of the virus to another cell [5]. Recently, Wang et al. [6] illustrated that the expression level of ACE2 is low and there may be other routes for viral entry which later they found to be CD147. They came across the activity of the spike protein of SARS-CoV-2 and CD147 receptor-mediated by meplazumab which is an anti-CD147 humanized antibody. The use of this antibody blocks the host cell receptor CD147 which in turn inhibits the amplification of SARS-CoV-2. This particular work revealed an important target that can be utilized in developing a therapeutic drug against COVID-19 [7]. Furthermore, the study conducted by Geng and a co-worker revealed CD147 as a universal receptor for SARS-CoV-2 as well as its other variants [8].

CD147 or basigin or extracellular matrix metalloproteinase inducer is an integral membrane glycoprotein that comprises mainly two immunoglobulin units present in the extracellular region. The two domains of immunoglobulin are a single transmembrane domain and a cytoplasmic domain [9]. Apart from facilitating SARS-CoV infection, it also facilitates tumor growth, Plasmodium capture, and infection of viruses and bacteria [7]. CD147 is thought to regulate cytokine production and leucocyte chemotaxis when it binds to cyclophilin A (CyPA) [10]. Cyclophilins belong to the PPlase family of intracellular proteins which is highly conserved. It is the most commonly expressed protein of this family consisting of seven types of which CyPA is the most abundant and dominant protein [4].

This work aims to design and identify inhibitors of human basigin to stop viral entry into the cell by developing an e-Pharmacophore model of receptor-ligand interaction between CD147 and various U.S. Food and Drug Administration (FDA)–approved drugs for COVID-19. Pharmacophore modeling is a part of computer-aided drug design (CADD) that plays a key role when there is no structure of receptor presents [11]. Pharmacophores may be described as the arrangement of a molecule that stores the essential characteristics necessary for the biological or pharmacological interaction of drugs. It is used for screening the molecules virtually that triggers a biological response. A pharmacophore model usually has two procedures, ligand-related and structure-related [12]. In ligand-related pharmacophores modeling, a set of active molecules are superposed in order to get the chemical activities that are necessary for biological activities [13]. However, in structure-related method a pharmacophore model is developed using the structural features of the protein. In this model, the interaction between ligands and macromolecular targets is taken into account [14].

The technique of molecular docking has been widely applied for many decades in the field of drug discovery and many other aspects of bioinformatics. The method takes into account the orientation of a molecule and how the molecule changes its conformation after binding to a particular molecular target [15]. It is a structure-based drug design where binding occurs between a receptor and a ligand. There are various software developed to date, among which are some well-known, AutoDock, AutoDock Vina, ZDOCK, RDOCK, Glide, and Gold [16]. The first step in molecular docking involves generating a three-dimensional structure of a ligand and receptor from various databases available like Protein Data Bank (PDB) or can be drawn using tools e.g., ChemSketch, and preparing it. The next step involves defining and calculating of grid where the ligand can be docked. And finally, the docking process can be carried out and the results can be visualized using a visualization software [17].

Methods

In this work, with the help of the Biovia Discovery Studio platform, an e-pharmacophore model was developed using the protein Human Basigin (PDB ID 7DCE) in a structure-based manner. The developed model was used for screening the drug targets. A total of eleven drugs (Fig. 1), drawn online using Chem-Space were used for screening. These drugs were selected based on the literature studies [18,19] that have shown activity in COVID-19 therapy. All the drugs act differently and have several mechanisms against coronavirus. Various in vitro studies showed the drugs to be helpful in blocking the viral production of this deadly virus. Therefore, these drugs have been utilized for repurposing studies and to understand their ability to block the CD147 receptor to treat COVID-19. Further, the reasons for selecting these eleven drugs are described below.
<table>
<thead>
<tr>
<th>Drug</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Azithromycin</td>
<td><img src="image" alt="Structure" /></td>
</tr>
<tr>
<td>2. Hydroxychloroquine</td>
<td><img src="image" alt="Structure" /></td>
</tr>
<tr>
<td>3. Chloroquine</td>
<td><img src="image" alt="Structure" /></td>
</tr>
<tr>
<td>4. Lopinavir</td>
<td><img src="image" alt="Structure" /></td>
</tr>
<tr>
<td>5. Ritonavir</td>
<td><img src="image" alt="Structure" /></td>
</tr>
<tr>
<td>6. Nafamostat</td>
<td><img src="image" alt="Structure" /></td>
</tr>
<tr>
<td>7. Camostat</td>
<td><img src="image" alt="Structure" /></td>
</tr>
<tr>
<td>8. Famotidine</td>
<td><img src="image" alt="Structure" /></td>
</tr>
<tr>
<td>9. Umifenovir</td>
<td><img src="image" alt="Structure" /></td>
</tr>
<tr>
<td>10. Nitazoxanide</td>
<td><img src="image" alt="Structure" /></td>
</tr>
<tr>
<td>11. Fluvoxamine</td>
<td><img src="image" alt="Structure" /></td>
</tr>
</tbody>
</table>

**Fig. 1.** Structure of drugs.
**Drugs**

**Azithromycin**

Azithromycin is an FDA-approved drug which is mainly used to treat respiratory diseases like pneumonia. A person suffering from COVID-19 shows the symptoms of pneumonia and acute respiratory problems where azithromycin presents a promising drug candidate [20]. This drug has antibacterial properties along with the properties of antiviral and immunomodulatory that could be an interest in the treatment of COVID-19. It is a bacteriostatic antibiotic showing in vitro activity against SARS-CoV-2 and acts on the viral cycle at different points [21].

**Hydroxychloroquine**

This medication has been used for a long period of time for the treatment of malaria. Apart from its antimalarial effects, it also shows antiviral and anti-inflammatory effects. Since hydroxychloroquine shows its effectiveness across a broad range of diseases, it is also thought to be useful in viral infection [22]. It also has certain metabolic benefits which include lipid profiles and lower blood glucose levels. The pharmacokinetics of this drug vary among different individuals and its oral bioavailability is about seventy-five percent [23].

**Chloroquine**

Recently, the benefits of chloroquine have been observed in treating the symptoms of COVID-19. Alike, hydroxychloroquine this drug also has antimalarial and antiviral effects [24]. Chloroquine phosphate and sulfate, both have been made commercially available in the market as antimalarial drugs of which chloroquine phosphate proved to be beneficial in treating COVID-19-associated pneumonia [25].

**Lopinavir**

Lopinavir belongs to the class of protease inhibitors which is an antiretroviral drug. It is usually used along with ritonavir for the treatment and therapy of human immunodeficiency virus (HIV) infections. This medication shows activity in the treatment of COVID-19 as it inhibits the protease activity of coronavirus. It also stops the replication cycle of MERS-CoV as it blocks the post-entrance steps [26].

**Ritonavir**

Ritonavir, along with lopinavir is used to treat HIV infections. This medication is marketed using the name Norvir and it behaves as a protease inhibitor. Ritonavir, when combined with lopinavir acts as an inhibitor of drug metabolism whereas lopinavir acts as a principal antiviral compound. Another compound named Atazanavir, when used in combination with ritonavir showed successful activity in hindering the SARS-CoV-2 replication cycle [27].

**Nafamostat**

This medication belongs to the group of serine protease inhibitors, majorly used for therapy of dispersed intraventricular coagulation and pancreatitis. Various publications and in vitro studies suggests that this medication is beneficial for curing pneumonia in individual suffering from coronavirus disease [28].

**Camostat**

Camostat mesylate is a drug available for oral consumption. It belongs to the group of serine protease inhibitors which potentially degrades transmembrane serine protease 2. This drug is hypothesized to be potentially used against the coronavirus. It mainly inhibits the fusion of virus and cell membrane which in turn inhibits the replication of the virus [29].

**Famotidine**

Famotidine is over a counter drug which is widely used to treat gastric acid. This drug is safe to use and does not react with any other drug. It is a histamine 2 receptor antagonist and can be presented as a candidate for the therapy of coronavirus. The study conducted by Janowitz and co-author on patients suffering from COVID-19 suggests that high doses of famotidine can prove to be beneficial for the treatment of coronavirus [30].

**Umifenovir**

Umifenovir (Arabidol) is an antiviral medication mainly used for treating the infection caused by influenza viruses as it targets the membrane fusion process of influenza virus. Arabidol therapy or Arabidol in combination with darunavir is thought to be a possible approaches for treating the deadly coronavirus [31]. This drug prevents the integration of the virus lipid cell with the plasma membrane which restricts the virus from entering the host cell. This drug is thought to inhibit coronavirus infection as it interferes with the release of SARS-CoV-2 from intracellular vesicles [32].

**Nitazoxanide**

Nitazoxanide is a drug that is used to treat infection caused by protozoan, for example, diarrhea which is caused by the protozoa Cryptosporidium and Giardia. This drug has also been efficient against broad range of viruses in vitro including rotavirus, norovirus, hepatitis B virus, and dengue virus. It has shown an acceptable in vitro activity against the coronavirus that makes it a probable
candidate for treating SARS-CoV-2 [33].

**Fluvoxamine**

Fluvoxamine is a selective serotonin reuptake inhibitor used for the treatment of the obsessive-compulsive disorder. It decreases cytokine production by activating the sigma-1 receptor that resides inside the endoplasmic reticulum of the cell. This drug proves to be a promising approach for the advanced treatment of COVID [34].

**Protein selection and preparation**

Fig. 2 shows the cryogenic electron microscopy (cryo-EM) structure of the human XKR8-Basigin complex bound to Fab fragment (PDB ID 7DCE) which was retrieved from PDB at a resolution of 3.80 Å. This protein is bound to a Fab fragment which will be utilized further to produce an e-pharmacophore model. The selected protein was firstly used for pharmacophore generation and later it was prepared for docking. Thereafter, it was subjected to energy minimization by applying CHARMm force field, and also the root-mean-square deviation (RMSD) value was calculated.

**Pharmacophore generation**

The first step in the generation of the pharmacophore model is removing the ligand present in the protein. After removing the ligand, an e-Pharmacophore model was developed in a structure-based manner utilizing the Biovia discovery studio platform. Later, this model was used to screen the above-mentioned eleven drug targets.

**Ligand preparation**

A total of seven drug targets (chosen after screening with the e-pharmacophore model) were presented to dock against the protein molecule as these are used for treating the symptoms of COVID-19. All the ligands were prepared to dock against the prepared protein.

**Protein ligand docking**

To begin with molecular docking, the first and foremost step was determining the site sphere of the protein to know our ligand binding site. Thereafter, the prepared protein and different ligands were docked using CDOCKER tool of Biovia Discovery Studio. CDOCKER (CHARMm-based DOCKER) tool utilizes CHARMm's position for producing highly accurate docking results as it provides full flexibility for the ligand which includes bonds, dihedrals, angles, thereby generating the various conformations of the docked receptor and ligand. Apart from advantageous ligand flexibility, CDOCKER tool also offers CHARMm family
force fields, flexible CHARMM engine, and efficient time for computation analysis [35]. The CDOCKER energy is taken into account to find the best conformations amongst various refined poses that were generated.

**Results**

**e-Pharmacophore model**

Fig. 3A and 3B show the e-pharmacophore model that was developed using the Biovia Discovery Studio platform. Nineteen non-bonded interactions were found between the ligand and the receptor and a total of 10 pharmacophores were generated.

**Protein preparation and minimization**

The active site sphere of the prepared protein was 101.44, 87.84, and 97.17 along with the radius being 15.33. After superimposing the original protein with prepared protein an acceptable RMSD value of 0.73 Å was obtained which confirmed that the prepared protein structure is similar to that of original protein. The protein minimization energy obtained was –30,328.81547 kcal/mol. The following values of different energies were generated, bond energy 651.327, angle energy 3,287.27, dihedral energy 2,970.2, Vanderwaal energy –6,016.41, electrostatic energy –54,180.5, and the value of CHARMM was obtained as –52,832.1.

**CDOCKER results**

All eleven drug targets were screened and only seven of them fitted the pharmacophore.

*Hydroxychloroquine*

After docking the prepared protein with hydroxychloroquine, 40 refined poses were generated in total of which three were selected based on docking score shown in Table 1. Fig. 4 depicts the best interaction that had the highest CDOCKER energy and CDOCKER interaction energy of –20.7452 and –35.2816 respectively.

*Lopinavir*

A total of 10 refined poses were generated after docking with lopinavir. The top three docking values are shown in Table 2 and the best fit is depicted in Fig. 5 as it displayed the highest CDOCKER energy and CDOCKER interaction energy of – 45.0255 and –57.952 respectively.

*Ritonavir*

In case of ritonavir, 20 refined poses were generated and the top three ranked are depicted in Table 3. The highest CDOCKER energy and CDOCKER interaction energy generated was –57.3072 and –53.3876 respectively shown in Fig. 6.

*Camostat*

Docking with camostat resulted in the generation of 20 refined poses and the top three are depicted in Table 4. The best fit, shown in Fig. 7 had the highest CDOCKER energy and CDOCKER interaction energy of –20.3898 and –35.6218 respectively.

---

Fig. 3. (A) Generated e-Pharmacophore model showing the pockets where ligands can be fitted. (B) Generated e-Pharmacophore model without protein.
Table 1. -CDOCKER energy and -CDOCKER interaction energy of hydroxychloroquine-CD147 complex

<table>
<thead>
<tr>
<th>-CDOCKER_ENERGY</th>
<th>-CDOCKER_INTERACTION_ENERGY</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.7451</td>
<td>35.2816</td>
</tr>
<tr>
<td>19.9686</td>
<td>35.973</td>
</tr>
<tr>
<td>16.7187</td>
<td>30.459</td>
</tr>
</tbody>
</table>

Table 2. -CDOCKER energy and -CDOCKER interaction energy of lopinavir-CD147 complex

<table>
<thead>
<tr>
<th>-CDOCKER_ENERGY</th>
<th>-CDOCKER_INTERACTION_ENERGY</th>
</tr>
</thead>
<tbody>
<tr>
<td>45.0255</td>
<td>57.952</td>
</tr>
<tr>
<td>42.0827</td>
<td>44.1381</td>
</tr>
<tr>
<td>42.0644</td>
<td>44.4358</td>
</tr>
</tbody>
</table>

Fig. 4. CD147-hydroxychloroquine interaction.

Fig. 5. CD147-lopinavir interaction.
Table 3. -CDOCKER energy and -CDOCKER interaction energy of ritonavir-CD147 complex

<table>
<thead>
<tr>
<th>-CDOCKER_ENERGY</th>
<th>-CDOCKER_INTERACTION_ENERGY</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.3072</td>
<td>53.3876</td>
</tr>
<tr>
<td>54.152</td>
<td>49.1827</td>
</tr>
<tr>
<td>53.9664</td>
<td>47.3915</td>
</tr>
</tbody>
</table>

Table 4. -CDOCKER energy and -CDOCKER interaction energy of camostat-CD147 complex

<table>
<thead>
<tr>
<th>-CDOCKER_ENERGY</th>
<th>-CDOCKER_INTERACTION_ENERGY</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.3898</td>
<td>35.6218</td>
</tr>
<tr>
<td>20.2663</td>
<td>35.175</td>
</tr>
<tr>
<td>20.0669</td>
<td>36.949</td>
</tr>
</tbody>
</table>

**Fig. 6.** CD147-ritonavir interaction.

**Famotidine**

Famotidine generated the highest refined poses amongst all which is 110 poses. The top three docking values is given in Table 5 and Fig. 8 displays the best fit of the highest CDOCKER energy and CDOCKER interaction energy which is –26.8616 and –28.8008 respectively.

**Umifenovir**

Upon docking the prepared protein with umifenovir, a total of 20 refined poses were generated amongst which the top three docking values are shown in Table 6. Fig. 9 shows the best-fit pose having the highest CDOCKER energy and CDOCKER interaction energy of –25.1386 and –35.2122, respectively.

**Fluvoxamine**

Fluvoxamine also resulted in 110 refined poses, as in case of famotidine. The top three docking values are shown in Table 7 and the best fit is depicted in Fig. 10 as it displayed the highest CDOCKER energy and CDOCKER interaction energy of –29.8226 and –39.2161, respectively.

**Discussion**

The cryo-EM structure of human XKR8-Basigin complex bound to Fab fragment (PDB ID 7DCE) was acquired from Protein Data Bank at a resolution of 3.80 Å. An e-pharmacophore model was developed using this protein (7DCE) to generate pockets where drug targets can fit accordingly. Eleven drugs, namely, azithromycin, hydroxychloroquine, chloroquine, lopinavir, ritonavir, nafamostat, camostat, famotidine, umifenovir, nitazoxanide, and fluvoxamine were considered for this work as these drugs are currently being utilized for the treating the symptoms of coronavirus. These drugs after preparation were screened in the pharmacophore model amongst which seven of them, viz, hydroxychloroquine, lopinavir, ritonavir, nafamostat, camostat, famotidine, umifenovir, and fluvoxamine fitted to the pharmacophore pockets. The rest of the drugs...
**Fig. 7.** CD147-camostat interaction.

**Fig. 8.** CD147-famotidine interaction.

**Table 5.** -CDOCKER energy and -CDOCKER interaction energy of famotidine-CD147 complex

<table>
<thead>
<tr>
<th>-CDOCKER_ENERGY</th>
<th>-CDOCKER_INTERACTION_ENERGY</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.8616</td>
<td>28.8008</td>
</tr>
<tr>
<td>26.5134</td>
<td>29.4275</td>
</tr>
<tr>
<td>26.3377</td>
<td>28.1334</td>
</tr>
</tbody>
</table>

**Table 6.** -CDOCKER energy and -CDOCKER interaction energy of umifenovir-CD147 complex

<table>
<thead>
<tr>
<th>-CDOCKER_ENERGY</th>
<th>-CDOCKER_INTERACTION_ENERGY</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.1386</td>
<td>33.2122</td>
</tr>
<tr>
<td>25.0374</td>
<td>33.6277</td>
</tr>
<tr>
<td>24.906</td>
<td>33.8427</td>
</tr>
</tbody>
</table>
Fig. 9. CD147-umifenovir interaction.

Fig. 10. CD147-fluvoxamine interaction.
i.e., azithromycin, chloroquine, nafamostat, and nitazoxanide could not screen as these drugs do not interfere with the CD147 receptor. Later, the protein was prepared and energy minimization was performed to determine a proper molecular arrangement. After defining the site sphere molecular docking was performed between the receptor (human basigin) and the screened drug targets using CDOCKER. According to the CDOCKER energy generated, ritonavir emerged as the best drug demonstrating a higher CDOCKER energy of –57.30 and its corresponding CDOCKER interaction energy of –53.38, respectively. In summary, our results contribute to the findings of relevant drugs for COVID-19 and can be of great importance for other scientific works in the future.

ORCID

Nisha Kumari Pandit: https://orcid.org/0000-0003-2899-939X
Simranjeet Singh Mann: https://orcid.org/0000-0002-7756-8601
Anee Mohanty: https://orcid.org/0000-0002-8300-5549
Sumer Singh Meena: https://orcid.org/0000-0002-3932-2643

Authors’ Contribution

Conceptualization: NKP, AM, SSM (Sumer Singh Meena). Data curation: NKP, SSM (Simranjeet Singh Mann). Formal analysis: NKP. Methodology: NKP, AM, SSM (Sumer Singh Meena). Writing – original draft: NKP. Writing – review & editing: NKP, AM, SSM (Sumer Singh Meena).

Conflicts of Interest

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

References

17. Santos LH, Ferreira RS, Caffarena ER. Integrating molecular

Table 7. -CDOCKER energy and -CDOCKER interaction energy of fluvoxamine-CD147 complex

<table>
<thead>
<tr>
<th>-CDOCKER_ENERGY</th>
<th>-CDOCKER_INTERACTION_ENERGY</th>
</tr>
</thead>
<tbody>
<tr>
<td>29.8226</td>
<td>39.2161</td>
</tr>
<tr>
<td>29.3781</td>
<td>34.895</td>
</tr>
<tr>
<td>28.8813</td>
<td>35.1303</td>
</tr>
</tbody>
</table>

https://doi.org/10.5808/gi.23005


Single-cell RNA sequencing identifies distinct transcriptomic signatures between PMA/ionomycin- and αCD3/αCD28-activated primary human T cells

Jung Ho Lee1, Brian H Lee1, Soyoung Jeong1, Christine Suh-Yun Joh1, Hyo Jeong Nam1, Hyun Seung Choi1, Henry Sserwadda1, Ji Won Oh2, Chung-Gyu Park1,3,4, Seon-Pil Jin5,6,7*, Hyun Je Kim1,8,9*

1Department of Biomedical Sciences, Seoul National University Graduate School, Seoul 03080, Korea
2Department of Anatomy, Yonsei University College of Medicine, Seoul 03722, Korea
3Department of Microbiology and Immunology, Seoul National University College of Medicine, Seoul 03080, Korea
4Transplantation Research Institute, Seoul National University College of Medicine, Seoul 03080, Korea
5Department of Dermatology, Seoul National University Hospital, Seoul 03080, Korea
6Department of Dermatology, Seoul National University College of Medicine, Seoul 03080, Korea
7Medical Research Center, Institute of Human-Environmental Interface Biology, Seoul National University College of Medicine, Seoul 03080, Korea
8Genomic Medicine Institute, Seoul National University College of Medicine, Seoul 03080, Korea
9Seoul National University Hospital, Seoul 03080, Korea

Immunologists have activated T cells in vitro using various stimulation methods, including phorbol myristate acetate (PMA)/ionomycin and αCD3/αCD28 agonistic antibodies. PMA stimulates protein kinase C, activating nuclear factor-κB, and ionomycin increases intracellular calcium levels, resulting in activation of nuclear factor of activated T cell. In contrast, αCD3/αCD28 agonistic antibodies activate T cells through ZAP-70, which phosphorylates linker for activation of T cell and SH2-domain-containing leukocyte protein of 76 kD. However, despite the use of these two different in vitro T cell activation methods for decades, the differential effects of chemical-based and antibody-based activation of primary human T cells have not yet been comprehensively described. Using single-cell RNA sequencing (scRNA-seq) technologies to analyze gene expression unbiasedly at the single-cell level, we compared the transcriptomic profiles of the non-physiological and physiological activation methods on human peripheral blood mononuclear cell–derived T cells from four independent donors. Remarkable transcriptomic differences in the expression of cytokines and their respective receptors were identified. We also identified activated CD4 T cell subsets (CD55+) enriched specifically by PMA/ionomycin activation. We believe this activated human T cell transcriptome atlas derived from two different activation methods will enhance our understanding, highlight the optimal use of these two in vitro T cell activation assays, and be applied as a reference standard when analyzing activated specific disease-originated T cells through scRNA-seq.

Keywords: scRNA-seq, T cell, T cell activation, transcriptome
Introduction

The human immune system has evolved to survive against a variety of infections. T cells, which are among the central players in the adaptive immune system, not only recognize and eradicate infected cells, but also interact with other immune cells through cytokines. Hence, the production of cytokines from activated T cells is an essential process for protecting the body, mediating inflammation, and regulating other types of immune cells. T cells need several simultaneous signals to be fully activated, including T cell receptor (TCR), costimulatory, and cytokine signals. To study the detailed activation steps and role of activated T cells in specific environments, immunologists have used two different methods to activate T cells: αCD3/αCD28 agonistic antibodies and phorbol 12-myristate 13-acetate (PMA)/ionomycin.

Anti-CD3 is an agonistic antibody that can physiologically stimulate the TCR, thereby activating ZAP-70, which is the initiator of T cell downstream signaling [1]. ZAP-70 delivers downstream signals through phosphorylation of its primary targets: linker for activation of T cell (LAT) and SH2-domain-containing leukocyte protein of 76 kD (SLP-76) [2,3]. After the activation of LAT and SLP-76, several signaling molecules are recruited, including phospholipase C-γ (PLC-γ) and AKT, which play a key role in cellular metabolism. In combination, anti-CD28 maximizes PLC-γ activation, generating two second messengers—diacylglycerol (DAG) and IP3, from PIP2—via the local production of PIP3 [4]. In contrast, PMA induces Ras and protein kinase C to activate nuclear factor-κB in a SOS- and CARMA1-dependent manner [5]. Ionomycin triggers a cytosolic influx of calcium ions, effectively simulating the role of PLC-γ. However, differences in the resulting T cell gene expression profiles between the two activation methods using single-cell transcriptomic techniques have not yet been explored.

In this study, we used single-cell RNA sequencing to compare resting, αCD3/αCD28 agonistic antibody-activated, and PMA/ionomycin-activated T cells isolated from peripheral blood mononuclear cells (PBMCs) from four healthy individuals. By analyzing a total of 35,736 resting, PMA/ionomycin-activated, and αCD3/αCD28 agonistic antibody-activated T cells, we aimed to establish a standard reference that would aid researchers in determining which activation method to apply and could be used to analyze specific disease-derived T cells after activation.

Methods

Isolation of PBMCs

The collection of healthy individuals’ blood was approved by the Seoul National University Hospital Institutional Review Board (SNUH IRB No. C-2205-189-1327). From each donor, 16 mL of whole blood was collected in heparin CPT tubes (cat No. 362753, BD, Franklin Lakes, NJ, USA). The collected blood was immediately processed for PBMC isolation. CPT tubes were centrifuged for 20 min at 1,800 ×g with minimum acceleration and deceleration, and the interphase was collected in complete medium (RPMI containing 10% fetal bovine serum [FBS]). After washing with complete medium twice, the PBMCs were frozen with Cell Banker I (cat No. 11888, AMSBIO, Abingdon, UK) and stored in liquid nitrogen until batch processing.

T cell enrichment and stimulation

T cells were enriched using the EasySep human T cell enrichment kit (cat No. 19051, STEMCELL Technologies, Vancouver, Canada). After T cell enrichment, the cells were suspended in complete medium (RPMI 1640, 50%, Gibco, Thermo Fisher Scientific, Waltham, MA, USA) containing NEAA, HEPES, L-GlutaMax, 2-mercaptoethanol, FBS, and antibiotics. Brefeldin A was added to all groups in order to lock the produced transcripts and proteins inside the cells. Interleukin (IL)-2 was added to all groups to ensure T cell survival during the incubation period. All donor samples were divided into three activation groups: group 1, resting control; group 2, 50 ng/mL PMA (cat No. P8139, 1 mg, Sigma Aldrich, St. Louis, MO, USA) and 1.34 µM ionomycin (cat No. 10634, 1 mg, Sigma) for chemical stimulation and group 3, Dynabeads human T-activator CD3/CD28 (cat No. 11131D, Gibco, Thermo Fisher Scientific) added at a 1:1 ratio (beads: cells) for physiological activation. Each group was incubated for 4 h at 37.5 °C in a 5% CO2 incubator. After 4 h of incubation, the cells were harvested and used in further experiments.

Purity and protein validation through flow cytometry

An aliquot of the T cell-enriched samples from each donor was stained with anti-human CD3 (cat No. 300426, clone UCHT1, BioLegend, San Diego, CA, USA) for 30 minutes at room temperature to validate the isolation of T cells. The aliquots were also stained with anti-human CD3 (cat No. 300426, clone UCHT1, BioLegend), anti-human CD8 (cat No. 560662, clone RPA-T8, BD Biosciences, Franklin Lakes, NJ, USA), and anti-human CD45RO (cat No. 562299, clone UCHL1, BD Biosciences). The cells were then treated with fixation/permeabilization concentrate
(cat No. 00-5123-43, Invitrogen, Waltham, MA, USA) to fix and permeabilize cells for 1 h at 4°C, and intracellular staining was conducted according to the manufacturer’s recommendations. The cells were stained with anti–human interferon γ (IFN-γ) (cat No. 56-7319-41, clone 4S.B3, Invitrogen), anti–human tumor necrosis factor α (TNF-α) (cat No. 25-7349-41, clone MAb11, Invitrogen), and anti–human IL-2 (cat No. 500307, clone MQ1-17H12, BioLegend). Sample data were acquired using the LSR Fortessa X20 and were analyzed using FlowJo software.

**Cell multiplexing**

For each experimental condition, samples from four donors were multiplexed. Immediately following T cell activation, all 12 samples (4 donors × 3 conditions each) were multiplexed in Cell Multiplexing Oligo (CMO) for 5 min at room temperature using the 3’ Cellplex kit set A (PN 1000261) according to the manufacturer’s recommendation to reduce batch effects. The labeled cells were washed several times to avoid multi-labeling after pooling of cells. The labeled cells were washed by centrifugation at 400 ×g at 4°C following the 10× Genomics protocol with RPMI containing 10% FBS.

**Single-cell RNA library construction and sequencing**

Gene expression and CMO libraries were constructed following the 10× Genomics guidelines. The 4150 TapeStation system (Agilent, Santa Clara, CA, USA) was used for quality control of the cDNA and cell multiplexing libraries. Sequencing was done by NovaSeq 6000 (Illumina, San Diego, CA, USA) at a depth of 50,000 and 10,000 reads per cell for gene expression and CMO libraries, respectively.

**Data analysis**

Demultiplexing and alignment to the human genome were performed using the Cell Ranger software (v6.1.2). The Seurat package (v4.2) was used for pipeline analysis of the aligned dataset. All data were integrated using Harmony (v1.0) to minimize batch effects. Doublets were excluded using Doublet Finder (v2), and cells with over 3,300 or fewer than 200 features of transcripts were excluded. Furthermore, cells showing a percentage of mitochondrial gene expression exceeding 7.5% were excluded. Non-T cells were excluded based on the absence of CD3E, CD3D, CD247, and CD3G expression. Normalization was applied using the log normalization method. We also applied the Louvain algorithm for clustering, and differentially expressed genes were identified using the Wilcoxon rank-sum test. The Monocle3 (v1.3.1) package was used to order the cells according to the pseudotime via trajectory analysis with 0.9 resolution, 10 principal components, and naïve T cells selected as root nodes.

**Data available**

The primary human T cell scRNA-seq data used in this study are available with links to BioProject accession number PRJNA948720 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/).

**Results**

**scRNA-seq data generation from multiplexed human activated T cells**

The experimental design is summarized in Fig. 1A. The results of sorting CD3+ T cells from PBMCs were validated by flow cytometry with an obtained sorting purity of over 85% T cells from all donors (Fig. 1B). Following activation in each experimental condition (control, PMA/ionomycin, and αCD3/αCD28), we obtained a total of 35,970 cells (8,432 cells in the resting control group, 14,587 cells in the PMA/ionomycin group, and 12,951 cells in the αCD3/αCD28 agonistic antibody group) (Fig. 1C).

**Frequency of annotated cell type across donors and experimental conditions**

T cells across all experimental conditions and donors were integrated using the top 1,000 highly variable genes (Fig. 2A). Unbiased clustering was performed using the Louvain clustering method, and T cell clusters were annotated using cell type-and activation type-specific gene markers (Fig. 2B). The frequencies of annotated cell types by each experimental condition and donor are shown in Fig. 2C. No significant differences were found in cell type frequency across donors. Activated CD4 T cells and activated CD8 T cells were found at higher frequencies in the PMA/ionomycin-stimulated activation group than in the agonistic antibody-stimulated group (Fig. 2D).

**Cytokine and cytokine receptor expression on activated CD4 T cells**

To further interrogate the various CD4 T cell activation states induced by the different stimulation methods, we grouped the resting and activated CD4 T cells (total of 18,513 cells) and applied unsupervised sub-clustering (Fig. 3A). We found broad differences between the stimulation methods regarding the gene expression of cytokines and their respective receptors (Fig. 3B). Cytokine expression, including TGFβ1, IL2, CSF1, CSF2, XCL1, XCL2, CCL20, CCL4, CCL3, CXCL8, CXCL3, IFNG, TNFSF14,
**Fig. 1.** Workflow of the experiment. (A) Overview of the experimental design. (B) CD3⁺ sorting purity validation using flow cytometry. (C) Design of the scRNA-seq multiplexing and total recovered cell number. All figures were created with BioRender.com. IL-2, interleukin 2; PBMC, peripheral blood mononuclear cell; scRNA-seq, single-cell RNA sequencing.
Fig. 2. Annotation of total T cells and comparison of frequencies across experimental conditions. (A) UMAP of 35,736 T cells clustered by cell type and activation markers. Clusters are color-coded. (B) Dot plot showing expression levels of canonical markers in each cluster. The percent expression of CD4\(^+\) and CD8\(^+\) T cell genes is represented based on the dot size, and the average expression is represented by color. (C) The proportions of CD4\(^+\) and CD8\(^+\) T cell clusters were grouped by experimental conditions (top) and donors (bottom). (D) UMAP of CD4\(^+\) and CD8\(^+\) T cells divided by different experimental conditions.
Fig. 3. Expression of cytokines and cytokine receptors on CD4⁺ T cells. (A) UMAP of CD4⁺ T cells (18,513 cells) clustered by different experimental conditions. (B) Volcano plot showing differentially expressed genes between αCD3/αCD28 antibody-activated CD4⁺ T cells and phorbol myristate acetate (PMA)/ionomycin-activated CD4⁺ T cells. Dot plot showing the expression of cytokine genes (C) and cytokine receptor genes (D) in activated CD4⁺ T cells between the two groups of distinct activation methods.
TNFSF8, FASLG, CD40LG, and TNF, was higher in the PMA/ionomycin group than in the αCD3/αCD28 agonistic antibody-treated group in terms of both percentages of expression and average expression per cell. The percentages of expression of MIF and LTA were similar between the two activated groups. However, the expression levels and percentages of expression of IL32 and CCL5 were higher in the αCD3/αCD28 agonistic antibody-treated group (Fig. 3C). For cytokine receptor expression, the TNF superfamily, including TNFRSF1A, TNFRSF1B, TNFRSF4, FAS, CD27, TNFRSF9, TNFRSF18, and TNFRSF25, showed overall elevated expression in the αCD3/αCD28 agonistic antibody-treated group. Furthermore, receptors including CXCR4, CCR7, IL-2RG, IL4R, IL6ST, IL7R, and IL27RA showed higher expression in the αCD3/αCD28 agonistic antibody-treated group (Fig. 3D). Conversely, IL2RA, IL1R1, and CXCR3 were higher in the PMA/ionomycin group. Among these, IFN-γ, TNF-α, and IL-2 were validated at the protein level using flow cytometry, and the findings were found to be consistent with the transcriptomic levels (Supplementary Fig. 1A).

Cytokine and cytokine receptor expression on activated CD8 T cells
We applied the same sub-clustering strategy as described above for sub-clustering CD8 T cells (Fig. 4A). We grouped and then sub-clustered a total of 7,033 resting and activated CD8 T cells, again finding unique expression patterns in terms of cytokines and their respective receptor expression (Fig. 4B). We found similarities in cytokine expression between activated CD8 T cells and CD4 T cells, but the expression of TNF was distinctly reduced in activated CD8 T cells (Fig. 4C). In addition, the expression patterns of cytokine receptors were similar as those in CD4 T cells, but IL2RA showed lower levels of expression, and IFNAR1 and IFNAR2 were distinctive in CD8 T cells (Fig. 4D). IL6R expression was elevated in the cells that received PMA/ionomycin stimulation. The protein levels of IFN-γ, TNF-α, and IL-2 were also validated in CD8 T cells by flow cytometry and were consistent with the transcriptome levels (Supplementary Fig. 1B).

Enrichment of CD55+ activated CD4 T cells by PMA stimulation
We grouped all the activated CD4 T cells across all stimulated groups and sub-clustered them for further analysis (a total of 13,573 cells). We annotated the subsets based on differential gene marker expression and found Th1, Th2, Th17, central memory helper T cell, activated naïve helper T cell, SOX4 T cell and CD55+ T cells (Fig. 5B). We found that CD55+ T cells were enriched specifically by PMA/ionomycin activation (Fig. 5C). Moreover, the Monocle 3 package was used to order the cells according to pseudotime via trajectory analysis (Fig. 5D). Using Monocle 3 for pseudotime trajectory analysis, we found that CD55+ T cells and SOX4 T cells could potentially be differentiated from naïve T cells instead of central memory T cells.

Discussion
Naïve T cells require two different extracellular signals in order to become activated. Each T cell has its antigen specificity, and if this receptor interacts with the antigen presented by the MHC-antigen peptide complex on the antigen-presenting cell (APC), an initial signal is generated. Signaling is transmitted into the cells via CD3 [6,7]. A secondary signal is also required from costimulatory molecules on the APC and the corresponding ligand on the T cell surface [8]. Therefore, stimulating T cells with αCD3/αCD28 antibodies closely resembles the natural downstream signal of the TCR, resulting in a more physiological method for activating T cells. However, αCD3/αCD28 agonistic antibodies may indirectly activate other classes of lymphoid-lineage cells, such as T cells, B cells, and natural killer cells, upon activation [9]. In addition, the heterogeneity of CD3 expression among T cell subsets may lead to varying degrees of T cell activation [10,11], and these factors should be considered for precise data interpretation.

PMA, a DAG analog, exerts its effect on protein kinase C by crossing the cell membrane. Ionomycin increases intracellular calcium ions and activates calcineurin. The shared signals of protein kinase C and calcineurin promote the activation of T cells [12]. PMA/ionomycin is widely employed to stimulate immune cells since it is relatively inexpensive and simple to optimize the conditions [13]. However, when applying this method for T cell activation, it should be noted that PMA/ionomycin stimulates cells through non-specific mechanisms and can be toxic to cells due to overstimulation, leading to activation-induced cell death.

As predicted, the percentage of activated T cells was higher in the stimulated groups than in the resting group. We also found that the PMA/ionomycin-mediated activation group had a higher frequency of activated CD4 T cells and proportionally fewer resting CD4 T cells than the αCD3/αCD28 agonistic antibody group (Fig. 2C). Activation markers, including CD69, IL2RA, CD40LG, ICOS, CTLA4, and PDCD1, were present at elevated levels on PMA/ionomycin-treated T cells when compared to the αCD3/αCD28 agonistic antibody-treated T cells (Supplementary Fig. 2). Taken together, these results indicate that PMA/ionomycin has a greater T cell-stimulating potential than the αCD3/αCD28 ago-
Fig. 4. Expression of cytokines and cytokine receptors on CD8$^+$ T cells. (A) UMAP of CD8$^+$ T cells (7,033 cells) clustered by different experimental conditions. (B) Volcano plot showing differentially expressed genes between αCD3/αCD28 antibody-activated CD8$^+$ T cells and phorbol myristate acetate (PMA)/ionomycin-activated CD8$^+$ T cells. Dot plot showing the expression of cytokine genes (C) and cytokine receptor genes (D) in activated CD8$^+$ T cells between the two groups of distinct activation methods.
Fig. 5. Characterization of activated CD4+ T cells. (A) UMAP of activated CD4+ T cells (13,573 cells) clustered by different CD4+ subsets in different experimental conditions. (B) Dot plot showing the expression of cytokine genes in CD4+ T cell subsets between the two groups of distinct activation methods. (C) Bar graph showing frequency of each cluster of activated CD4+ T cells using the two distinct activation methods. (D) UMAP of trajectory analysis on activated CD4+ T cells reported by Monocle 3.
nistic antibodies under the same duration of incubation. However, not all genes related to T cell activation showed higher expression in the PMA/ionomycin-stimulated group in analyses of cytokines and their receptors.

Cytokines are critical to T cell function in terms of maturation, growth, and response to stimuli. Antigen stimulation induces T cell activation and the production of numerous cytokines in peripheral T cells (Figs. 3 and 4) [14]. We identified several cytokines among the differentially expressed genes in activated T cells according to the activation method (Figs. 3C and 4C). Generally, cytokine expression was higher in the PMA/ionomycin-treated group than in the αCD3/αCD28 agonistic antibody-treated group, including the TNF family. Since the members of the TNF family (TNFSF14, TNFSF8, FASLG, and TNF) are related to apoptosis, this result is aligned with a previous report stating that PMA/ionomycin stimulation results in activation-induced cell death via Fas/Fas ligand up-regulation [15]. A study using CD4 T cells from intestinal biopsies also showed that PMA/ionomycin induced larger amounts of IFN-γ and IL-17, while IL-10 production was predominant following αCD3/αCD28 stimulation [13].

Cytokines deliver signals upon binding to their receptors. A greater number of receptors on a cell indicates that it is more sensitive to a cytokine. Here, we found that, in general, the expression of cytokine receptors in the αCD3/αCD28 agonistic antibody-treated group was greater than that in the PMA/ionomycin-treated group. Only the genes for a few receptors, such as IL2RA, IL1R1, TNFRSF10B, and TNFRSF10A, were elevated in the PMA/ionomycin-treated group.

As for both cytokines and their receptors, activated CD8 T cells showed similar findings to activated CD4 T cells upon two different stimulation methods, suggesting a general stimulatory effect on T cell subsets (Fig. 4C and 4D).

When further investigating the activated CD4 T cell subsets, the proportions of Th1, Th2, and Th17 were not remarkably different despite the higher potential of PMA/ionomycin to induce cytokines (Fig. 5C). However, we found an increase in activated CD4 T cells expressing CD55 in response to PMA/ionomycin, and trajectory analysis showed that their potential source was naïve CD4 T cells (Fig. 5C and D). Further studies are needed to elucidate the role of this population.

The main limitation of this study is that we utilized a single fixed experimental condition for the two distinct stimulation methods. The results may vary depending on the concentration and length of incubation. Nonetheless, we anticipate that our research will aid in determining the most appropriate stimulation methods for T cell research based on the objectives of the study.

ORCID
Jung Ho Lee: https://orcid.org/0000-0002-0789-0259
Brian H Lee: https://orcid.org/0000-0002-1291-3208
Soyoung Jeong: https://orcid.org/0000-0003-0911-0567
Christine Suh-Yun Joh: https://orcid.org/0000-0003-0913-1202
Hyo Jeong Nam: https://orcid.org/0000-0003-4987-0747
Hyun Seung Choi: https://orcid.org/0000-0002-6217-4714
Henry Sserwadda: https://orcid.org/0000-0001-9647-139X
Ji Won Oh: https://orcid.org/0000-0001-5742-5120
Chung-Gyu Park: https://orcid.org/0000-0003-4083-8791
Seon-Pil Jin: https://orcid.org/0000-0002-8290-4278
Hyun Je Kim: https://orcid.org/0000-0003-4467-0949

Authors’ Contribution
Conceptualization: HJK. Data curation: JHL, HJN. Formal analysis: JHL, BHL. Funding acquisition: HJK, SPJ, CGP. Methodology: HJK, JHL, HJN. Writing – original draft: JHL, SPJ, BHL, SJ, CSYJ. Writing – review & editing: HJK, BHL, SPJ, JWO, SJ, CSYJ, HS, HSC.

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

Acknowledgments
This work was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science & ICT (2022M3A9D30316848), and by the Basic Science Research Program through the National Research Foundation (NRF) funded by the Ministry of Education (2020R1F1A1073692 and 2022R1F1A1075235).

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

References


Cell line–specific features of 3D chromatin organization in hepatocellular carcinoma

Yeonwoo Kim, Hyeokjun Yang, Daeyoup Lee*

Department of Biological Sciences, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea

Liver cancer, particularly hepatocellular carcinoma (HCC), poses a significant global threat to human lives. To advance the development of innovative diagnostic and treatment approaches, it is essential to examine the hidden features of HCC, particularly its 3D genome architecture, which is not well understood. In this study, we investigated the 3D genome organization of four HCC cell lines—Hep3B, Huh1, Huh7, and SNU449—using in situ Hi-C and assay for transposase-accessible chromatin sequencing. Our findings revealed that HCC cell lines had more long-range interactions, both intra- and interchromosomal, compared to human mammary epithelial cells (HMECs). Unexpectedly, HCC cell lines displayed cell line-specific compartmental modifications at the megabase (Mb) scale, which could potentially be leveraged in determining HCC subtypes. At the sub-Mb scale, we observed decreases in intra-TAD (topologically associated domain) interactions and chromatin loops in HCC cell lines compared to HMECs. Lastly, we discovered a correlation between gene expression and the 3D chromatin architecture of SLC8A1, which encodes a sodium-calcium antiporter whose modulation is known to induce apoptosis by comparison between HCC cell lines and HMECs. Our findings suggest that HCC cell lines have a distinct 3D genome organization that is different from those of normal and other cancer cells based on the analysis of compartments, TADs, and chromatin loops. Overall, we take this as evidence that genome organization plays a crucial role in cancer phenotype determination. Further exploration of epigenetics in HCC will help us to better understand specific gene regulation mechanisms and uncover novel targets for cancer treatment.

Keywords: chromatin loops, compartments, hepatocellular carcinoma, in situ Hi-C, TADs, 3D chromatin organization

Introduction

In 2022, the American Cancer Society estimated that approximately 800,000 people were diagnosed with liver cancer and 700,000 died from the disease worldwide. Liver cancer is the third leading cause of cancer death, partly reflecting that the lack of accurate liver cancer-specific diagnostic tools has limited early diagnosis [1,2]. There has recently been increasing interest in immunotherapy and targeted therapy [3], but the diagnosis and treatment of liver cancer remain challenging. Developing novel diagnostic tests and treatments is critical for improving the prognosis of patients. Therefore, we must better understand the genomic characteristics of liver cancer.

The most common type of primary liver cancer is hepatocellular carcinoma (HCC), which is a malignant tumor of hepatocytes. The causes of HCC are very diverse; they include chronic infection with hepatitis B virus or hepatitis C virus, alcohol abuse, and met-
This heterogeneity makes standardized chemotherapy ineffective, resulting in frequent metastases to nearby organs, poor prognoses, and high mortality rates [5].

Researchers have thoroughly scrutinized and categorized the transcriptomes of numerous HCC tissue samples [6-8], but we still know relatively little about the relevant three-dimensional (3D) interactions between distant gene loci, which have been shown to play a crucial role in regulating gene expression [9].

The mammalian genome maintains a highly organized and dynamic 3D form that arises from two kinds of interactions [10]. At the megabase (Mb) scale, the genome is partitioned into A and B compartments. Active chromatin, which possesses high-transcription histone markers and many genes, is called the A compartment, whereas inactive chromatin, which is gene-poor and possesses gene-silencing histone markers, is called the B compartment [11]. At the sub-Mb level, each compartment is divided into smaller, self-interacting domains called TADs (topologically associated domains), which are insulated from neighboring domains [12-14]. Although the precise function of TADs is not yet fully understood, it is believed that they contribute to regulating gene expression by helping ensure that cis-regulatory elements and their target promoters are brought into proximity with one another [15]. Therefore, from the perspective of epigenetics, studying 3D chromatin organization is essential to understanding the mechanisms of gene expression regulation.

To characterize the 3D chromatin landscape of HCC cell lines, we performed in situ Hi-C and assay for transposase-accessible chromatin sequencing (ATAC-seq) of four HCC cell lines and compared the results to those of a normal epithelial cell line. This comprehensive analysis of the chromatin interactions in HCC extends our knowledge of genome topology and epigenetics in hepatocellular carcinoma.

**Methods**

**Cell culture**

This research used four HCC cell lines: Hep3B, Huh1, Huh7, and SNU449 (Table 1). The HCC cell lines were cultured and harvested in the laboratory of Kyung Hyun Yoo at Sookmyung Women’s University. We used human mammary epithelial cells (HMECs) from normal female mammary tissue as non-cancer control samples (primary mammary epithelial cells; normal, human; ATCC PCS-600-010). HMECs were cultured in mammary epithelial cell basal medium (ATCC PCS-600-030) supplemented with components from a mammary epithelial cell growth kit (ATCC PCS-600-040). All cell lines were grown at 37°C in a humidified incubator with 5% CO₂.

**Cell harvest**

We harvested cells with 0.05% trypsin-EDTA, resuspended 5 million cells in 5 mL 1x phosphate buffered saline (PBS), added 274 μL of 36.5% formaldehyde (final concentration, 2%), mixed the solution thoroughly by inversion, and incubated the suspended cells at room temperature for 10 min. We added 400 μL of 2.5 M glycine and placed the sample on ice for 15 min. Cells were pelleted by centrifugation at 500 ×g for 5 min, the supernatant was discarded, and the cells were resuspended in 5 mL 1x PBS. The cells were equally distributed to five new tubes (1 million cells/tube) and pelleted by centrifugation, and the supernatant was discarded.

**In situ Hi-C and library sequencing**

We followed the Arima-HiC protocol (cat No. A160259, Arima Genomics, Inc., San Diego, CA, USA) to perform in situ Hi-C with 1 million harvested cells. We generated a library using an Arima-HiC kit (cat No. A510008, Arima Genomics, Inc.) and sequenced the library using the 150 bp paired-end method of the Illumina NovaSeq 6000 system (Illumina, San Diego, CA, USA).

**Assay for transposase-accessible chromatin sequencing**

Harvested nuclei of 50,000 cells were incubated in 25 μL fresh TD buffer (10 mM Tris-HCl, pH 8.0, 5 mM MgCl₂, and 10% dimethylformamide) with 2.5 μL Tn5 transposase for 30 min at 37°C. We purified DNA fragments with a QIAquick PCR purification kit (cat No. 28106, Qiagen, Hilden, Germany) and amplified the library using a KAPA HiFi HotStart ReadyMix (KK2061, Roche, Mannheim, Germany) as described in the provided manual, with adjustment of the PCR cycle number. The resulting libraries were purified with a QIAquick PCR purification kit. The purified libraries of HCC cell lines were sequenced via the 150 bp paired-end method of the Illumina NovaSeq 6000 system. The 150 bp paired-end method of the Illumina HiSeq 2500 system was used for HMECs.

**Table 1. Characteristics of cell lines**

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Host</th>
<th>Morphology</th>
<th>HBV integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hep3B</td>
<td>Black male, 8 y</td>
<td>Epithelial</td>
<td>Positive</td>
</tr>
<tr>
<td>Huh1</td>
<td>Japanese male, 53 y</td>
<td>Epithelial-like</td>
<td>Positive</td>
</tr>
<tr>
<td>Huh7</td>
<td>Japanese male, 57 y</td>
<td>Epithelial-like</td>
<td>Negative</td>
</tr>
<tr>
<td>SNU449</td>
<td>Korean male, 52 y</td>
<td>Epithelial</td>
<td>Positive</td>
</tr>
<tr>
<td>HMEC</td>
<td>Adult female breast tissue</td>
<td>Epithelial</td>
<td>Negative</td>
</tr>
</tbody>
</table>

HBV, hepatitis B virus.
Data processing and analysis

In situ Hi-C analysis

The in situ Hi-C sequencing data were analyzed using HiC-Pro [16]. Each raw file was aligned to the human genome (hg19) and filtered. Replicate data were merged. Contact matrices were built from the merged data, and iterative correction and eigenvector decomposition (ICE) normalization were applied. Various resolutions (10, 20, 40, 100, and 500 kb) of ICE-normalized Hi-C matrices were generated, and annotation files indicating genomic bins were developed. Contact probability calculation and principal component analysis (PCA) to define the compartment were done using Cooltools at a 100 kb resolution [17]. To define TAD boundaries, reciprocal insulation (RI) scores were calculated using CaTCH at 20 kb resolution [18]. For most of the Hi-C analyses, including relative contact probability analysis, TAD insulation scoring, aggregate TAD analysis, and aggregate peak analysis, GENOVA was used [19]. To identify chromatin loop interactions, the HiCCUPS algorithm of juicer tools was used with default parameters [20].

ATAC-seq analysis

For ATAC-seq analysis, the adaptors of raw reads were trimmed with Cutadapt [21], and the trimmed sequences were mapped to the human genome (hg19) via Bowtie2 (version 2.5.0) with default parameters [22]. The aligned bam files were merged and sorted with SAMtools [23]. The bam2wig tool of the RSeQC tool was used to generate bigwig files [24].

Total RNA-sequencing analysis

For the total RNA-sequencing (RNA-seq) analysis, raw reads were aligned to the human genome (hg19) using STAR (version 2.7.10) with default parameters [25]. Cufflinks (fr-firststrand) was used to analyze differential expression levels [26]. CummeRbund was used to create certain plots [27]. Additional plots, including box plots, were drawn with the R package, ggplot2 [28]. Heatmaps were drawn with Java Treeview [29], and the examples of genome-wide data were visualized using the Integrative Genomics Viewer (IGV) [30].

Public data acquisition

Publicly released total RNA-seq data were downloaded from NCBI Gene Expression Omnibus (GEO) datasets. These files were obtained in fastq format. The total RNA-seq data of Hep3B, Huh1, Huh7, and SNU449 cells (GSM2551564, GSM2551568, GSM2551570, and GSM2551589, respectively) were obtained from GSE97098 [31]. The total RNA-seq data of HMECs (GSM5667415) were obtained from GSE187119 [32].

Data availability

The Hi-C and ATAC-seq datasets have been deposited in the NCBI GEO; http://www.ncbi.nlm.nih.gov/geo/ under accession number GSE226217 (SuperSeries). This SuperSeries (GSE226217) is composed of two SubSeries: GSE226215 (ATAC-seq) and GSE226216 (Hi-C).

Results

Long-range interactions are increased in HCC cell lines

To investigate the 3D chromatin organization of HCC cell lines, we performed in situ Hi-C in four HCC cell lines and HMECs (Fig. 1), with two biological replicates for each cell line. We compared genome-wide Hi-C features between HCC cell lines and HMECs. We found significantly more long-range interactions, including intrachromosomal and interchromosomal interactions, in all four HCC cell lines relative to HMECs. Long-range interactions especially increased in Hep3B and Huh7, compared to Huh1 and SNU449.

First, the trans-interactions (i.e., interchromosomal interactions) were explored. Our results revealed that HCC cell lines had higher observed versus expected (obs/exp) trans-contact ratios (Fig. 1A) than HMECs. All trans-contact counts were much lower than expected in the case of HMECs, but not in HCC cell lines. The whole-contact maps, which compared the ratio of interactions throughout the genome between HCC cell lines and HMECs, also reflected this difference (Fig. 1B). Most trans-bins indicated higher contact frequencies in HCC cell lines. Along with these differences in the contact maps, we observed that all chromosomes of HCC cell lines had decreased cis-contact percentages compared to HMECs (Fig. 1C). In other words, in HCC cell lines, the number and ratio of trans-contacts in most chromosomes were consistently higher than those of HMECs. Previous studies reported that cancer genomes undergo various chromosomal rearrangements, including chromosome duplications, deletions, and translocations [33,34]. To prevent chromosomal rearrangement from affecting our result, we calculated copy number variations (CNVs) using HiCnv [35] (Supplementary Table 1, Supplementary Fig. 1). Consistent with our previous results, HCC cell lines had higher trans-contacts in contact maps with no CNV regions (Supplementary Fig. 1A).

Markedly, the contact probabilities of HCC cell lines and HMECs changed as the genomic distance passed 1 Mb (Fig. 1D and 1E). When the genomic distance was shorter than or equal to 1 Mb, the contact probability was higher for HMECs than for
Fig. 1. Long-range interactions are increased in hepatocellular carcinoma (HCC) cell lines. (A) Genome-wide Hi-C contact maps representing trans-contacts between chromosomes at a 500 kb resolution, normalized by expected contact counts. (B) Contact maps depicting a log2-fold change in contacts of HCC cell lines compared to those of human mammary epithelial cells (HMECs). (C) The percentages of cis-contacts per chromosome in HCC cell lines and HMECs are shown (left). A bar plot of averaged cis-contact ratios across the whole genome, with the ratio of cis-contacts in HMECs presented as a dashed line (right). (D) Averaged contact probabilities according to the genomic distance are shown, with dashed lines representing points at which the contact probabilities of HMECs and HCC cell lines are the same (top). The derivatives of the contact probability are shown in the subpanel, and the regions of minimum and maximum points are marked with dashed lines (bottom). (E) Relative contact probability plots represent a log2-fold change in contact probability between HMECs and HCC cell lines according to the genomic distance. The dashed lines mark the points with no fold change. (F) A box plot of the ratios between short-range cis contacts (shorter than or equal to 1 Mb) and long-range cis contacts in HCC cell lines and HMECs. p-values were calculated using the Wilcoxon rank sum test (*p < 0.005, **p < 1.0e-4, ***p < 1.0e-5).
HCC cell lines. As the genomic distance reached and then passed 1 Mb, the contact probability of HMECs shrank until the distance reached 40 Mb. Based on this observation, we divided intrachromosomal contacts into short- and long-range contacts based on a genomic distance of 1 Mb. The ratios between short vs. long-range interactions of HCC cell lines were significantly lower than those of HMECs, indicating that HCC cell lines had dominant long-range interactions (Fig. 1F, Supplementary Fig. 1D). In conjunction with frequent interchromosomal contact, HCC cell lines also revealed intensified contacts in long-range intrachromosomal interactions.

From the derivative of the contact probability plot, we could further infer the size of TADs and the linear density of cohesin [36]. The average TAD sizes, determined from the maximum points in the derivative plots, were slightly increased in HCC cell lines (Fig. 1D subpanel). The cohesin linear density, which was determined by the depth of the minimum point, was decreased in HCC cell lines; this could be interpreted as indicating the presence of weaker intra-TAD interactions based on the previous study [36], which found stronger TADs and chromatin loops were associated with more robust insulation and a higher density of cohesins. In summary, in HCC cell lines, shorter contacts (e.g., intra-TAD interactions) were decreased, and more elongated contacts (e.g., interchromosomal contacts) were increased. This change might cause the abnormal cancer phenotype of HCC cell lines.

Distinguishing compartment landscapes in HCC cell lines
Having observed a discrepancy in genome-wide chromosome contact between HMECs and HCC cell lines, we next defined compartmental domains to explore potential relationships between contact differences and 3D chromatin organization. We also analyzed open chromatin regions via ATAC-seq using two technical replicates.

We analyzed compartments through PCA analysis of Hi-C contacts at a 100 kb resolution. Positive PC values were taken as defining A compartments, and negative values were taken as delimiting B compartments [37]. The PC1 values of HCC cell lines showed moderate positive correlations with those of HMECs (Fig. 2A). Based on these scores, we calculated the Pearson correlations for hierarchical clustering (Fig. 2B). Huh1 and SNU449 cells had the strongest correlation with one another. Moreover, Huh1 and SNU449 cells had the lowest correlation scores with Huh7. To compare these results to the correlation scores obtained from the ATAC peaks, we mapped the trimmed ATAC-seq fastq files onto the hg19 genome and merged the mapping results of the replicates. After sorting and indexing, we calculated the Pearson correlation scores at a 100 kb resolution. The ATAC-seq peaks showed close correlations between SNU449 and Huh1 cells, followed by Huh7 and Hep3B cells, and thus supported the PCA analysis results (Fig. 2C).

Regarding compartmentalization changes between HMECs and HCC cell lines, we found that almost 50% of the compartments in HMECs were altered (A-to-B or B-to-A; called changed compartments or CCs) in HCC cell lines (Fig. 2D). Intriguingly, the CCs were unique in each HCC cell line. For example, some B-to-A domains in Hep3B were not changed in the other HCC cell lines. Accordingly, we analyzed the CCs in each cell line. Surprisingly, each cell line had only 25% CCs compared to HMECs, compared to the 50% difference between all HCC cell lines and HMECs (Fig. 2F). Thus, the CCs appeared to be distinctive to each HCC cell line.

Next, we generated heatmaps that sorted bins according to the comparison between HMECs and each HCC cell line (Fig. 2E, Supplementary Fig. 2). The A-to-B domains of Hep3B cells were mostly consistent with those of the other HCC cell lines, whereas the B-to-A domains of Hep3B cells only marginally overlapped with those of other HCC cell lines. For instance, in chromosome 13, there was a B-to-A domain in Hep3B cells that corresponded to a static B domain in the other HCC cell lines (Fig. 2G). Additionally, the PC values of most compartment B domains were lower in HCC cell lines than in HMECs. Unlike the compartment scores for compartment B, those for compartment A did not decrease by more than half in HCC cell lines compared to HMECs (Fig. 2E).

To summarize, compartment alterations in HCC cell lines were mostly found in the B compartment of HMECs, which were weakened or changed to A compartment regions in the HCC cell lines. Notably, these B-to-A domain changes were cell line-specific. Our results suggest that compartment analysis could potentially be used to classify the subtype of HCC, which is critical for selecting a cancer treatment strategy.

Intra-TAD interactions are decreased in HCC cell lines
Next, we explored shorter-range neighborhood interactions, namely TADs and chromatin loop interactions, in greater depth. Based on our initial results, we expected smaller TAD sizes and weaker intra-TAD interactions in HCC cell lines compared to HMECs. To analyze this in more detail, we defined TADs by RI analysis using CaTCH [18]. HMECs had more abundant (n = 9,944) TADs than HCC cell lines (Hep3B = 8,473, Huh1 = 9,159, Huh7 = 7,693, SNU449 = 9,553). As we expected, according to ATAC analysis, the intra-TAD interactions were diminished in HCC cell lines compared to HMECs (Fig. 3A). Subsequently, we calcu-
Fig. 2. Distinguishing compartment landscapes in hepatocellular carcinoma (HCC) cell lines. (A) Scatter plots of the compartment scores (PC1) of HCC cell lines relative to human mammary epithelial cells (HMECs) at a 100 kb resolution. Linear regression lines and correlation coefficients ($R^2$) are presented. (B) A heatmap showing Pearson correlations of PC1 correlation coefficients with hierarchical clustering. (C) A Pearson correlation heatmap generated from ATAC peaks with hierarchical clustering. (D) An accumulative column graph of the ratio of compartmental changes between HCC cell lines and HMECs (left). The genomic bins of CCs are depicted as a heatmap with hierarchical clustering (right). (E) Heatmaps of compartment scores in each cell type, sorted according to compartment alteration. The genomic bins were sorted by compartmental changes between HMECs and Hep3B cells (left) and between HMECs and Huh7 cells (right). (F) Accumulative column graphs of the percentages of different compartment transitions from HMECs to each HCC cell line. (G) Example of compartmental domains of HMECs and HCC cell lines. The example domains of cell line-specific CCs are marked with a dotted box.
labeled the insulation score, which provided an aggregate of interactions within a sliding square across the interval [18]. The insulation scores' local minima were considered to be the TAD boundaries. The insulation scores for TAD interactions of HCC cell lines were also higher (less negative), i.e., had smaller signal amplitude than those of HMECs, implying that the insulating abilities at TAD borders were affected under HCC (Fig. 3B). Huh7 and SNU449 cells exhibited particularly smaller signal amplitude at TAD boundaries than Huh1 and Hep3B cells. Weaker TAD borders could explain why the inter-TAD interactions and TAD sizes were increased in HCC cell lines compared to HMECs (Fig. 3C). Hep3B and Huh7 cells had greater inter-TAD strength and longer average TAD lengths than Huh1 and SNU449 cells.

Another critical interaction is the chromatin loop interaction. Chromatin loops are formed by the interaction between the pairs of loci that show significantly higher contact frequencies than their neighbors [11]. We defined loops via the HiCCUPS algorithm of juicer tools [20]. We detected 30,325 chromatin loops in HMECs and 21,646, 20,477, 14,280, and 15,056 chromatin loops in Hep3B, Huh7, Huh1, and SNU449 cells. Similar to the weaker TAD interactions, we observed that HCC cell lines had fewer and weaker chromatin loops than HMECs (Fig. 3D and 3E). The intensity and number of chromatin loops in SNU449 and Huh1 cells were significantly lower than in Hep3B and Huh7 cells.

To prevent chromosomal rearrangement from affecting our result, we also analyzed TAD domains and chromatin loops after excluding CNVs from the genome (Supplementary Table 1, Supplementary Fig. 1B–1F). The results were consistent with our previous results about the characteristics of TAD and chromatin loops of HCC cell lines compared to HMECs.

We also examined the similarities between chromatin loop interactions. Unexpectedly, only 4.4% of loop interactions were common to all four HCC cell lines and HMECs (Fig. 3F). Hep3B and Huh7 cells had the highest number of common peaks, followed by Huh1 and SNU449 cells. When we combined the results of all Hi-C analyses, including those of genome-wide contacts, compartments, ATAC-seq peaks, and TAD characteristics, the HCC cell lines clearly segregated into two subgroups based on chromatin 3D organization: one group comprising SNU449 and Huh1 cells and one comprising Hep3B and Huh7 cells. This tendency was also observed in the previous study, which identified six HCC subgroups through unsupervised transcriptome analysis [38].

**Alteration of 3D chromatin organization can disturb gene expression**

Finally, we assessed the relationship between 3D chromatin organization and gene expression. We used the public total RNA-seq results for HMECs, Hep3B, Huh1, Huh7, and SNU449 cells and mapped them to the hg19 genome using STAR [25]. We applied Cufflinks to identify differentially expressed genes (DEGs) of five cell lines compared to each other [39] and used cummeRbund to plot the results of our analysis [27]. We calculated the correlation of DEGs in HCC cell lines and HMECs (Fig. 4A left panel). The DEGs of Hep3B and Huh7 cells had the strongest relationship with each other, while those of HMECs had the weakest relationship with the DEGs of the HCC cell lines. This result was also depicted as a dendrogram (Fig. 4A right panel). There are three published strategies for subtyping HCC based on transcriptome analysis, namely the subgroupings reported by Boyault et al. [38] (G1 to G6, G-standard), Hoshida et al. [40] (S1 to S3, S-standard), and Caruso et al. [41] (CL1 to CL3, CL-standard). Based on G-standard, Huh1 and SNU449 are G3, Hep3B is G1, and Huh7 is G2. Based on S-standard, Hep3B, Huh1, and Huh7 are S2 and SNU449 is S2. Finally, based on CL-standard, Hep3B, Huh1, and Huh7 are CL1 and SNU449 is CL3. When taking together all these previous studies, SNU449 has the least similarity with other HCC cell lines, and Hep3B and Huh7 tend to be the most similar in the transcriptome. In our RNA-seq analysis, Hep3B and Huh7 had the highest correlation, and SNU449 had the lowest correlation with other HCC cell lines. Since our RNA-seq analysis results were closely related to three previous subtypings, we could conclude that this public RNA-seq data and our analysis were sufficient to support the observation of transcription level change. Importantly, as described earlier, Hep3B and Huh7 cells demonstrated the highest correlation in 3D chromatin organization and RNA-seq analysis, followed by Huh1 and SNU449 cells. This implies that 3D chromatin organization could also be another HCC subtyping standard.

Our RNA-seq analysis revealed that Hi-C analysis results could potentially be used as subtyping standards and supported the connection between 3D chromatin organization and gene expression level. Additionally, we found the expression level of SLC8A1 (encoding Solute Carrier Family 8 A1; chr2: 40,097,270-40,611,053) was lower in all tested HCC cell lines compared to HMECs (Fig. 4B). Notably, the TAD domain at this locus was altered in HCC cell lines relative to HMECs (Fig. 4C). In HMECs, the boundaries of the TAD containing SLC8A1 were located at 40 Mb and 40.46 Mb of chromosome 2. The two most highly ranked enhancers for SLC8A1 are GH02J0400449 and GH02J040511, located at 40,449,400 and 40,511,348 bp, respectively, and are thus found in the same TAD as SLC8A1 in HMECs. In HCC cell lines, however, the TAD of SLC8A1 was enlarged. We speculate that this altered
Fig. 3. Diminished intra-TAD (topologically associated domains) interactions in hepatocellular carcinoma (HCC) cell lines. (A) Aggregate TAD analysis (ATA) of the normalized contacts (top) and differential interactions (bottom) of HCC cell lines compared to human mammary epithelial cells (HMECs) in 9,944 HMEC TADs. (B) Averaged insulation scores at TAD boundaries of HMEC within ±1 Mb. Dashed lines mark the insulation scores of HMECs (black) and HCC cell lines at HMEC TAD boundaries. Calculated p-values with the Wilcoxon rank sum test are shown. (C) TAD n+1 plots (top) showing inter-TAD interactions with neighbor TADs of HCC cell lines compared to HMECs. The dashed line is located at zero. The box plot shows the TAD length distribution for each cell line (bottom). The medians of the TAD lengths are represented with white lines, and the median for HMECs is shown with a black dashed line. p-values were calculated using the Wilcoxon rank sum test (*p < 0.05, **p < 0.001, ***p < 1.0e-6). (D) Aggregate peak analysis of the normalized peaks (top) and differential peaks (bottom) of HCC cell lines compared to HMECs at 30,325 HMEC peaks within ±50 kb. (Continued to the next page)
TAD might not be able to support proximity between the enhancers and the promoter of \textit{SLC8A1} and that the smaller TAD might be necessary for the proper expression of \textit{SLC8A1}.

Moreover, the chromatin loops annotated in the \textit{SLC8A1} TAD of HMECs were not detected in the HCC cell lines. These discrepancies in 3D chromatin organization might form a basis for the differential expression of \textit{SLC8A1} in HCC cell lines. Furthermore, we found another gene, \textit{CHDR1} (encoding Cadherin related family member 1, chr10:84,194,635–84,219,621), that expression level reduction and TAD disruption both occurred (Supplementary Fig. 3). The low expression of \textit{CHDR1} is an unfavorable prognostic factor, and overexpression of \textit{CHDR1} could inhibit glioma cell growth [42]. In Hep3B, Huh1, and Huh7, the TAD at the \textit{CHDR1} gene locus enlarged. In SNU449, \textit{CHDR1} TAD was separated into two smaller TADs. Moreover, all HCC cell lines lost chromatin loop interaction at the \textit{CHDR1} locus. These disruptions of TADs and chromatin loops might cause the suppression of gene expression. These examples support the idea that there is a relationship between gene expression and 3D chromatin organization and further emphasize the importance of characterizing 3D chromatin organization in cancer.

**Discussion**

The study of epigenetics in HCC is essential for several reasons. Firstly, as epigenetic changes play a critical role in the development and progression of HCC [43], the study of such changes could critically help us understand the mechanisms underlying this disease. Secondly, targeting specific epigenetic changes could be a strategy for improving the diagnosis and prognosis of HCC [44]. Finally, knowledge of HCC epigenetics could facilitate the development of new therapies and personalized medicine, such as epideugs, by guiding researchers in leveraging relevant DNA methylation inhibitors, histone deacetylases, and other chromatin-modifying enzymes [45]. Therefore, the study of epigenetics in HCC has the potential to provide new insights into the nature of HCC and reduce the burden of this devastating disease worldwide.

\textit{In situ} Hi-C allowed us to characterize the 3D genome organization of four HCC cell lines (Hep3B, Huh1, Huh7, and SNU449) compared to HMECs. The HCC cell lines were found to have the following changes relative to HMECs: a higher frequency of long-range (>1 Mb) interactions, such as trans-interactions and intrachromosomal interactions; cell line-specific compartmental changes relative to HMECs; and reductions in the number and strength of TADs and chromatin loops. Using the correlation scores for all 3D chromatin structures, we could divide the four HCC cell lines into two subgroups: one comprising Hep3B and Huh7 cells and one comprising Huh1 and SNU449 cells. This subtyping was supported by the results of our RNA-seq analysis. Finally, we revealed that the gene expression level and 3D chromatin organization are linked in the case of \textit{SLC8A1}, which encodes a sodium-calcium antiporter that plays a crucial role in inducing apoptosis by increasing the influx of calcium ions [46-48].

We previously studied 3D chromatin organization in different breast cancer cells and tissues [49]. Both breast cancer and HCC cell lines showed increases in trans-contacts and distant interactions relative to those of HMECs. The TAD insulation scores and peak strength at chromatin loops were decreased in cancer cells.
Fig. 4. Changes in 3D chromatin organization can disturb gene expression. (A) A JS Distance correlation plot (left) and a dendrogram (right) of differentially expressed genes in hepatocellular carcinoma (HCC) cell lines and human mammary epithelial cells (HMECs). Three previous HCC subtyping strategies are indicated by colored boxes. (B) FPKM (fragments per kilobase of transcript per million) gene-level plot of SLC8A1 in HMECs and HCC cell lines. (C) Normalized Hi-C contact maps with TADs (topologically associated domains; green lines) indicated, spanning 34 to 41 Mb of chromosome 2 in the five cell lines. The white dashed lines mark the SLC8A1 locus, and the yellow dashed line marks the location of enhancers. (D) Normalized Hi-C contact maps indicating TADs (straight lines) and chromatin loops (dots) of HCC cell lines and HMECs. The TAD boundaries of each cell line and chromatin loops are marked with the following colors: HMEC, black; Hep3B, blue; Huh1, red; Huh7, green; and SNU449, yellow. The gray dots in the upper triangle mark the HMECs’ chromatin loops.

compared to HMECs. Reduced local contacts and increased global contacts may be common features of cancer cells. However, there were some differences between the two cancer types. In the breast cancer study, the BT549 cell line, which is a triple-negative breast cancer cell line, showed the most distinctive CCs from other breast cancer cells. In the present study, in contrast, the HCC cell lines all had similar A-to-B compartmental changes but cell line-specific B-to-A changes. This distinct B-to-A compartmental change may be a biomarker of HCC cell lines and could potentially be used to classify HCC cell lines.

Although we found some apparent cell line-specific 3D chromatin organizations in HCC, our study has limitations. We used HMECs as a control; as these cells originated from mammary tissue, our findings may not be cancer-specific features but rather liver-specific. Furthermore, we do not report on other epigenetic data, such as histone ChIP-seq results; thus, the results of our Hi-C analysis must be confirmed by additional experiments. Nevertheless, we herein reported distinguishing features of chromatin orga-
nization between HCC cell lines and showed that their contact characteristics coincided with those of the previously studied breast cancer cell lines. Finally, the utilized Hi-C analysis pipeline was one that had been previously verified by many studies; thus, our findings should be meaningful in a general sense.

Our results provide new epigenetic perspectives into HCC pathology and cancer biology that can be further explored through future research. There is a long-standing debate over whether cancer cell lines demonstrate phenotypes identical to cancer cells obtained from tissues [50], and studies have shown that a given cell line can display varying characteristics depending on the culture environment [51]. As a result, researchers are making significant efforts to build in vitro microenvironments for cancer cells, such as with 3D culture techniques [52]. As an extension of our research, by comparing the monolayer-cultured cancer cell lines, 3D-cultured cancer cell lines, and cancer cells from patients’ tissue samples, we will be able to distinguish three distinct epigenetic phenotypes. Moreover, we will identify the universal features of cancer cells compared to normal cells and determine which epigenetic markers can be used to discriminate cancer cell lines from tissues and normal cells. These findings can serve as a foundation for further research in cancer epigenetics and epigenetic drug development.

ORCID

Yeonwoo Kim: https://orcid.org/0000-0002-9410-4980
Hyeokjun Yang: https://orcid.org/0000-0003-0152-8520
Daeyoup Lee: https://orcid.org/0000-0003-2006-1823

Authors’ Contribution

Conceptualization: DL. Data curation: YK. Formal analysis: YK. Funding acquisition: DL. Methodology: HY. Writing – original draft: YK. Writing – review & editing: YK, DL.

Conflicts of Interest

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

Acknowledgments

We sincerely thank Professor Kyung Hyun Yoo of Sookmyung Women’s University for providing the four HCC cell lines, which were instrumental in this work. This research was supported by a National Research Foundation (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A2C3003115) and by an NRF grant funded by the Ministry of Science and ICT (MSIT) (2018R1A5A1024261, SRC).

Supplementary Materials

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

References


https://doi.org/10.5808/gi.23015


The genomic landscape associated with resistance to aromatase inhibitors in breast cancer

Kirithika Sadasivam, Jeevitha Priya Manoharan, Hema Palanisamy, Subramanian Vidyalakshmi*

Department of Biotechnology, PSG College of Technology, Coimbatore 641004, Tamil Nadu, India

Aromatase inhibitors (AI) are drugs that are widely used in treating estrogen receptor (ER)-positive breast cancer patients. Drug resistance is a major obstacle to aromatase inhibition therapy. There are diverse reasons behind acquired AI resistance. This study aims at identifying the plausible cause of acquired AI resistance in patients administered with non-steroidal AIs (anastrozole and letrozole). We used genomic, transcriptomic, epigenetic, and mutation data of breast invasive carcinoma from The Cancer Genomic Atlas database. The data was then separated into sensitive and resistant sets based on patients’ responsiveness to the non-steroidal AIs. A sensitive set of 150 patients and a resistant set of 172 patients were included for the study. These data were collectively analyzed to probe into the factors that might be responsible for AI resistance. We identified 17 differentially regulated genes (DEGs) among the two groups. Then, methylation, mutation, miRNA, copy number variation, and pathway analyses were performed for these DEGs. The top mutated genes (FGFR3, CDKN2A, RNF208, MAPK4, MAPK15, HSD3B1, CRYBB2, CDC20B, TP53TG5, and MAPK8IP3) were predicted. We also identified a key miRNA - hsa-mir-1264 regulating the expression of CDC20B. Pathway analysis revealed HSD3B1 to be involved in estrogen biosynthesis. This study reveals the involvement of key genes that might be associated with the development of AI resistance in ER-positive breast cancers and hence may act as a potential prognostic and diagnostic biomarker for these patients.

Keywords: acquired drug resistance, ER-positive breast cancer, non-steroidal Ais, pathway analysis, TCGA-BRCA patient data

Introduction

According to the World Health Organization (2020), about 2.3 million women were diagnosed with breast cancer around the world, and approximately 685,000 individuals are currently deceased [1]. There are different types of breast cancer, out of which 75% are hormone receptor–positive (HR-positive) breast cancer [2], that may be either estrogen receptor–positive (ER-positive) or progesterone receptor–positive. Most females across the world suffer from ER-positive breast cancer during their post-menopausal and pre-menopausal stages [3].

ER-positive breast cancer cells contain receptors for estrogen, which upon binding to the hormone help in the proliferation of cancer. This information is vital in deciding the treatment method for ER-positive breast cancer [4]. The existing breast cancer treatment targets the inhibition of estrogen and the prevention of hormone-receptor binding using ER modulators (tamoxifen), inhibitors (aromatase inhibitors), and ER degrading agents...
like fulvestrant [5,6]. Generally, combinatorial therapy is preferred which comprises surgery, endocrine therapy, chemotherapy, immunotherapy, etc. During such multiple treatments, aromatase inhibitors (AIs) are often administered to both pre and postmenopausal women as part of the endocrine therapy. Currently, AIs are reported to be one of the efficient drugs that are used in treating ER-positive breast cancer [7,8].

Aromatase inhibitors are drugs that inhibit the enzyme aromatase [9]. It is responsible for catalyzing the conversion of testosterone and androstenedione to estradiol and estrone, the rate-limiting step in estrogen biosynthesis [10]. Estradiol and estrone are the biologically active forms of estrogen detected in non-pregnant women. Researchers realized the need to target aromatase enzymes to improve ER-positive breast cancer treatment and started the development of AIs in the early 1970s [11]. Currently, there are three generations of AIs named in the chronological order of discovery. Aminoglutethimide is one of the well-known first-generation AIs. Fadrozole and rogeletimide belonging to the second-generation AIs were reported to be efficient in treating breast cancer with few undesirable pharmacokinetic properties [12]. The frequently used drugs like anastrozole (Arimidex), letrozole (Femara), and Exemestane (Aromasin) fall under the third generation of AIs [13,14]. The third generations of AIs are further classified based on their mechanism of action and chemical structure into steroidal (type 1) and non-steroidal (type 2) AIs [15–17]. Steroidal AIs such as Aromasin (third generation AIs) compete with the substrate of aromatase and bind covalently to the active sites resulting in an irreversible inhibition [18]. Whereas, non-steroidal AIs (anastrozole and letrozole) bind non-covalently to the heme moiety of the enzyme resulting in a reversible inhibition [19].

Even though AIs are very effective in treating breast cancer, drug resistance seems to be a major problem while administering them. There are two types of resistance that can be observed in patients, namely de novo/primary resistance and acquired drug resistance [15]. When a patient does not respond to the initial administration of the drug, then the individual possesses de novo/primary drug resistance. Whereas, when the patient is sensitive to the drug initially and later develops resistance after 2–3 years, it is categorized as an acquired drug resistance. Patients with acquired drug resistance often relapse. The current strategy that is being used to treat AI-resistant patients is to change the type of drug that is being administered to them [20–22]. For example, if a patient being treated with Aromasin develops acquired resistance, then the individual is continued with the administration of another AI like anastrozole. The development of acquired drug resistance is intricate with a complex interplay of multiple genetic profiles and hence, the exact mechanism of drug resistance must be explored. This would help in planning for better management of ER-positive breast cancer.

We aim to determine the plausible genetic and epigenetic factors that might play a vital role in the development of non-steroidal AI resistance. The breast cancer patients treated with non-steroidal drugs like anastrozole and letrozole were chosen for this analysis from the The Cancer Genome Atlas (TCGA) database. Complete genomic, transcriptomic, and epigenomic analyses were performed with this data set for the resistant and sensitive groups and compared to identify the underlying mechanism of drug resistance.

Methods

NCI’s Genomic Data Common (GDC) data portal was used to access TCGA project (https://portal.gdc.cancer.gov/). The Breast Invasive Carcinoma project (TCGA-BRCA) was chosen from the list of TCGA projects for this study. The entire study design is shown in Fig. 1.

Clinical data retrieval

Clinical data of patients administered with anastrozole and letrozole were downloaded from the TCGA database, using the TCGA-biolinks package in RStudio [23]. The retrieved sample data were further classified into resistant and sensitive groups based on the patient’s responsiveness to the drug and recurrence of the tumor. The sensitive group is defined by the patient showing response to the drugs after 2 years of initial treatment [24] and the individual not reporting a relapse. The resistant group is categorized as the lack of patients’ response to the drug after initial treatment and also shows signs of cancer relapse.

Transcriptomic analysis

Identification of differentially expressed genes by gene expression analysis

Processed RNA-sequencing data were retrieved from the TCGA database using TCGA-Assembler in RStudio [25]. A total of 464 genes were chosen for analysis from an exhaustive literature analysis [20,26–33]. The significant (p<0.05) differentially expressed genes (DEGs) with <0.5 log 2-fold change and >–0.5 log 2-fold change were identified and a volcano plot was constructed to represent the data.

Methylation data analysis

For methylation analysis, corresponding probes for differentially expressed genes were taken from TCGA-Illumina human DNA methylation 450K platform. Probes having differential beta values...
were calculated by subtracting the average beta value of each probe in the sensitive sample from the average beta value of the same in resistant samples. Probes having an absolute beta value difference of more than 0.2 and less than –0.2 with a significant p-value (p ≤ 0.05) were considered to be hypomethylated or hypermethylated probes respectively.

**miRNA analysis of the DEGs: miRNA analysis**

Using the expression data available for miRNA for the sensitive and resistant samples in the TCGA dataset, the fold change of each miRNA in the above-fetched genes were obtained. Then, miRNAs that were differentially expressed in the TCGA dataset were subjected to analysis of both validated and predicted gene targets in the Mirwalk database (http://mirwalk.umm.uni-heidelberg.de/) [34]. These gene targets were then verified using the miRDB (http://mirdb.org/cgi-bin/search.cgi) [35], and the Targetscaen databases (http://www.targetscan.org/vert_80/) [36]. From this analysis, a list of miRNAs targeting certain differentially expressed genes were obtained. The miRNA which was reciprocally regulated as compared to their respective gene expression patterns were represented in the results.

**Genomic analysis**

**Mutational analysis**

Mutation data of the resistant and sensitive sets were also downloaded using TCGA-Assembler in R Studio from the TCGA database. The retrieved data was then filtered and the top 10 mutated genes were identified by plotting a Plotmaf summary in RStudio using the TCGAbiolinks package. A Venn diagram showing the mutated genes in sensitive and resistant data sets was plotted. To identify the percentage of mutation within the sensitive and resistant sets, an oncoplot was plotted for the top mutated genes using ggplot2 package (version 3.3.5). Further, lollipop (g3viz package-version 1.1.4) plots were plotted for the identified top mutated genes that were unique in the resistant set using g3viz package (version 1.1.4).

**Copy number variation analysis of the DEGs**

The TCGA-Assembler was used to download the copy number variation data of the identified DEGs from the TCGA database. The Genomic Analysis of Important Aberrations (GAIA) plots were drawn for the processed copy number variation (CNV) in RStudio using GAIA package version 2.36.0.

---

Fig. 1. Outline of the computational analysis performed for identifying the genomic landscape for non-steroidal aromatase inhibitors resistance in breast cancer. TCGA, The Cancer Genome Atlas.
Protein-protein interaction network
A protein-protein interaction network (PPI) was constructed for the DEGs (17) using the STRING database (https://string-db.org/) [37] with a minimum confidence of 0.150. The genes in the built network were then clustered by the K-means clustering module of the STRING server.

Pathway analysis of DEGs
Pathway enrichment of the identified DEGs was performed using the Reactome browser (https://reactome.org/) [38,39]. Significant pathways of the DEGs were explored and the predominant genes involved in estrogen biosynthesis were identified.

Results
Genetic and epigenetic alterations of DEGs involved in resistance to non-steroidal aromatase in breast cancer
A total of 322 patient barcodes were retrieved from the TCGBA-BRCA project (TCGA database). Among the 322 patient barcodes, 150 patient barcodes were segregated as sensitive and 172 as resistant categories and were used for further analyses. We identified the DEGs among the resistant and sensitive categories. It was found that 121 genes were significantly expressed, of which 16 genes were up-regulated and 1 gene was down-regulated (Fig. 2, Supplementary Table 1).

The epigenetic mechanisms behind differential regulation were analyzed for the DEGs. Differential methylation analysis indicated that the DEGs were not subjected to epigenetic alterations.

To understand the role of miRNAs in the regulation of the differentially expressed genes, Mirwalk was used to analyze the predicted and validated miRNAs targeting these genes. In our analyses, miRNAs that were upregulated in the TCGA dataset targeting down-regulated genes and vice versa were focused on. Using the TCGA expression dataset for miRNA and mRNA, it was found that a total of 20 miRNAs were differentially regulated, 19 miRNAs being up-regulated and 1 miRNA being down-regulated. The gene CD20B and miRNA hsa-mir-1264 were found to obey the reciprocal rule between the target gene and miRNA expression.

Fig. 2. Volcano plot displaying the differentially regulated genes (DEGs) obtained in gene expression analysis. The upregulated (16) and the downregulated genes (1) are represented in red and blue dots respectively. The green dots indicated the non-regulated genes in the array.
Genetic mutations of DEGs that might drive non-steroidal aromatase resistance in breast cancer

To investigate mutations in the DEGs identified in this study, we analyzed the mutations in these genes in the whole-exome sequencing (WES) data of breast cancer samples from TCGA. Analysis of TCGA WES data of the samples indicated that there were eight genes significantly mutated in the resistant samples and six genes significantly mutated in the sensitive samples. The results obtained from the analysis are depicted as a Venn diagram shown in Fig. 3. The genes TP53TG5 and MAPK8IP3 were distinctively mutated in the sensitive set (Fig. 3). Similarly, the genes CDKN2A, MAPK15, HSD3B1, and CRYBB2 were distinctively seen to be mutated in the patients showing resistance (Fig. 3). These unique genes (not mutated commonly in both groups) are proposed to influence the drug resistance mechanism. Hence further analyses were carried out with these unique genes.

The oncoplot shows the genes that are highly mutated in the sensitive and resistant set (Fig. 4A and 4B). The plot indicates the percentage population possessing the mutation along with the type of mutation for the corresponding gene. Fig. 4 represents the predominantly mutated genes in the population. The unique genes reported above are also seen to be predominantly mutated in both the sensitive and the resistant sets. Genes such as TP53TG5 and MAPK8IP3 are present in the top mutated genes in the oncoplot for the sensitive set (Fig. 4A). Similarly, CDKN2A, MAPK15, HSD3B21, and CRYBB2 are present among the top mutated genes in the oncoplot plotted for the resistant set (Fig. 4B).

Further, the lollipop plots reveal the location of the mutation present in these unique genes (Supplementary Table 2). Supplementary Fig. 1 shows the location of mutations in the unique genes observed in the sensitive samples and Supplementary Fig. 2 for the resistant samples. A missense mutation was seen in the TP53TG5 gene outside the TP53IP5 domain region, indicating an alteration in its function (Supplementary Fig. 1A). Additionally, a missense mutation was seen in the MAPK8IP3 gene at the PARP domain (Supplementary Fig. 1B).

A frameshift deletion and a frameshift insertion mutation were observed in the gene CDKN2A (Supplementary Fig. 2A). A missense mutation was noticed after 200 bp in the CRYBB2 gene (Supplementary Fig. 2B). The domain 3-beta HSD of the HSD3B1 gene tended to have a missense mutation in the resistant group (Supplementary Fig. 2B).

Fig. 3. Venn diagram for mutational analysis of differentially regulated genes. The genes that are uncommon between the sensitive and resistant set are represented.

Fig. 4. Oncoplots for the sensitive and resistant barcodes of the genes that are predominantly mutated in the population. (A) Sensitive group. (B) Resistant group.
Similarly, a missense mutation was noted at 544 bp of the MAPK15 gene (Supplementary Fig. 2D).

**CNV analysis**

The GAIA plots obtained for the CNV data of the two groups help us in visualizing the variation in the DEGs. Fig. 5 represents the GAIA plots of sensitive and resistant groups. The plots reveal that the genes MAPK15 and MAPK8IP3 located in chromosomes 8 and 16 respectively were amplified in both sensitive and resistant samples. It was also seen that the gene GSTM2P1 located in chromosome 6 was deleted in both the sample sets, whereas the gene CDK2NA present on chromosome 9 was deleted only in the sensitive set.

**Protein-protein interaction of the DEGs**

The PPI network of the DEGs was built with 16 nodes and 17 edges with a medium confidence score of 0.150 and an enriched p-value of 0.000478 (Supplementary Fig. 3) using the STRING server. Clustering of DEGs resulted in three clusters with an average local clustering coefficient of 0.719 (Fig. 6). The top cluster (cluster 1) comprised the proteins MAPK4, CDKN2A, MAPK15, FGFR3, and MAPK8IP3. Whereas cluster 2 and cluster 3 included CRYBB2, CRYBA4, CDC42EP5, CRYGS, CRYBB3 and COMTD1, HSD3B1, RNF151, RNF208, respectively.

**Pathway analysis of DEGs**

The significant pathways associated with the DEGs were explored using the Reactome tool. The DEGs (17) were enriched in 105 pathways (data not shown), where 7 pathways (Table 1) corresponded to estrogen biosynthesis. The remaining pathways of DEGs were related to cancer, disease, cell cycle, signaling, and other regulating pathways. From the pathway analysis, it was observed that the gene HSD3B1 was indirectly involved with estrogen biosynthesis.

**Discussion**

AI drug resistance is one of the major problems caused while treating ER-positive breast cancer. Identification of a significant biomarker for predicting non-steroidal AI drug resistance will help clinicians with the problems caused due to resistance. Although several studies have been done to delineate AI resistance, the genetic mechanism of resistance is still not uncovered. This study was aimed at identifying the prognostic and diagnostic biomarkers for AI resistance through available data sets and computational analysis. The breast cancer patients’ data was downloaded and differential gene expression was analyzed. It was found that 17 genes were differentially expressed, further genomic and transcriptomic analyses revealed the reason behind the dysregulation of the identified DEGs. Mutational analysis of the DEGs further narrowed down the genes that need to be focused. We found that epigenetic mechanisms are not the underlying reason behind the differential expression of the genes identified in this study. Besides, we found the involvement of the miRNA hsa-mir-1264 in regulating the expression of the CDC20B gene. It has been previously reported that the over-expression of CDC20 resulted in poor response in patients undergoing endocrine therapy and hence it acts as a biomarker for endocrine therapy resistance in ER-positive breast cancer patients [40].

---

Fig. 5. Genomic Analysis of Important Aberrations (GAIA) plots drawn for visualization of variation in differentially regulated genes. (A) Sensitive dataset. (B) Resistant dataset.
Fig. 6. Clustering of differentially regulated genes by STRING database to visualize the genes that are similar.

### Table 1. Estrogen biosynthesis pathways of the differentially expressed gene (HSD3B1) predicted using the Reactome tool

<table>
<thead>
<tr>
<th>Pathway name</th>
<th>Entities found</th>
<th>Entities total</th>
<th>Entities p-value</th>
<th>Reactions found</th>
<th>Reactions total</th>
<th>Species name</th>
<th>Submitted entities found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mineralocorticoid biosynthesis</td>
<td>1</td>
<td>22</td>
<td>0.0364</td>
<td>2</td>
<td>7</td>
<td>Homo sapiens</td>
<td>HSD3B1</td>
</tr>
<tr>
<td>Androgen biosynthesis</td>
<td>1</td>
<td>27</td>
<td>0.0444</td>
<td>1</td>
<td>9</td>
<td>Homo sapiens</td>
<td>HSD3B1</td>
</tr>
<tr>
<td>Glucocorticoid biosynthesis</td>
<td>1</td>
<td>29</td>
<td>0.0476</td>
<td>2</td>
<td>9</td>
<td>Homo sapiens</td>
<td>HSD3B1</td>
</tr>
<tr>
<td>Metabolism of steroid hormones</td>
<td>1</td>
<td>72</td>
<td>0.1143</td>
<td>5</td>
<td>40</td>
<td>Homo sapiens</td>
<td>HSD3B1</td>
</tr>
<tr>
<td>Metabolism of steroids</td>
<td>1</td>
<td>328</td>
<td>0.4277</td>
<td>5</td>
<td>244</td>
<td>Homo sapiens</td>
<td>HSD3B1</td>
</tr>
<tr>
<td>Metabolism of lipids</td>
<td>1</td>
<td>1444</td>
<td>0.9227</td>
<td>5</td>
<td>954</td>
<td>Homo sapiens</td>
<td>HSD3B1</td>
</tr>
<tr>
<td>Metabolism</td>
<td>1</td>
<td>3646</td>
<td>0.9992</td>
<td>5</td>
<td>2260</td>
<td>Homo sapiens</td>
<td>HSD3B1</td>
</tr>
</tbody>
</table>

Moreover, in our study, the *CDC20B* gene, a homolog of CDC20 was noticed to be downregulated and was also significantly mutated. Therefore, we hypothesize the role of CDC20B in acquired non-steroidal AI resistance in breast cancer patients.

Pathway analysis indicated the involvement of the gene *HSD3B1* in estrogen biosynthesis. *HSD3B1* codes for 3 beta-hydroxysteroid dehydrogenases and is responsible for catalyzing delta-5-3-beta-hydroxysteroid precursors into delta-4-ketosteroid through an oxidation reaction [41] in steroid biosynthesis. It is also responsible for the conversion of dehydroepiandrosterone (DHEA) to androstenedione in estrogen biosynthesis [42]. As delta-4-ketosteroid and androstenedione are essential for the synthesis of all steroid hormones, it can be speculated that HSD3B1 is the key factor influencing the conversion of DHEA to androstenedione. We found that the HSD3B1 gene was upregulated in the resistant samples, indicating the increased expression of HSD3B1 in breast cancer patients on treatment with the non-steroidal AI drugs (anastrozole and letrozole). The overexpression of HSD3B1 could result in the increased production of androstenedione, which in turn elevates the estrogen hormone levels of the tumor cell. This overall mecha-
nism would lead to the proliferation of cancer cells and pave way for non-steroidal AI resistance. The mechanism of HSD3B1 in the development of AI resistance in HR-positive breast cancer has been documented earlier [43]. These findings strengthen our prediction of HSD3B1 as a significant biomarker in AI resistance.

From the PPI network analysis, the DEGs were noticed to be highly interacting with one another. This indicates that the expression of DEGs can be influenced by one another, although further studies are required to understand the mechanism of interaction.

Additionally, it was noticed in our CNV analysis that the gene CDKN2A has been deleted in sensitive patients whereas it remained unaltered in resistant patients. CDKN2A gene is responsible for coding several proteins, especially p16, a cell division regulating protein, and a tumor suppressor gene [44]. Moreover, it was proven that the presence of mutation, differential expression, or copy number variation in the CDKN2A gene enhances tumor formation [45-47]. Surprisingly in our CNV analysis, CDKN2A gene deletion was restored in patients who are resistant to the non-steroidal AIs. Even though researchers are suggesting that CDKN2A plays a vital role in drug resistance in a variety of cancer types [48,49], its exact role and mechanism remain imprecise.

Furthermore, our findings included some upregulated genes (MAPK4, MAPK15, and MAPK8IP3) that were predicted to be mutated majorly. Previous studies have implicated the role of the MAPK pathway in AI resistance [5,20,50], but the exact mechanism behind it is still a conundrum. Studies have also shown that the increased activity of the MAPK pathway and the dysregulation of the genes involved in the pathway might contribute to the resistance of AIs [15]. Moreover, AI resistance is reported to be associated with the activation of the MAPK pathway in ER-positive breast cancer [51]. Hence, the activation of the identified DEGs (MAPK4, MAPK15, and MAPK8IP3) might trigger the MAPK pathway and thereby influence drug resistance during cancer therapy. Overall, MAPK4, MAPK15, and MAPK8IP3 may be considered significant genes in developing drug resistance toward AIs in ER-positive breast cancer patients.

The role of other DEGs (CDC42EP5, COMT1, CRYBA4, CRYBB2, CRYBB3, CRYGS, FGFR3, GSTM21P1, RNF151, RNF208, and TP53TGS) in ER-positive cancer as well as in developing AI resistance is yet to be explored. Further in vitro and in vivo studies are required to support the involvement of CDC20B, HSD3B1, and CDKN2A, as possible prognostic biomarkers of non-steroidal AI resistance in breast cancer.

The development of drug resistance imposes a greater difficulty in the treatment of cancer and thus finding a suitable biomarker for drug resistance and proposing a feasible mechanism is crucial. In this study, we have identified a few significant genes like CDC20B, HSD3B1, CDKN2A, MAPK4, MAPK15, and MAPK8IP3 that can act as potential biomarkers for non-steroidal AI resistance in ER-positive breast cancer patients. Further studies warrant a clear understanding of the mechanism behind the resistance conferred by these marker genes.

**ORCID**

Kirithika Sadasivam: https://orcid.org/0009-0006-3856-4954
Jeevitha Priya Manoharan: https://orcid.org/0000-0001-6292-6018
Hema Palanisamy: https://orcid.org/0009-0007-8643-3315
Subramanian Vidyalakshmi: https://orcid.org/0000-0003-1702-9064

**Authors’ Contribution**

Conceptualization: KS, JPM, HP, SV. Data curation: KS, JPM. Formal analysis: NA. Methodology: KS, JPM, SV. Writing – original draft: KS, JPM. Writing – review & editing: KS, JPM, HP, SV.

**Conflicts of Interest**

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

**Acknowledgments**

We thank the management of PSG College of Technology, Coimbatore for their support and infrastructural facilities provided to carry out this work.

**Supplementary Materials**

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

**References**

3. Surakasula A, Nagarjunapu GC, Raghavaiah KV. A comparative...


33. Kurokawa H, Lenferink AE, Simpson JF, Piscacce PI, Sliwkowski MX, Forbes JT, et al. Inhibition of HER2/neu (erbB-2) and mi-


Introduction

Fusobacterium nucleatum strain ATCC 25586 is an anaerobic, gram-negative, and opportunistic pathogen which belongs to Bacteroidaceae family. *Fusobacterium* mostly inhabits the oral cavity and throat of the diseased as well as normal individuals by adhering (through FadA protein) and invading the epithelial cells of mouth and gut. This bacterium forms a biofilm and alters the host immune response through the process of adhesion and invasion of critical organs like head, neck, lung, liver, heart, and brain. *Fusobacterium* infection can cause diseases like periodontitis, gingivitis, and appendicitis [1].

It has been reported that *Fusobacterium* crosses the placenta and causes preterm and still birth in women having pregnancy-associated gingivitis [2]. This bacterium is also...
found to be associated with colorectal cancer progression in those patients who have been suffering from long-term intestinal bowel disease [3]. Antibiotics therapy and surgical treatment (in some cases) are used to treat several diseases caused by F. nucleatum. Continuous administration of these antibiotics can cause antimicrobial resistance in bacteria. Therefore, the identification of new therapeutic target and development of new drugs against this bacterium can help in reducing the burden of disease. Shotgun genome sequencing of F. nucleatum strain ATCC 25586 reveals 2.17 Mb of genome containing a single circular chromosome having 27% GC content and 2,067 open reading frames [4]. Some of these open reading frames have been listed as “hypothetical” or “uncharacterized” proteins. These hypothetical proteins (HPs) are functionally and structurally uncharacterized and are classified into uncharacterized protein families and domain of unknown function classes. While many proteins are characterized during the sequencing itself, some of them remain uncharacterized due to lack of better sequence homolog or structurally related protein. It is therefore important to revisit such uncharacterized proteins and assign them functions in the context of new scientific knowledge. These previously uncharacterized proteins may yield interesting results and shed some light on functionality of a cell [5]. Many researchers have used this computer guided approach to functionally annotate the uncharacterized protein or HP from different organisms [6-8]. In the present study, we have attempted the functional annotation of uncharacterized proteins present in the genome of F. nucleatum strain ATCC 25586. Out of 398 uncharacterized proteins listed in F. nucleatum genome, we have assigned functions to 39 sequences with high confidence and another 7 with relatively low confidence. The receiver operating characteristics (ROC) analysis performed to evaluate the methodology adopted, yielded an average accuracy of 83% across the parameters.

Methods

Sequence retrieval
Proteome data of F. nucleatum strain ATCC 25586 was downloaded using Proteome ID UP000002521 format from UniProt database. Proteome of F. nucleatum strain ATCC 25586 contains 2,046 proteins, out of which 398 proteins are listed as 'uncharacterized proteins'. FASTA sequences of these 398 uncharacterized proteins were used for further analysis.

Physicochemical properties
Physical and chemical properties of a protein such as molecular weight, extinction coefficient, isoelectric point, grand average of hydropathicity, etc. were estimated through Expasy’s ProtParam program [9]. ProtParam computes physicochemical properties using the protein sequence only. Negative grand average value of hydropathicity (GRAVY) shows the hydrophilic nature of protein and vice-versa. The instability index of less than 40 indicates a stable protein.

Sub-cellular localization
Prediction of protein localization was done through Cello server. It uses a hybrid approach i.e., support vector machines model and a structural homology approach for localization prediction [10]. SignalP 5.0 was used for predicting signal peptide and cleavage site in a protein's sequence. Signal peptide is a small sequence present on the protein which directs them for movement to target position in the cell. Signal peptides are generally cleaved by signal peptidases after the translocation. SignalP 5.0 uses a deep neural network-based approach to identify the signal peptides [11]. TMHMM server was used for predicting the presence of transmembrane helices in a protein [12].

Domain identification
The protein sequences were then subjected to domain identification using the InterProScan, Motif, Smart, HMMER, NCBI CDART (Conserved Domain Architecture Retrieval Tool), and BlastP programs.

InterProScan server classifies the protein sequences into homologous superfamilies and identifies the functional domain based on information compiled from different databases. Besides this, InterProScan can also identify the presence of signal peptide and transmembrane helices in protein sequences [13]. Motif server was used for identification of motifs in a protein sequence which is available in GenomeNet database. SMART (Simple Modular Architecture Research Tool) web server in combination with UniProt, Ensembl, and String database analyze the domain architecture using per-species protein clustering procedure (normal mode) and completely sequenced genome (genomic mode) [14]. HMMER web server uses jack-hammer algorithm for the annotation of protein sequence based on identified domain [15]. NCBI CDART performs an RPS-BLAST against Entrez protein database for domain similarity in query protein sequence [16]. BlastP uses heuristic approach to identify the sequence similarity between input sequences and database sequences [17].

Performance assessment
ROC, a web-based calculator was used to check the accuracy, sen-
sitivity, and specificity of the different servers used in this study (Supplementary Table 1) [18]. About 50 proteins with known functions were randomly selected from *F. nucleatum* (Supplementary Table 1) and their functions were predicted against the same databases as used for the prediction of HPs in this study. Six-level classification of predictions were done using binary (0,1) format in which 0 represents true negative and 1 represents true positive. Integers ‘2’, ‘3’, ‘4’, and ‘5’ were used for the diagnosis of efficacy, higher the integer, higher the efficacy. ROC web server generates a ROC curve between sensitivity and 1-specificity, where area under the curve represents the effective measures of accuracy ranging from 0 to 1. The average accuracy and ROC area of the used database/s were determined to be 83.6% and 0.90, respectively (Supplementary Table 2).

String analysis
The functional partners reveal important information about a protein and its function. To search for such information, we subjected the 46 annotated proteins to the string database search [19]. After preliminary analysis, proteins with a confidence score of >1 were listed.

Homology detection with the human proteins
All the 46 annotated proteins were searched in BLASTp program against the human proteins (taxon id: 9606) in non-redundant database. These proteins were also searched in DrugBank database for identification of any similar druggable candidates [20].

Structure prediction and modeling
Homology-based structural modeling was carried out for the annotated proteins using Swiss PDB [21] and Phyre2 servers [22]. Templates with most sequence coverage were subsequently used for model building. Structure models were predicted for 25 annotated proteins with identity ranging from 14% to 97%. Based on the annotations, we identified several proteins which might play important role in cell survival and therefore, can be a potential drug target. Models of some of these proteins were explored further. The structure models were uploaded on PDBSum [23] page and their structural quality was assessed by PROCHECK [24].

Virulence prediction
The annotated protein sequences were analyzed for virulence factor prediction using the VICMPred [25] and VirulentPred [26] softwares. Proteins which were determined as virulent factors by both programs were analyzed further.

The characterization of the previously uncharacterized proteins employed a large number of available programs and servers which predict the essential parameters such as localization, domains, motifs, interactions, etc. The inference of the probable function for these proteins is based on the collective results of all these programs (Fig. 1).

---

**Fig. 1.** Flowchart of methodology. The methodology adopted for functional annotation of uncharacterized proteins of *Fusobacterium nucleatum* included the servers for sequential characterization, sub-cellular location, domain and motif identification, structure prediction, and virulence factor identification. Only the results with high confidence are taken and the rest are excluded.
Results and Discussion

Physico-chemical parameters
Assigning the function of HPs bridges the gap in the knowledge of protein structure-function relationship and may reveal information about novel pathways responsible for pathogenesis. Based on the structural and functional information, they can be used as a drug targets or biomarkers for disease identification. Using different online servers, we tried to annotate the HPs present in genome of *Fusobacterium nucleatum* subsp. *nucleatum*. Predicted physico-chemical parameters of all uncharacterized protein/HP are tabulated in Supplementary Table 3. These parameters provide an insight into the protein, such as pI value, extinction coefficient, etc. HPs with following accession IDs Q8RDL8, Q8REK2, Q8RH50, and Q8RIC0 have not shown extinction coefficient value due to absence of cysteine, tryptophan, and tyrosine residue. Most of the proteins are hydrophilic in nature as they have low GRAVY value. Forty-three percent of the total uncharacterized proteins had the acidic pH (pH < 7) while the rest 57% were basic (pH ≥ 7) in nature.

Localization
Cello server does not rely solely on the homology of the sequences but on the combination of two-level support vector machine classifiers to determine the subcellular location and thus, reduces the bias while increasing the accuracy [10]. Among 398 uncharacterized proteins, most of the proteins (74%) were predicted to be localized in cytoplasm whereas 12% and 7% of proteins were localized in the inner and outer membrane of the cell respectively. Only 1% of the proteins were found to be localized extracellular and 6% of proteins were localized in periplasm (Fig. 2, Supplementary Table 4). Presence of a signal peptide in a protein determines the translocation of protein inside or outside the cell. Signal peptide prediction was done using the SignalP 5.0 server which predicts its presence along with the location of their cleavage sites in bacterial proteins [27]. A total of 47 proteins were predicted to have signal peptide at their N-terminal, among which 30 proteins have standard signal peptide cleaved by signal peptidase I and the rest (17) have lipoprotein signal peptide cleaved by signal peptidase II (SPII) (Supplementary Table 4). One hundred forty-one proteins were predicted to be transmembrane proteins which might be involved in transport and signal transduction. Transmembrane proteins, especially the outer membrane proteins of the gram-negative bacteria behave as virulence factors and also help the pathogen in escaping defence mechanism of host [28].

Domain identification
Annotation of uncharacterized proteins using InterProScan, Motif, SMART, HMMER, NCBI CDART, and BlastP search led to the identification of 90 proteins having functional domains (Supplementary Table 5). For increasing the accuracy of the results, we assigned the probable function to only those protein sequences whose conserved domains were predicted by two or more databases. As per this convention, out of 90 HPs with functional domains, functions were successfully assigned to 39 proteins with high confidence (Table 1) and other 7 proteins with relatively low confi-

![Fig. 2. Sub-cellular localization. Pie-chart depicting the localization of uncharacterized proteins in *Fusobacterium nucleatum* as determined by the Cello server.](https://doi.org/10.5808/gi.22065)
dence (Table 2).

**Predicted function**

We could predict the function of 46 HPs out of which 17 proteins (37%) are enzymes, 5 (13%) are binding proteins, 10 regulatory proteins (21%), 3 transport proteins (6%), 2 are phage related (4%), 2 are membrane proteins (4%), and 7 proteins are involved in other functions (15%) (Fig. 3).

**Enzymes**

Enzymes are the proteins that catalyze various metabolic pathways essential for the survival of an organism. *F. nucleatum* uncharacterized enzymes were identified as enzymes: Q8R669, Q8REG3, Q8REM4, Q8RF86, Q8RFU1, Q8RG23, Q8RDY5, Q8R699, Q8RH78, Q8RM7, Q8RGP8, Q8RHS6, Q8RII7, Q8RE80, Q8RF13, Q8RG9, Q8RH12, and Q8RF1 (Table 1).

Q8R669 belongs to a nucleoside phosphorylase superfamily involved in S-adenosylmethionine mediated reaction. Enzymes of this family play a vital role in biofilm formation and pathogenesis of an organism. Q8R699 might be involved in these functions and can be used as a drug target for antimicrobial treatment [29].

Q8REG3 is D-component of 2-hydroxyglutaryl-CoA dehydratase (HGD-D) which undergoes dehydration to form enoyl CoA for the fermentation of α-amino acid. HGD consist of 2 components: component A which acts as an activator and component D which is a dehydratase enzyme. component A transfers electron to component D which in turn transfers electron to its substrate and thus perform the elimination of hydroxyl group [30].

Q8REM4 is a PGAP-1 like protein which encodes for glycosyl phosphatidylinositol (GPI) inositol deacylase responsible for deacylation and transport of GPI-anchored protein from endoplasmic reticulum to Golgi [31].

Q8REM4 is a PGAP-1 like protein which encodes for glycosyl phosphatidylinositol (GPI) inositol deacylase responsible for deacylation and transport of GPI-anchored protein from endoplasmic reticulum to Golgi [31].

Q8RFU1 is a LpxI metal dependent hydrolase that catalyzes water mediated hydrolysis of β-phosphate of UDP-2,3-diacylglycerol (Table 2).
amine into lipidX in lipid A biosynthesis. Lipid A is essential for pathogenesis and viability of bacteria, thus, making Q8RFU1 a therapeutic target [33,34].

Q8RGP8 was identified as arginine deiminase, a homolog of DDAH (N,N-dimethylarginine dimethylaminohydrolases) which belongs to superfamily amidinotransferase. Arginine deiminase is involved in arginine metabolism in which NH$_4^+$ is produced. This NH$_4^+$ protects the bacteria from host acidic environment by raising the cytoplasmic pH. Based on the above function, this enzyme can be used as a probable drug target [35,36]. Q8RH78 is predicted to be an acyl-coenzyme A:6-aminopenicillanic acid acyl-transferase, known as the last enzyme that catalyzes the synthesis of antibiotic in penicillin biosynthesis pathway [37].

Q8RHS6 was characterized as a type IS restriction endonuclease that protects bacterial cell by recognizing and cleaving asymmetric sequences of bacteriophage DNA [38]. Q8RJ7 was identified as a YmdB-like protein which belongs to calcineurin-like phosphatase/phosphodiesterase family. YmdB contains binuclear metal center which helps in biofilm formation and motility regulation [39].

Q8RE08 is an O-antigen ligase required for O-antigen ligation reaction in which lipid A attaches to core oligosaccharide and O antigen for the formation of lipopolysaccharide layer [40].

Q8RF9 is a DNA repair enzyme protecting cell from the cytotoxic and mutagenic alkylating agents [41].

Q8RH84 belongs to the PD-(D/E)XXK nuclease superfamily 9 involved in a variety of functions such as DNA restriction, repair, modification, tRNA splicing, transposon excision, Holliday junction resolving, Pol I termination, etc. [42].

Q8RJ95 is a permuted papain-like amidase enzyme thought to be involved in host-pathogen interactions and could be a potential drug target [43]. Q8RE1 is identified as small electron transfer protein known as flavodoxin. Flavodoxin proteins contain a non-covalently bonded flavin mononucleotide molecule as cofactor which also acts as a redox site [44]. This protein is involved in different metabolic pathways like nitrogen fixation and has the potential to be used as a therapeutic target [45].

Q8REJ6 was characterized as a thioredoxin protein having a conserved ‘thioredoxin motif’. Thioredoxins are involved in transferring of electrons from NADPH to thioredoxin via thioredoxin reductase. Thioredoxins have a role in DNA synthesis, protein repair, sulfur assimilation and in oxidative stress [46]. Q8RE14 is a cysteine protease PrP, responsible for cleaving L27 protein for efficient functioning of ribosome. Defective PrP leads to uncleaved L27 protein resulting in inhibition of bacterial growth [47].

**Regulatory proteins**

A total of 10 proteins (Q8RDY5, Q8REC7, Q8RFD4, Q8RG23, Q8RH3Q, Q8R1D9, Q8REK7, Q8RG29, Q8RF29, and Q8RE9) were identified as regulatory proteins performing different functions. Q8RDY5 is a translocon component of type I secretion system which enhances the serine sensitivity in a cell as serine is known to cause the inhibition of bacterial growth [48]. Q8REC7 was identified as a Cas7 or DevR protein which along with DevS has a regulatory role in fruiting body development in Myxococcus xanthus [49]. Q8RFD4 may act as a RelB regulatory protein which is an anti-toxin component of type I toxin-antitoxin complex. RelB inhibits RelE (toxin) functioning and binds to Rel operator thus allowing the transcriptional auto regulation [50]. Q8RG23 protein was identified as a ParD antitoxin protein, cognate of ParDE toxin-antitoxin system. ParE toxin inhibit the DNA synthesis and cell growth of bacteria. This activity of ParE is prevented by ParD anti-

![Probable functions of uncharacterized proteins](https://doi.org/10.5808/gi.22065)
Bacterial exotoxin mediates the ADP-ribosylation in target eukaryotes. Proteins with this domain, have diverse roles in regulation of ADP-ribosylation, DNA repair, and transcriptional regulation. Bacterial exotoxin mediates the ADP-ribosylation in target protein of host cell, thus contributing to the onset of infection [2].

Q8RHD9 was predicted to be a RecG helicase which is a double-stranded DNA translocase. RecG regulates DNA transcription and avoids origin-independent pathological DNA synthesis by targeting Holliday junctions, three strand junction, R-loops, and D-loops [53,54]. This protein also possesses a Sclafen domain which binds to DNA and is involved in various functions such as DNA metabolism, DNA repair, and protecting cell from foreign elements [55]. Q8RF29 and Q8REK7 are the transcriptional repressor DNA binding winged-helix-turn-helix proteins belonging to Rrf2 transcriptional regulator family. Transcriptional repressor proteins contain [2Fe-2S]⁺ cluster which can repress the expression of the gene encoding for the Fe-S cluster assembly protein [56]. RseC/MucC is a transcriptional regulator localized in the inner membrane of the cell. Q8RGG0 is identified as RseC/MucC protein which positively regulates the sigma (RpoE) transcription factor. RpoE plays important role in regulating the gene expression of proteins having extracellular functions [57].

Q8REE9 is a FtsL/DivC protein involved in bacterial cell division. FtsL is a short protein which forms complex with 11 other proteins engaged in the synthesis of peptidoglycan wall [58].

**Binding proteins**

We have identified five proteins (Q8RDP1, Q8RF86, Q8RGQ9, Q8RG53, and Q8RF83) involved in DNA, RNA, and protein binding. Q8RDP1 has a β-propeller domain that belongs to Kelch-repeat superfamily. Based on the function of Kelch-repeat, this protein is involved in a range of functions such as transcriptional and cytoskeletal regulation, signal trafficking and can also act as a substrate adapter for E3 ubiquitin ligase [59]. Q8RF86 was identified as a DNA helicase that catalyzes the separation of double stranded DNA when bound to a specific sequence in an ATP-dependent process [60]. Q8RGQ9 protein catalyzes ATP-dependent phosphorylation of 4-carbon acid sugars and nucleotides by binding them through N & C terminal domains respectively [61]. Q8RG53 and Q8RF83 contain a tettracopeptide repeat and it has been identified that proteins with this repeat are involved in virulence related functions [6,62].

**Transport proteins**

Q8RGCC0, Q8RHR3, and Q8RHI2 were predicted as transport proteins. Q8RGCC0 is a PelG protein involved in transport of polysaccharides outside the cell for the synthesis of the biofilm. Formation of biofilm causes resistance to anti-microbial treatment and host defense mechanism thus increasing the survival chance of bacteria [63]. Q8RGCC0 can also serve as a therapeutic target against *F. nucleatum* infection. Q8RHR3 contains a HEAT/Armadillo repeat which is present in nuclear protein transport complex [64]. Q8RHI2 belongs to an ABC transporter family that transports nickel and cobalt. These transition metals act as cofactor for prokaryotic enzymes which are involved in various metabolic processes [65].

**Membrane proteins**

Q8RE69 and Q8RGB9 are the only proteins that were identified as membrane proteins. Q8RE69 possesses a glycine zipper domain responsible for right-handed helix packing in the structure of membrane protein [66] whereas Q8RGB9 was predicted as a PagP β-barrel outer membrane protein of gram-negative bacteria. OMPLA and PagP are the two β-barrel protein enzymes involved in bacterial lipid metabolism. PagP-mediated lipid metabolism promotes infection by providing the resistance to antimicrobial peptides [67].

**Phage-related protein**

Q8REC4 and Q8REB2 proteins were recognized as the phage resistance proteins. These proteins may form a part of abortive infection system which is involved in degradation of phage mRNA. This degradation of phage mRNA is brought about by halting the synthesis of phage proteins or by activation of the intracellular sensors that activate other proteins of further pathways for the abolishment of phage infection [68,69].

**Other proteins**

Q8REQ3 protein of *F. nucleatum* belongs to SatD (secretion and acid tolerance) family having a role in acid resistance. This acid resistance may act as a virulence factor for survival of cariogenic bacteria [70,71]. Proteins Q8R6K0, Q8RGM7, and Q8RIP2 were predicted to be adhesion protein FadA, which is known to be involved in attachment and invasion of the host cells [72].

Q8RER4 was found to be involved in biosynthesis of colicin V, a bacteriocin, secreted by some bacteria for the intake of essential nutrients by inhibiting the growth of related strains [73]. Q8RECG6 belongs to the CRISPR-Cas’s system involved in defense mechanism of prokaryotes against foreign substances [49]. Q8RHE9
protein was identified as a PilN, subunit of type IV pili which is well known for cell attachment, biofilm formation, twitching motility, and pathogenesis. PilN subunit upon binding with PilM causes a structural change in PilM and induces the type IV pilus system function [74].

String analysis
String database contains the information on the functional and physical partners of the protein in a cell. Out of the 40 proteins searched on the string database, we got 30 with the confidence score > 1. Out of these 30 proteins, 9 had the confidence score ≥ 2.5 and had the maximum interacting partners.

Q8RDP1, Q8REC7, Q8RER4, Q8RFU1, Q8RG0C, Q8RHE9, Q8RHHQ2, and Q8RE80 showed the maximum interacting partners with Q8RHHQ2 having 29 interactions (Supplementary Table 6).

Homology detection with the human proteins
Using protein BLAST to search for the homologous proteins in humans (Homo sapiens [taxid:9606]) revealed that out of the 36 annotated uncharacterized proteins only 1 (Q8RHHQ2) has homology with any human protein. Rest 45 proteins have no similarity and can be used as probable drug target. Interestingly, on searching the DrugBank database for target sequences, out of all annotated proteins, only one protein which had similarity to F. nucleatum protein Q8RHHQ2 showed interaction with several drugs. No other protein was listed in the DrugBank database.

Structure prediction and modeling
Out of the 25 proteins (Table 3) for which we could find templates for homology modeling, we analyzed the structure of few important proteins. Three proteins Q8RG07, Q8RHHQ2, and Q8RII7 shared maximum homology with their respective templates (93%, 49%, and 50%, respectively). Another six functionally important proteins based on their annotation were also modeled.

Q8RG07 shares about 93% identity with the template protein and the model was prepared with 99.75% confidence. The structure showed two large α-helices running antiparallel to each other joined by a loop. FadA exists in two forms in F. nucleatum; as 129 amino acid long non-secreted pre-FadA and 111 amino acid long mature secreted FadA. Ramachandran analysis shows all residues of the model in favorable region (Fig. 4A).

Q8RHHQ2 was identified as macrodomain containing protein with a function in ADP ribosylation. The structural modeling with Phyre2 returned the template as 4IQY which records the structure of a human protein-proximal ADP-ribosyl-hydrolase MacroD2 [75]. Q8RHHQ2 shares 49% sequence identity with the template protein. The final structure model comprised 172 amino acids containing 5 α-helices and 6 β-sheets. Ramachandran plot analysis showed 98.7% residues in allowed regions (Fig. 4B).

Q8RII7 is annotated as YmdB-like protein belonging to calcineurin-like phosphatase/phosphodiesterase family which helps in biofilm production. Phyre2 search identified crystal structure of Bacillus subtilis YmdB (4B2O) [39] as the template for the modeling with 50% identity to the query sequence. Similar to the template structure, the Q8RII7 model contained the conserved αββα architecture which is crucial for the co-ordination of two divalent metal ions. PROCHECK analysis showed 99.6% residues in allowed regions and only 0.4% in disallowed region of Ramachandran plot (Fig. 4C).

Q8RFU1 was modeled based on template with PDB id 4GGM (LpxI from Caulobacter crescentus) [33]. The prediction model was seen as a dimer. Q8RG08 structure was obtained using the structure of ArgZ (PDB id: 6JUY), an arginine dehydrodase en-

---

**Table 3. List of protein structures from Protein Data Bank used as templates for homology modeling of the uncharacterized proteins of Fusobacterium nucleatum**

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Accession Id</th>
<th>Template</th>
<th>Sequence identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Q8RDP1</td>
<td>5YY8</td>
<td>18.9</td>
</tr>
<tr>
<td>2</td>
<td>Q8RE69</td>
<td>7MUS</td>
<td>33.9</td>
</tr>
<tr>
<td>3</td>
<td>Q8REC7</td>
<td>7TR9</td>
<td>17.8</td>
</tr>
<tr>
<td>4</td>
<td>Q8REM4</td>
<td>6WRY</td>
<td>15.5</td>
</tr>
<tr>
<td>5</td>
<td>Q8RF86</td>
<td>3LMM</td>
<td>14.1</td>
</tr>
<tr>
<td>6</td>
<td>Q8RFD4</td>
<td>4HV0</td>
<td>19.1</td>
</tr>
<tr>
<td>7</td>
<td>Q8RFU1</td>
<td>4GGM</td>
<td>26.1</td>
</tr>
<tr>
<td>8</td>
<td>Q8RG23</td>
<td>7ETR</td>
<td>26.8</td>
</tr>
<tr>
<td>9</td>
<td>Q8RG07</td>
<td>3ETW</td>
<td>97.3</td>
</tr>
<tr>
<td>10</td>
<td>Q8RG08</td>
<td>6JUY</td>
<td>17.2</td>
</tr>
<tr>
<td>11</td>
<td>Q8RG09</td>
<td>4XFM</td>
<td>22.4</td>
</tr>
<tr>
<td>12</td>
<td>Q8RH78</td>
<td>2X1C</td>
<td>17.1</td>
</tr>
<tr>
<td>13</td>
<td>Q8RHHQ2</td>
<td>4IQY</td>
<td>49.7</td>
</tr>
<tr>
<td>14</td>
<td>Q8RHR3</td>
<td>4RV1</td>
<td>16.2</td>
</tr>
<tr>
<td>15</td>
<td>Q8RHHQ2</td>
<td>2EFV</td>
<td>19.2</td>
</tr>
<tr>
<td>16</td>
<td>Q8RAD9</td>
<td>3LMM</td>
<td>19.3</td>
</tr>
<tr>
<td>17</td>
<td>Q8RII7</td>
<td>4B2O</td>
<td>50.4</td>
</tr>
<tr>
<td>18</td>
<td>Q8RERK7</td>
<td>7BOC</td>
<td>19.1</td>
</tr>
<tr>
<td>19</td>
<td>Q8RP2</td>
<td>3ETW</td>
<td>27.9</td>
</tr>
<tr>
<td>20</td>
<td>Q8RER1</td>
<td>5MJ1</td>
<td>19.5</td>
</tr>
<tr>
<td>21</td>
<td>Q8RER4</td>
<td>2G0I</td>
<td>20.9</td>
</tr>
<tr>
<td>22</td>
<td>Q8RH12</td>
<td>4HZU</td>
<td>22.8</td>
</tr>
<tr>
<td>23</td>
<td>Q8RG00</td>
<td>6RAJ</td>
<td>20.3</td>
</tr>
<tr>
<td>24</td>
<td>Q8RF29</td>
<td>7BOC</td>
<td>22.4</td>
</tr>
<tr>
<td>25</td>
<td>Q8R6K0</td>
<td>3ETW</td>
<td>19.6</td>
</tr>
</tbody>
</table>
Fig. 4. Structure modeling. Homology structure modeling and Ramachandran analysis of 3D structure using the Swiss-Model and PROCHECK were done for Q8RGM7 (A), Q8RHQ2 (B), and Q8RII7 (C).
zyme found in the ornithine-ammonia cycle in cyanobacteria [76]. Q8RGP8 also takes the 5-fold-pseudosymmetric structural arrangement which is the signature motif of guanidino-group modifying enzyme superfamily. Our annotation has identified Q8RGP8 as arginine deaminase enzyme. Q8RHS6 homology model was prepared using the template PDB 2EWF containing the structure of the larger subunit N.BspD6I of restriction endonuclease of Bacillus sp. [77]. Q8REI4 model was based on the PDB 2G0I showing the crystal structure of protein SMU.848 having unknown function from Streptococcus mutans [78]. Q8RFD4 has been annotated as a regulatory protein and its structure was predicted using a homologous structure of AvtR, a novel transcriptional regulator from a hyperthermophilic archael Lipothrixivirus (PDB 4HV0) [79]. Q8RG23 was modelled on the structure of ParE SO Cop-A SO toxin-antitoxin system (PDB 7ETR) of Shewanella oneidensis [80] (Supplementary Fig. 1A–1F).

Virulence prediction

Q8RGP8 and Q8REQ3 were the common proteins listed as virulence factors by both VICMpred and VirulentPred software. Q8RGP8 is annotated as Arginine deiminase which is important for defence mechanism of the bacteria. Arginine deiminase increased expression has been found to be associated with the antibiotic resistance in Streptococcus bacteria by regulating the pH [81]. It has also been found to be important for survival of bacteria causing oral and dental infections [82,83]. String analysis of these two proteins revealed their interacting partners. Q8RGP8 interacts with three other neighboring proteins namely Q8RGQ3 (FN0235) which is an ABC transporter ATP binding protein, Q8RGQ0 (FN0236), an ATP transporter substrate binding protein and finally Q8RGP9 (FN0237), an ABC transporter permease protein (Fig. 5A). Arginine is known to be transported through a specialized ATP dependent transport system in Escherichia coli [84]. Similar system could also be present in the F. nucleatum which can be a lucrative drug target. Q8REQ3 is identified as SatD family protein having a role in acid resistance. Acid tolerance is essential for the survival of the cariogenic bacteria which forms a biofilm on teeth [71,85]. String analysis of Q8REQ3 displayed many interacting partners with this protein. Q8RE18 (FN1313) is one such protein which is an oligopeptide binding protein oppA which is involved in quorum sensing (Fig. 5B). Social behavior of the bacteria in biofilms is found to be regulated by quorum sensing proteins and thus

![Image](https://doi.org/10.5808/gi.22065)
the interaction of Q8REQ3 with the oppA protein is of much significance [86]. The homology model of Q8RGP8 has already been discussed in section above.

The functional annotation of non-characterized genes using recent software and programs provides new insight into the probable functions of previously uncharacterized proteins. *F. nucleatum* strain ATCC 25586 genome has 398 uncharacterized proteins and we annotated 46 out of them. Physico-chemical parameter determination was done for all the uncharacterized proteins leading to prediction of their pI and approximate molecular weight. About 43% of the proteins were shown to be acidic with pH of less than 7 while 53% have pH of more than 7. Subcellular location of the proteins is an important determinant of its function, and it was predicted using the Cello, SignalP 5.0, and TMHMM servers. The majority of the proteins (75%) were seen to be localized in cytoplasm while 12% were outer membrane proteins. Searching for conserved sequential and structural features (domains and motifs) using the combined results obtained from InterProScan, Motif, SMART, HMMER, and NCBI-CDART programs identified 17 proteins as enzymes, 10 have regulatory roles, 5 as binding proteins and final 7 have other functions as transport, membrane proteins, etc. Another seven proteins were also annotated based on the result of at least three of these programs. ROC analysis of the software and programs used for annotation show a reliable confidence on the approach. Structural modeling was performed for 9 proteins for which suitable templates with good homology were obtained using the Swiss-Model and Phyre2 servers. Two probable virulence-related proteins (Q8RGP8 and Q8REQ3) were also identified which provide excellent opportunity for their further detailed analysis as potent drug targets. This study has resulted in the identification of many interesting proteins which were previously mentioned as uncharacterized and can provide deep understanding of the biology of the pathogen with related experimental study.

**ORCID**

Kanchan Rauthan: https://orcid.org/0000-0001-8600-7972
Saranya Joshi: https://orcid.org/0000-0002-6591-3914
Lokesh Kumar: https://orcid.org/0000-0002-5614-0331
Divya Goel: https://orcid.org/0000-0002-8374-9222
Sudhir Kumar: https://orcid.org/0000-0001-9673-5842

**Authors’ Contribution**


**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**

9. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR,


Rauthan K et al. • Assignment of function to uncharacterized proteins

Genomics & Informatics


Bioinformatic analyses reveal the prognostic significance and potential role of ankyrin 3 (ANK3) in kidney renal clear cell carcinoma

Keerakarn Somsuan¹,², Siripat Aluksanasuwan¹,²*¹

¹School of Medicine, Mae Fah Luang University, Chiang Rai 57100, Thailand
²Cancer and Immunology Research Unit (CIRU), Mae Fah Luang University, Chiang Rai 57100, Thailand

Kidney renal clear cell carcinoma (KIRC) is one of the most aggressive cancer type of the urinary system. Metastatic KIRC patients have poor prognosis and limited therapeutic options. Ankyrin 3 (ANK3) is a scaffold protein that plays important roles in maintaining physiological function of the kidney and its alteration is implicated in many cancers. In this study, we investigated differential expression of ANK3 in KIRC using GEPIA2, UALCAN, and HPA databases. Survival analysis was performed by GEPIA2, Kaplan-Meier plotter, and OS-kirc databases. Genetic alterations of ANK3 in KIRC were assessed using cBioPortal database. Interaction network and functional enrichment analyses of ANK3-correlated genes in KIRC were performed using GeneMANIA and Shiny GO, respectively. Finally, the TIMER2.0 database was used to assess correlation between ANK3 expression and immune infiltration in KIRC. We found that ANK3 expression was significantly decreased in KIRC compared to normal tissues. The KIRC patients with low ANK3 expression had poorer survival outcomes than those with high ANK3 expression. ANK3 mutations were found in 2.4% of KIRC patients and were frequently co-mutated with several genes with a prognostic significance. ANK3-correlated genes were significantly enriched in various biological processes, mainly involved in peroxisome proliferator-activated receptor (PPAR) signaling pathway, in which positive correlations of ANK3 with PPARA and PPARG expressions were confirmed. Expression of ANK3 in KIRC was significantly correlated with infiltration level of B cell, CD8+ T cell, macrophage, and neutrophil. These findings suggested that ANK3 could serve as a prognostic biomarker and promising therapeutic target for KIRC.

Keywords: ankyrin 3 (ANK3), biomarker, kidney renal cell carcinoma, peroxisome proliferator-activated receptor, prognosis

Introduction

Kidney renal clear cell carcinoma (KIRC) or clear cell renal cell carcinoma is the most common histological type of kidney cancer, accounting for 70% of all cases [1]. According to global cancer statistics, there were 431,288 new cases and 179,368 new deaths for kidney cancer worldwide in 2020 [2]. Although localized KIRC can be cured by surgical treatment, patients frequently present with metastasis at diagnosis or develop recurrence after treatment, resulting in high mortality rate and limited therapeutic options [3,4]. Therefore, the identification of biomarkers is beneficial to improve diagnosis and prognosis for KIRC patients.
Ankyrin 3 (ANK3), also known as ankyrin G, belongs to the ankyrin protein family. It is a scaffold protein that regulates the organization of membrane and cytoskeletal components [5]. ANK3 is the most abundant ankyrin in kidney [6] and plays a crucial role in membrane assembly, epithelial cell polarization, and regulation of ion channels [7-9]. In cancers, genetic and expression alterations of ANK3 have been reported in several studies [10-16]. Decreased ANK3 expression was associated with poor survival outcome in prostate cancer [14] and androgen receptor-positive breast cancer [15]. It has been shown that ANK3 regulates cell cycle and inhibits cell invasion in prostate cancer cells [14]. Overexpression of ANK3 promotes cell apoptosis and suppresses epithelial-mesenchymal transition in papillary thyroid carcinoma cells [16]. These findings indicate the prognostic value and tumor suppressive function of ANK3 in cancers. Nevertheless, its prognostic significance and role in KIRC remain largely unknown.

In the present study, we performed an integrative bioinformatic analysis of molecular and clinical data from the publicly available datasets through various online databases. The expression of ANK3 and its relationship to clinicopathologic outcomes in KIRC were explored in Gene Expression Profiling Interactive Analysis 2 (GEPIA2), University of Alabama at Birmingham CANcer data analysis portal (UALCAN), and Human Protein Atlas (HPA) databases. Prognostic significance of ANK3 for KIRC was assessed by GEPIA2, Kaplan-Meier (KM) plotter, Online consensus Survival analysis for KIRC (OSKirc), and Tumor Immune Estimation Resource (TIMER) databases. ANK3 mutations and co-mutations in KIRC were analyzed by the cBioPortal database. Potential roles of ANK3 in KIRC carcinogenesis and immune infiltration were also investigated using GeneMANIA, Shiny GO, and TIMER2.0 databases.

Methods

Differential expression analysis

Differential expression of ANK3 in KIRC compared to normal tissues was explored using GEPIA2 database (http://gepia2.cancer-pku.cn/) [17] and the UALCAN database (http://ualcan.path.uab.edu) [18]. The mRNA expression of ANK3 was analyzed in KIRC (n = 533) and normal (n = 72) tissue samples in The Cancer Genome Atlas (TCGA) dataset using GEPIA2. The differential expression of ANK3 protein in KIRC (n = 110) and normal (n = 84) tissue samples was examined in the Clinical Proteomic Tumor Analysis Consortium (CPTAC) dataset using UALCAN. In addition, the protein expression levels of ANK3 in renal cancer and normal kidney tissues were explored in the HPA database (http://www.proteinatlas.org) [19,20].

Analysis of the association of ANK3 expression and clinicopathological features of KIRC patients

Relationships between ANK3 mRNA expression and clinicopathological features, including age, race, sex, cancer stage, tumor grade, and nodal metastasis status in KIRC patients were analyzed in the TCGA dataset using the UALCAN database.

Survival analysis

Survival analysis of ANK3 expression in KIRC patients was performed using various databases, including GEPIA2, KM plotter (https://kmplot.com/analysis/) [21], OSKirc (https://bioinfo.henu.edu.cn/KIRC/KIRCList.jsp) [22]. In GEPIA2, patients (n=516) were split into low- and high-expression groups based on median expression value. Survival analysis by KM plotter was conducted for 530 KIRC patients. Low- and high-ANK3 expression groups were divided using “Auto select best cut-off” option. For OSKirc, a total of 629 KIRC patients from combined data sources (TCGA, GSE22541, GSE29609, and GSE3) were subjected to survival analysis with the patients split by “upper 50%” option. The KM curves of overall survival of KIRC patients were plotted along with the log-rank p-value and hazard ratio (HR). Multivariable Cox proportional hazard regression analysis to assess an independent predictive value of ANK3 expression was performed using the TIMER database (https://cistrome.shinyapps.io/timer/) [23,24].

Genetic alteration analysis

Genetic alterations of ANK3 were explored using cBioPortal for Cancer Genomics database (https://www.cbioportal.org/) [25,26]. The ANK3 mutations and co-mutations in KIRC were analyzed in 1,496 samples in TCGA datasets (TCGA, Firehose Legacy; TCGA, Nature 2013; TCGA, PanCancer Atlas). Heatmap representing HR and the KM curve of ANK3 co-mutated gene expression for overall survival of KIRC patients were created by GEPIA2.

Interaction network and functional enrichment analyses of ANK3-correlated genes

The top 50 genes that were positively correlated with ANK3 in KIRC based on Pearson correlation coefficient, were retrieved from GEPIA2 and subjected to further analyses. Interaction networks of ANK3-correlated genes were constructed using GeneMANIA (https://genemania.org/) [27]. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed and graphically visualized using ShinyGO (version 0.76.3) (http://bioinformatics.sdstate.edu/go/) [28]. The significance threshold for the enrichment was set at the false discovery rate ≤ 0.05.
Analysis of relationship between ANK3 and peroxisome proliferator-activated receptors in KIRC

Correlations of ANK3 expression and peroxisome proliferator-activated receptors (PPARs), including PPARα (PPARA), PPARβ/δ (PPARδ), and PPARγ (PPARG) expression in KIRC were analyzed in 516 samples using GEPIA2. KM curves of PPARα, PPARδ, and PPARG expressions for overall survival of KIRC patients were generated by GEPIA2. The protein expression levels of PPARα and PPARG in renal cancer and normal kidney tissues were explored in the HPA database.

Immune infiltration analysis

Correlation between ANK3 expression and abundance of tumor-infiltrating immune cells, including B cells, CD8+ T cells, CD4+ T cells, neutrophils, macrophages, and dendritic cells in KIRC was estimated by TIMER, TIDE, CIBERSORT, CIBERSORT-ABS, QUANTISEQ, XCELL, MCPCounter, and EPIC algorithms through TIMER 2.0 database (http://timer.cistrome.org/) [29] with tumor purity adjustment. A heatmap representing the partial Spearman's correlation coefficient was plotted using GraphPad Prism version 8.0.1 (GraphPad Software, San Diego, CA, USA). Scatter plots of ANK3 expression level and infiltration level of immune cells were visualized by the TIMER2.0 database.

Statistical analysis

Differential expression analysis of ANK3 was performed using one-way ANOVA in the GEPIA database and Student's t-test in the UALCAN database. Survival analysis was performed with the Kaplan-Meier method and log-rank test. Multivariate analysis was conducted by Cox's proportional hazard model. Genetic alterations were analyzed by one-sided Fisher’s exact test in the cBioportal database. Pearson’s correlation analysis was used to evaluate the correlation between two genes expression. Correlation between ANK3 expression and immune infiltration level was evaluated by the purity-adjusted partial Spearman's correlation test. The p-value less than 0.05 was considered statistically significant.

Results

Differential expression of ANK3 in KIRC and normal tissues

Differential expressions of ANK3 in KIRC compared to normal tissue at mRNA and protein levels were investigated using GEPIA2 and UALCAN, respectively. GEPIA2 analysis showed that ANK3 mRNA expression was significantly down-regulated in KIRC compared to normal tissues in the TCGA dataset (Fig. 1A).

Similarly, a significant decrease of ANK3 protein expression in KIRC was observed from CPTAC dataset in UALCAN (Fig. 1B). Moreover, result from HPA database also demonstrated a decrease of ANK3 protein level in renal cancer compared to normal kidney tissues (Fig. 1C). These findings indicated that ANK3 expression was significantly decreased in KIRC compared to normal tissues at both mRNA and protein levels.

Association between ANK3 expression and clinicopathological features of KIRC patients

We assessed the associations between ANK3 mRNA expression and clinicopathological features of KIRC patients using UALCAN. Based on the TCGA dataset, ANK3 expression was not significantly associated with patient’s age and race (Fig. 2A and 2B). Male patients had a significantly lower level of ANK3 expression compared to female patients (Fig. 2C). In addition, the data showed that ANK3 expression was significantly correlated with cancer stage (Fig. 2D), tumor grade (Fig. 2E), and nodal metastasis status (Fig. 2F). These findings suggested that the decreased expression of ANK3 may be a predictive indicator for KIRC severity and progression.

Prognostic impact of ANK3 expression in KIRC

We analyzed an association between ANK3 expression and overall survival of KIRC patients with low- and high-ANK3 expression using GEPIA2, KM plotter, and OSkirc databases. The data from GEPIA2 showed that KIRC patients with low ANK3 expression had significantly shorter overall survival than those with high ANK3 expression (Fig. 3A). Significant associations of ANK3 expression with overall survival of KIRC patients were consistently observed in KM plotter (Fig. 3B) and OSkirc (Fig. 3C) databases. In order to assess an independent predictive value of ANK3 expression, multivariate analysis was performed using the TIMER database. The analysis results confirmed that ANK3 expression was an independent prognostic factor for KIRC (Table 1). Thus, the low ANK3 expression could indicate poor prognosis in KIRC patients.

Genetic alteration of ANK3 in KIRC

Genetic alteration of ANK3 in KIRC patients was analyzed using cBioPortal. Based on TCGA datasets, ANK3 mutations were found in about 2.4% (36 of 1495 cases) of KIRC patients (Fig. 4A). There were 36 mutations distributed across the gene, in which missense mutations were the most frequent (28 of 36), followed by truncating (7 of 36) and splicing (1 of 36) mutations (Fig. 4B). In order to gain more insights into the underlying mo-

https://doi.org/10.5808/gi.23013
Fig. 1. Expression of ANK3 mRNA and protein in KIRC tissues. (A) Boxplot of ANK3 mRNA expression in KIRC (red) and normal (grey) tissues from The Cancer Genome Atlas dataset (GEPIA2 database). (B) Boxplot of ANK3 protein expression in KIRC (red) and normal (blue) tissues from the CPTAC dataset (UALCAN database). (C) Representative immunohistochemical image of ANK3 protein expression in normal kidney and renal cancer tissues (HPA database). ANK3, ankyrin 3; KIRC, kidney renal clear cell carcinoma; GEPIA2, Gene Expression Profiling Interactive Analysis 2; CPTAC, Clinical Proteomic Tumor Analysis Consortium; UALCAN, University of ALabama at Birmingham CANcer data analysis portal; HPA, Human Protein Atlas; RCC, renal cell carcinoma; TPM, transcripts per million. *p < 0.05, **p < 0.01.
Fig. 2. Relationship between ANK3 mRNA expression and clinicopathological features in KIRC. Boxplot of ANK3 mRNA expression in KIRC based on patient’s age (A), patient’s race (B), patient’s sex (C), individual cancer stage (D), tumor grade (E), and nodal metastasis status (F) from The Cancer Genome Atlas dataset (UALCAN database). ANK3, ankyrin 3; KIRC, kidney renal clear cell carcinoma; UALCAN, University of ALabama at Birmingham CANcer data analysis portal; TCGA, The Cancer Genome Atlas. *p < 0.05, **p < 0.01.
Fig. 3. Relationship between ANK3 mRNA expression and survival outcomes of KIRC patients. KM curves for overall survival in KIRC patients with low- and high-ANK3 expression obtained from GEPIA2 (A), KM plotter (B), and OSkirc (C) databases. ANK3, ankyrin 3; KIRC, kidney renal clear cell carcinoma; KM, Kaplan-Meier; GEPIA2, Gene Expression Profiling Interactive Analysis 2; OSkirc, Online consensus Survival analysis for KIRC; HR, hazard ratio; CI, confidence interval.
Table 1. Multivariable Cox proportional hazard regression analysis of factors affecting overall survival of KIRC patients (TIMER database)

<table>
<thead>
<tr>
<th>Variable</th>
<th>HR (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1.029 (1.015–1.044)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sex (male)</td>
<td>0.837 (0.607–1.154)</td>
<td>0.278</td>
</tr>
<tr>
<td>Race (Black)</td>
<td>2.235 (0.282–17.736)</td>
<td>0.447</td>
</tr>
<tr>
<td>Race (White)</td>
<td>2.142 (0.293–15.664)</td>
<td>0.453</td>
</tr>
<tr>
<td>Stage2</td>
<td>1.203 (0.644–2.247)</td>
<td>0.562</td>
</tr>
<tr>
<td>Stage3</td>
<td>2.172 (1.427–3.305)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Stage4</td>
<td>6.257 (4.237–9.238)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ANK3 expression</td>
<td>0.739 (0.669–0.816)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

KIRC, kidney renal clear cell carcinoma; TIMER, Tumor Immune Estimation Resource; HR, hazard ratio; CI, confidence interval; ANK3, ankyrin 3.

lecular mechanisms of cancer development, we further analyzed co-mutation pattern of ANK3 in KIRC. Genetic alterations of 124 genes were significantly identified in KIRC patients with ANK3 mutations (Fig. 4C, Supplementary Table 1). Top 10 genes with the most significantly co-mutated with ANK3 were MCRS1 (microspherule protein 1), SDAD1 (SDA1 domain containing 1), TTN (titin), NFAT5 (nuclear factor of activated T cells 5), PCLO (piccolo presynaptic cytomatrix protein), AJUBA (Ajuba LIM protein), NPASC (neurofascin), MSH6 (MutS homolog 6), HOXA9 (homeobox A9), and SLC7A6 (solute carrier family 7 member 6) (Fig. 4D). Among these genes, SDAD1, NFAT5, PCLO, AJUBA, and MSH6 had a significant prognostic impact on overall survival for KIRC (Fig. 4E and 4F). These data suggested that mutations of ANK3 and its co-mutated genes may involve cancer development and predict a high risk of poor prognosis in KIRC patients.

Interaction network, prognostic impact, and functional enrichment of ANK3-correlated genes in KIRC

In this study, we obtained the top 50 genes with the highest correlation with ANK3 in KIRC dataset from GEPIA2 for further analyses to define the possible roles of ANK3 in KIRC development and progression. A list of these genes is provided in Supplementary Table 2. The interaction network of ANK3-correlated genes was analyzed using GeneMANIA. As shown in Fig. 5A, these correlated genes closely interacted with each other in the network. The interactions among these genes were co-expression (91.98%), co-localization (3.53%), predicted (2.09%), physical interactions (1.85%), shared protein domains (0.3%), and genetic interactions (0.25%). Many of the genes in the network were significantly involved in several biological functions related to fatty acid and lipid metabolisms. Survival analysis using GEPIA2 revealed that most ANK3-correlated genes (49 of 50 genes) had a significant prognostic impact on overall survival for KIRC (Fig. 5B). In addition, functional enrichment analysis was also performed using ShinyGO. The data showed that these correlated genes were mainly enriched in GO biological process terms, such as “fatty acid beta-oxidation”, “carboxylic acid catabolic process”, and “fatty acid catabolic process” (Fig. 6A). The significantly enriched GO cellular component terms were predominantly involved with “peroxisome” and “microbody” (Fig. 6B). There was no significant enrichment of GO molecular function term in these correlated genes. For KEGG pathway enrichment analysis, ANK3-correlated genes were significantly enriched in several pathways, mainly including “PPAR signaling pathway”, “fatty acid degradation”, and “valine, leucine and isoleucine degradation” (Fig. 6C). These enriched pathways were closely connected with each other (Fig. 6D). Taken together, these findings suggested that ANK3 and its correlated genes may play a role in KIRC through PPAR signaling pathways and lipid metabolism.

Relationship between ANK3 and PPAR genes in KIRC

In order to explore a relationship between ANK3 and PPAR signaling pathway in KIRC, we employed GEPIA2 to analyze the correlation between ANK3 expression and three subfamilies of PPARs, including PPARA, PPARD, and PPARG in KIRC [30]. As shown in Fig. 7A, ANK3 expression was significantly positively correlated with PPARA and PPARG expressions, but not correlated with PPARD expression in KIRC. Furthermore, survival analysis using GEPIA2 showed that KIRC patients with low expressions of PPARA and PPARG had significantly shorter overall survival compared to high-expression groups. There was no significant association between PPARD expression and overall survival in KIRC patients (Fig. 7B). The results from the HPA database confirmed a decrease of PPARA and PPARG protein in renal cancer compared to normal kidney tissues (Fig. 7C). These results demonstrated a possible relationship of ANK3 to PPARa and PPARy signaling pathways in KIRC pathogenesis and prognosis.

Correlation between ANK3 expression and immune cell infiltration in KIRC

Because PPAR signaling pathway does not only involve energy homeostasis, but also plays a crucial role in regulating immune function and response in cancers [31,32]. Therefore, we further investigated the correlation between ANK3 and immune cell infiltration in KIRC using TIMER2.0. As shown in Fig. 8, ANK3 expression was consistently and significantly correlated with B cells, macrophages, neutrophils, and CD8+ T cells in KIRC. These findings
Fig. 4. ANK3 mutations and co-mutations in KIRC. (A) Genetic alteration frequency of ANK3 in KIRC patients (cBioportal database). (B) Distribution of mutations along the ANK3 gene in KIRC (cBioportal database). (C) Volcano plot of mutated genes in KIRC patients with and without ANK3 alterations (cBioportal database). (D) Bar graph representing the alteration frequency of the top 10 genes with the most significantly co-mutated with ANK3 in KIRC patients (cBioportal database). (E) Heatmap representing HR of each ANK3 co-mutated gene for overall survival of KIRC patients (GEPIA2 database). (F) KM curves of ANK3 co-mutated genes with a significant prognostic impact on overall survival for KIRC (GEPIA2 database). ANK3, ankyrin 3; KIRC, kidney renal clear cell carcinoma; GEPIA2, Gene Expression Profiling Interactive Analysis 2; KM, Kaplan-Meier; HR, hazard ratio.
suggested that ANK3 expression was associated with abundance of tumor-infiltrating immune cells in KIRC tissue. The ANK3 expression may be related to anti-tumor immunity and therapeutic responses in KIRC.

**Discussion**

ANK3 is the major form of ankyrin which is widely expressed in all nephron segments of the kidney [6,33]. It plays an important role in maintaining structural and physiological integrities of the kidney [7-9]. Recently, several studies have demonstrated that ANK3 expression is positively associated with patient’s prognosis and exerts a tumor-suppressive function in many cancers [14-16]. Therefore, ANK3 is an interesting target for further investigations on its prognostic value and role in KIRC.

In this study, our data analyses demonstrated that ANK3 mRNA and protein expression levels were significantly decreased in KIRC compared to normal tissues. Decreased ANK3 expression was positively correlated with disease stage and progression. The patients with low ANK3 expression had poor survival outcomes. These findings indicated that ANK3 expression had a favorable prognostic impact on KIRC. Our results are in line with previous studies in other types of cancer, including prostate and breast cancers [14,15], where low ANK3 expression was associated with poor survival outcomes. Therefore, ANK3 expression could serve as a predictive indicator for progression and prognosis in KIRC patients.

A number of genetic alterations have been described in KIRC. The von Hippel-Lindau (VHL) mutation is considered as an initi-
Fig. 6. Functional enrichment of ANK3-correlated genes in KIRC. Bar graphs representing significantly enriched GO biological process terms (A) and GO cellular component terms (B) of ANK3-correlated genes in KIRC (ShinyGO database). (C, D) Dot plot and network of significantly enriched Kyoto Encyclopedia of Genes and Genomes pathway of ANK3-correlated genes in KIRC (ShinyGO database). ANK3, ankyrin 3; KIRC, kidney renal clear cell carcinoma; GO, gene ontology; FDR, false discovery rate.

ating factor for KIRC development [3]. Mutations in several genes, including polybromo 1 (PBRM1), SET domain containing 2 (SETD2), and BRCA-associated protein 1 (BAP1), are frequently identified and closely associated with the prognosis of KIRC [34]. To our knowledge, ANK3 mutation and its functional impact on KIRC have not been previously reported. In this study, mutations were distributed throughout the ANK3 gene. Among these, missense and frameshift mutations were found in ZU5 domain. It has been shown that the ZU5 domain of ANK3 serves as a binding site for β-spectrin to organize membrane components [35] and also plays a role in regulation of apoptosis [35,36]. These findings implied that ANK3 mutations may affect its function in kidney homeostasis and carcinogenesis. ANK3 mutations were found in only a small number of KIRC patients (2.4%), suggesting that such mutations might not directly influence ANK3 expression level. Thus, epigenetic mechanisms could play a role in regulation of ANK3 expression in KIRC and merit further studies. Because ANK3 mutations occur at a very low frequency in KIRC patients, they may not have a direct association with prognosis. However, patients with ANK3 mutations frequently carry additional mutations in several genes with strong favorable prognostic impact on overall survival for KIRC. Among these ANK3-comutated genes, mutations and loss of expression of MSH6 have been reported and thought to be related to KIRC development [37,38]. In addition, MSH6 has been identified as a predisposition gene in early-onset colorectal cancer and sporadic triple-negative breast cancer [39,40]. On the basis of these findings, it was suggested that mutations of ANK3 and its co-mutated genes might involve cancer development and predict a high risk of a poor prognosis. However, further investigations are needed to confirm their clinical relevance in KIRC patients.

Previous studies have demonstrated that ANK3 regulates vari-
Fig. 7. Correlation of PPAR expression with ANK3 expression and patient’s overall survival in KIRC. (A) Scatter plots representing correlation of ANK3 expression with PPARA, PPARD, and PPARG expressions in KIRC (GEPIA2 database). (B) KM curves for overall survival in KIRC patients with low- and high-expression of PPARA, PPARD, and PPARG (GEPIA2 database). (C) Representative immunohistochemical image of PPARA and PPARG protein expressions in normal kidney and renal cancer tissues (Human Protein Atlas database). PPAR, peroxisome proliferator-activated receptor; ANK3, ankyrin 3; KIRC, kidney renal clear cell carcinoma; GEPIA2, Gene Expression Profiling Interactive Analysis 2; KM, Kaplan–Meier; TPM, transcripts per million; HR, hazard ratio.
Fig. 8. Correlation between ANK3 expression and immune cell infiltration level in KIRC. (A) A heatmap representing the partial Spearman's correlation coefficient for correlation of ANK3 expression with infiltration level of immune cells, estimated by different algorithms in the TIMER2.0 database with tumor purity adjustment. (B) Scatter plots representing significant correlation of ANK3 expression and infiltration level of immune cells (TIMER2.0 database). ANK3, ankyrin 3; KIRC, kidney renal clear cell carcinoma; TIMER, Tumor Immune Estimation Resource; TPM, transcripts per million.
rious cellular processes in cancer cells, including cell cycle, apoptosis, and invasion [14-16]. Its tumor suppressive mechanisms are related to modulation of androgen receptor signaling pathway [14,15] and suppression of epithelial-mesenchymal transition process [16]. However, the precise role of ANK3 and its mechanisms in KIRC carcinogenesis remains largely unknown. In this study, functional enrichment analyses of ANK3-correlated genes revealed potential involvement of ANK3 in PPAR signaling pathways and lipid metabolism in KIRC. PPARs are nuclear receptor transcription factors which are classified into three main subfamilies: PPARα, PPARδ, and PPARγ. They play a major role in regulation of lipid metabolism and energy homeostasis [30]. Several lines of evidence have indicated that PPARs have a strong implication in cancers and have been recognized as promising therapeutic targets [31]. PPARα and PPARγ are widely considered to exert tumor suppressive function, whereas PPARβ/δ seems to play oncogenic role in many types of cancers [31]. In KIRC, PPARα and PPARγ have been shown to regulate tumor growth and metastasis via modulations of lipid and other metabolic pathways [41-44], but there was no study reporting the role PPARβ/δ in this cancer. In concordance with these findings, our analyses revealed that ANK3 expression was positively correlated with PPARα and PPARγ, but not PPARδ expression in KIRC. Low expressions of PPARα and PPARγ were associated with poor prognosis in the patients. Decreased protein levels of PPARα and PPARγ were also confirmed in renal cancer tissue. Taken together, our results implied that PPARα and PPARγ play a more prominent role than PPARβ/δ in carcinogenesis and prognosis of KIRC. A previous study has reported that C-terminal region of ANK3 protein binds to sterol regulatory element-binding protein (SREBP), which is a transcriptional factor involved in regulation of fatty acid metabolism [45]. SREBP can activate PPARγ through stimulating the production of its endogenous ligand [46]. These findings suggested that ANK3 might exert its tumor suppressive role in KIRC through modulation of PPARα and PPARγ pathways.

Emerging evidence indicates that PPARs play a crucial role in regulation of immune cell function and response [32]. Cancer progression, prognosis, and treatment outcomes of the patient with KIRC are strongly influenced by immune cells in tumor microenvironment [47-52]. Therefore, it was hypothesized that ANK3 expression might be related to tumor-infiltrating immune cells in KIRC. In our study, we found a significant correlation between ANK3 expression and infiltration level of various immune cells, including B cell, macrophage, neutrophils, and CD8+ T cell. However, there was a weak to moderate correlation observed from our analysis. Therefore, further experimental validations should be conducted to confirm the potential of ANK3 as an indicator for immune infiltrate and response in KIRC.

To our knowledge, our study is the first to show the potential role of ANK3 in prognosis and its possible relationships with PPARα/PPARγ signaling pathway and immune infiltration in KIRC. However, there are several limitations to this study that should be considered. First, we conducted bioinformatics analysis with a limited number of publicly available datasets. Although TCGA is a large and comprehensive dataset, it may not fully represent all KIRC patients. Future perspective and more independent cohort studies would help to confirm the prognostic significance of ANK3 in KIRC. Second, further in vitro/in vivo experiments are needed to address the oncogenic role of ANK3 and its mechanisms involved in regulating PPARα/PPARγ signaling pathways in KIRC. Finally, clinical relevance and underlying mechanism of ANK3 in modulating immune response in KIRC requires further investigations.

In conclusion, our findings demonstrated the prognostic significance of ANK3 and its potential involvement with PPARα/PPARγ signaling pathway and immune cell infiltration in KIRC. ANK3 could serve as a prognostic biomarker and promising therapeutic target for KIRC.

**ORCID**

Keerakarn Somsuan: https://orcid.org/0000-0002-9848-8010
Sirapat Aluksanasuwan: https://orcid.org/0000-0002-3542-5880

**Authors’ Contribution**


**Conflicts of Interest**

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

**Acknowledgments**

This work was supported by the Office of the Permanent Secretary, Ministry of Higher Education, Science, Research and Innovation (OPS MHESI), Thailand Science Research and Innovation (TSRI) and Mae Fah Luang University.
Supplementary Materials

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

References


ing multidimensional cancer genomics data. Cancer Discov
2012;2:401-404.
36. Frisch SM, Schaller M, Cieply B. Mechanisms that link the onco-
44. Sanchez DJ, Steger DJ, SkulI N, Bansal A, Simon MC. PPARgamma is dispensable for clear cell renal cell carcinoma progression. Mol Metab 2018;14:139-149.
Insight from sirtuins interactome: topological prominence and multifaceted roles of SIRT1 in modulating immunity, aging, and cancer

Nur Diyana Zulkifli, Nurulisa Zulkifle*

Department of Biomedical Sciences, Advanced Medical and Dental Institute, Universiti Sains Malaysia, Bertam, 13200 Kepala Batas, Penang, Malaysia

The mammalian sirtuin family, consisting of SIRT1–SIRT7, plays a vital role in various biological processes, including cancer, diabetes, neurodegeneration, cardiovascular disease, cellular metabolism, and cellular homeostasis maintenance. Due to their involvement in these biological processes, modulating sirtuin activity seems promising to impact immune- and aging-related diseases, as well as cancer pathways. However, more understanding is required regarding the safety and efficacy of sirtuin-targeted therapies due to the complex regulatory mechanisms that govern their activity, particularly in the context of multiple targets. In this study, the interaction landscape of the sirtuin family was analyzed using a systems biology approach. A sirtuin protein-protein interaction network was built using the Cytoscape platform and analyzed using the NetworkAnalyzer and stringApp plugins. The result revealed the sirtuin family’s association with numerous proteins that play diverse roles, suggesting a complex interplay between sirtuins and other proteins. Based on network topological and functional analysis, SIRT1 was identified as the most prominent among sirtuin family members, demonstrating that 25 of its protein partners are involved in cancer, 22 in innate immune response, and 29 in aging, with some being linked to a combination of two or more pathways. This study lays the foundation for the development of novel therapies that can target sirtuins with precision and efficacy. By illustrating the various interactions among the proteins in the sirtuin family, we have revealed the multifaceted roles of SIRT1 and provided a framework for their possible roles to be precisely understood, manipulated, and translated into therapeutics in the future.

Keywords: aging, cancer, immunity, protein interaction network, sirtuins

Introduction

Sirtuins are a highly conserved family of proteins that play critical roles in various biological processes, including immunity, aging, and cancer. In the 1990s, the discovery of the silent information regulator 2 (Sir2) gene in yeast, which was found to be required for lifespan extension in response to calorie restriction, sparked interest in sirtuins and their potential impact on human health [1]. Sirtuins are NAD+-dependent deacetylases that modulate chromatin structure and protein function through their deacetylase activity, leading to changes in gene expression and cellular behavior [2]. In humans, seven sirtuin members were identified: SIRT1, SIRT2, SIRT3, SIRT4, SIRT5, SIRT6, and SIRT7 [3],...
each with unique functions and subcellular localizations [4]. For example, SIRT1 is predominantly located in the nucleus and regulates gene expression through deacetylation of histones and transcription factors, whereas SIRT3 is primarily found in the mitochondria and controls metabolic processes through deacetylation of metabolic enzymes [5,6].

SIRT1 has been shown to play a critical role in the differentiation and function of regulatory T cells, which are key immune cells that help maintain immune tolerance and prevent autoimmune diseases [7]. Moreover, it has been demonstrated to play a vital role in extending lifespan in various model organisms, including yeast, flies, and mice [8-11], and has been implicated in regulating age-related diseases, such as neurodegenerative disorders, metabolic disorders, and cardiovascular disease [12]. In the context of cancer, the role of SIRT1 is significant, but its specific function is still a subject of debate as its effect can vary depending on the cellular environment, target proteins in specific signaling pathways, or cancer types, and can either act as a tumor suppressor or promoter [13]. While SIRT1 has been the most extensively studied member of the family, other members such as SIRT3, SIRT6, and SIRT7 have also been found to play important roles in immunity, aging, and cancer [14-16].

Due to their involvement in these biological processes, modulating sirtuin activity has the potential to impact the pathways related to immune- and aging-related diseases, as well as cancer [17]. However, despite the promise shown in preclinical studies, much remains unknown about the safety and efficacy of sirtuin-targeted therapies [18]. Although progress has been made in understanding the roles of sirtuins in disease pathogenesis, there is still much to learn about the complex regulatory mechanisms that govern their activity, particularly in the context of multiple targets and systems. Therefore, in this study, we take a systems biology approach to examine the intricate and multifaceted roles of sirtuins in modulating immunity, aging, and cancer. Specifically, we use a protein-protein interaction (PPI) network analysis to gain insights into the sirtuins’ interaction landscape and the topological prominence of SIRT1, the most extensively studied sirtuin, and its interactions with other key proteins in these biological processes. We aim to provide a comprehensive understanding of the mechanisms by which SIRT1 influences immune response, aging, and cancer development. The systems-level analysis also seeks to shed light on the functional diversity of sirtuins and their potential as therapeutic targets for immune- and aging-related diseases as well as cancer. By revealing the intricate and complex interplay between sirtuins and other proteins in these processes, our study provides a foundation for the development of novel therapies that can target sirtuins with precision and efficacy.

### Methods

**Data retrieval and network construction**

Sirtuins PPI network was constructed using Cytoscape ver3.6.0 [19] and the UniProt accession numbers of all sirtuins as queries. The sirtuins direct interaction partners were extracted from experimentally derived protein interaction databases IntAct, MINT, and IMEx consortium [20-22] using the Cytoscape plugin PSICQUIC web service client [23]. To obtain a reliable network, all nodes in the integrated PPI network were filtered to only include *Homo sapiens* species, and redundant nodes and edges, as well as self-interactions were removed. Interconnectivity among proteins in the sirtuins PPI network was identified by submitting the list of proteins as queries on stringApp plugin [24] with the confidence score cutoff set at 0.7 and the maximum number of additional interactors set at zero.

**Annotating aging-, immune-, and cancer-associated genes on sirtuins PPI network**

In order to identify proteins associated with immune response, aging and cancer from the sirtuin PPI network, a list of genes related to each condition along with their detailed annotations was obtained from InnateDB, HAGR’s GeneAge and COSMIC Cancer Gene Census database ver90 [25-27]. The list was then imported into the node data column of sirtuin PPI network from a tabular file.

**Network analysis**

The network’s hub and bottleneck nodes were identified by analyzing the degree and betweenness centrality (BC) value of each node using NetworkAnalyzer [28]. Network clustering and functional enrichment analysis were performed using the stringApp plugin. For clustering, stringApp used the clusterMaker2 algorithm to run Markov clustering (MCL) [29,30] with a granularity parameter value set at 4 and functional enrichment analysis was carried out with a false discovery rate (FDR) threshold of 5%.

**Results and Discussion**

**Each sirtuin targets a specific set of protein**

Sirtuin members SIRT1, SIRT2, SIRT3, SIRT4, SIRT5, SIRT6, and SIRT7 have been implicated in various physiological and pathological processes such as genomic stability, gene expression, reprogramming, obesity, neurodegeneration, diabetes, and cancer [31]. To gain a better understanding of the interaction landscape of sirtuins and their targets, we set out to construct a PPI network...
of the sirtuin family using three resources: IntAct, MINT, and IMEx. Following network construction, data from InnateDB, GeneAge, and COSMIC were integrated into the network, resulting in a completely informed PPI network with 385 nodes and 395 edges (Fig. 1A). SIRT1 has the most interaction partners with 135 edges, followed by SIRT4, SIRT6, SIRT3, SIRT2, and the fewest, SIRT5 and SIRT7 with six and five edges, respectively. The network clearly demonstrates that each sirtuin is linked to a distinct set of proteins and rarely shares interaction partners. With the exception of SIRT7, the remaining sirtuins are nevertheless still linked indirectly through a variety of intermediate proteins.

The target specificity portrayed by each sirtuin is justified as each sirtuin serves a distinct function and takes part in a different biological and pathological event. For instance, SIRT1 is involved in metabolism and inflammation, whereas SIRT2 is involved in cell cycle and tumorigenesis. Like SIRT1, SIRT3 is also involved in metabolism while SIRT4 is involved in insulin secretion, SIRT5 in ammonia detoxification, SIRT6 in DNA repair, metabolism, and TNF secretion and lastly SIRT7 in rRNA transcription [32-35].

Sirtuins’ diverse localization at different subcellular sites is a result of evolutionary divergence dating back to Archaea and contributes to the differences in their roles. These days, SIRT1, SIRT6, and SIRT7 are mostly found in the nucleus, while SIRT3, SIRT4, and SIRT5 are mostly found in the mitochondria, whereas SIRT2 is found mainly in the cytoplasm [36]. The target specificity and distinct localization of sirtuins are explained by variations in the sequence and length of their N- and C-terminal domains, which cause variations in their 3D structure, affecting their binding interfaces and, as a result, influencing their binding partners [32].

Fig. 1. (A) Sirtuins protein–protein interaction network shows each sirtuin has an almost unique set of interaction partners. Grey nodes indicate that the genes are associated with immunity, aging and/or cancer. (B) Interconnectivity among proteins in (A): Nodes are colored according to the number of interactors, ranging from 1 (green) to 66 (red). The node size represents the value of BP; the bigger the node, the higher the value. BP, biological process.
Topological significance of SIRT1 in sirtuins PPI network

Next, the STRING database was used to analyze the interconnectivity of sirtuins PPI network, yielding 338 nodes and 1,038 edges. Topological network analysis suggested that SIRT1 is significantly important based on the node's degree and BC measurement. As illustrated in Fig. 1B, the color scale from green to yellow to orange to red denotes the number of connections for each node (range, 1 to 66); thus, the more reddish nodes indicate the highest connectivity and are considered hub nodes. TP53 is the most connected node in the network, followed by SIRT1. Meanwhile, identification of bottleneck nodes using measurement of BC is indicated by node size increment; the larger the node, the higher the BC value. SIRT1 size is observed to be on the large side, trailing behind TP53, TUFM, HSPA9, HSP90AA1, ATP5A1, and HSPA8.

Highly networked, as evidenced by a node with a large number of interactors, implying that it regulates a wide range of cellular functions via its target. Meanwhile, a high BC value indicates that this protein serves as a bridge between network clusters more frequently than others. If each network cluster represents a distinct functional unit, SIRT1 is very likely to be involved in the crosstalk of several distinct pathways. This observation is critical when considering SIRT1 as a therapeutic target because such topological position in a network suggests that perturbing SIRT1 for one pathway may have unintended consequences for other pathways. Interestingly, while SIRT1 exhibited very prominent topological characteristics, other sirtuins, in contrast, did not.

The topological importance of SIRT1 in our PPI network should be taken with a grain of salt, as it may be the result of studies that focused on SIRT1 in the past. The PPI network was built using reported interactions from the database, which raises the possibility that other sirtuins' interactions are still not well-characterized. However, there is a strong possibility that our PPI network is reliable because studies have shown that SIRT1 has more disordered segments in its N- and C-terminal regions, implying that it is capable of forming broad and specific interactions with multiple proteins [33,37]. To obtain a fair analysis from a system biology standpoint, an equivalent amount of research on the remaining sirtuins must be conducted.

Gene ontology terms and Kyoto Encyclopedia of Genes and Genomes pathways analysis

The stringApp cluster analysis extracted at least 27 clusters with at least three nodes. Cluster 1 is made up of SIRT1, SIRT2, SIRT5, SIRT6, and 40 other proteins, whereas SIRT3, SIRT4, and SIRT7 are not part of any cluster. Functional enrichment analysis was then performed to characterize Cluster 1, which consisted of SIRT1, SIRT2 SIRT5, and SIRT6 among the gene set, yielding a list of 159 statistically significant terms spanning four categories: gene ontology (GO) biological process (BP), GO molecular function (MF), GO cellular component (CC) and Kyoto Encyclopedia of Genes and Genomes pathways. Of these, the five most significant terms were nucleoplasm (CC) with FDR value $2.78 \times 10^{-24}$, chromatin binding (MF) (FDR = $6.8 \times 10^{-15}$), transcription factor binding (MF) (FDR = $6.4 \times 10^{-15}$), nuclear chromatin (CC) (FDR = $2.16 \times 10^{-10}$) and cellular response to oxygen (BP) (FDR = $3.35 \times 10^{-10}$), which covered 42, 20, 18, 19 and 11 out of the 44 proteins in the cluster, respectively. Split donut charts were drawn around the nodes to show which proteins are annotated with which of these terms (Fig. 2).

SIRT1 and SIRT6 are primarily localized in the nucleus, where they play a role in regulating gene expression through their deacetylation activity on histones and transcription factors. This includes the regulation of chromatin structure and gene transcription in response to various stimuli such as oxidative stress, DNA damage, and nutrient availability, which are all related to the cellular response to oxygen. SIRT2, although mainly localized in the cytoplasm, can transiently enter the nucleus during mitosis [38]. Hypoxia is a well-established cellular response to low oxygen levels, which can trigger a range of physiological responses, including alterations in gene expression and chromatin structure. SIRT1 is a well-known regulator of HIF1α and HIF2α that either enhances or reduces HIF-dependent genes depending on the type of cell thereby influencing the cell's hypoxic response [39]. SIRT2 also participates in the hypoxia-induced stress response by repressing the protective response and enhancing the toxicity caused by hypoxia [40].

SIRT1 association with aging-, cancer-, and immune-associated genes

Going back to Fig. 1A, the integration of the network with aging, cancer, and immunity data revealed that 45 out of 385 nodes are aging-, cancer- and immune-associated genes, with 44 of them being SIRT1 interaction partners. In Fig. 3, the network demonstrates that while some modes are involved in one of the pathways, others act as multifaceted modulators, as indicated by color combination in each node. There are 29, 22, and 26 genes that are either exclusively or promiscuously associated with aging, immune, and cancer, respectively. Among these, nine are involved solely in cancer, five are immune-related and six are linked to aging. Whereby, 15 genes are observed to modulate two pathways, and nine genes are associated with all three pathways.

The network demonstrated an association of SIRT1 with cancer, innate immune response, and aging through eight different
nodes, namely TP53, PIK3R1, PPARγ, FOXO3, mTOR, AKT1, CTNNB1, and HS9OA1. Cancer, immune response, and aging are biological processes that are inextricably linked. Firstly, the immune system plays a crucial role in recognizing and eliminating cancer cells. Immune cells such as T cells, B cells, and natural killer cells are responsible for identifying and attacking cancer cells. However, aging causes the immune system’s function to decline, making it less effective at this particular task, which is one of the reasons why cancer incidence increases with age \[41\]. Secondly, chronic inflammation, which is a hallmark of aging, can contribute to the development of cancer. Inflammatory cells produce reactive oxygen species and other damaging molecules that can damage DNA and other cellular components, leading to mutations and ultimately cancer. Thirdly, as we observed in our study, many of the same genes and pathways that regulate aging and immune response also play a role in cancer development.

The best-known cancer-related protein is p53, and its gene is thought to be the most frequently mutated gene in cancer. Several downstream targets, including p21, MDM2, GADD45, cyclin G, and Bax, which induce cell cycle arrest, are transcriptionally activated by p53, making it the cellular gatekeeper for cell growth and division \[42\]. The p53 status in cancer cells significantly impacts the immune response. Cells with p53 loss or mutation can affect the activity and recruitment of myeloid and T cells, allowing immune evasion and accelerating cancer growth \[43\]. There is also strong evidence of p53 associations in aging. According to Feng et al. (2007) \[44\], p53 response to γ-irradiation declines significantly in various tissues of aging mice due to a decrease in its protein stabilization after the stress. Furthermore, p53 also regulates AMPKβ1, TSC2, and PTEN that interact with the IGF-1–AKT–mTOR signaling pathway, which is involved in regulating cell growth and proliferation. p53 is reported to play a tissue-specific role in regulating these genes, and the loss of p53 function with age can contribute to the development of age-related diseases such as cancer \[45\].
PI3K/AKT pathway is one of the kinases that can activate p53, and PIK3R1 is a regulatory subunit of this pathway. Two important cellular processes regulated by PI3K/AKT are proliferation and apoptosis [46]. In this pathway, the phosphorylated PTK2 provides a binding site for the SH2 domain of the regulatory subunit (PIK3R1) of PIK3CA. Subsequent production of PI3,4,5-P3 provides a binding site for the PH domain of both PDK1 and AKT. After its activation by PDK1, AKT phosphorylates a large number of proteins that directly or indirectly regulate cell death, such as p53 and FOXO3. mTOR is also regulated by the PI3K/AKT pathway and is a downstream effector of AKT1. Dysregulation of the PI3K/AKT pathway can lead to immune dysfunction, such as autoimmunity and immunodeficiency, and contribute to the development of various diseases, including cancer [47].

PPARγ and SIRT1 can interact and modulate each other’s activity. PPARγ induces the expression of SIRT1, which in turn can deacetylate and activate PPARγ. Additionally, SIRT1 has been shown to inhibit the activity of PPARγ coactivator-1 alpha, which is involved in the regulation of mitochondrial biogenesis and energy metabolism [48]. The interaction between PPARγ and SIRT1 has been implicated in various physiological and pathological processes, including adipocyte differentiation, insulin sensitivity, lipid metabolism, inflammation, and cancer [49]. When cancer develops, cells’ metabolisms must be reprogrammed to cope with a depleted supply of oxygen and nutrients to support their rapid proliferation and biomass production. A progressive rise of oxidative stress and related inflammatory reaction appears to be the hallmarks of the aging process and many age-related diseases [50].

Meanwhile, mutations in the CTNNB1 gene, which encodes β-catenin, a crucial component of the Wnt signaling pathway, have been associated with early events in carcinogenesis as well as a scarcity of immune cells in the tumor microenvironment and a poor clinical response to immunotherapy [51,52]. Additionally, restoring Wnt/β-catenin signaling has shown promise as a thera-

---

**Fig. 3.** SIRT1 protein interaction partner: A, Proteins that are associated with neither aging, cancer nor immunity pathway; B, Gene involved in either aging, cancer, or immunity pathway; C, Genes involved in combination of the two pathways; D, Genes involved in all three pathways.
peutic strategy for age-related diseases such as Alzheimer’s disease by promoting enhanced synaptic plasticity, neuronal survival, and neurogenesis [53].

In a nutshell, the finding that p53, PIK3R1, PPARγ, FOXO3, mTOR, AKT1, and β-catenin interact with SIRT1 and are associated with cancer, aging, and immune response suggests that targeting one of these proteins could have effects on the others. Targeted therapy directed at a specific protein can have both beneficial and harmful effects, as it can disrupt the normal functions of these genes and their interactions in different pathways. For example, targeting p53 for cancer therapy may also inhibit the immune response, leading to increased risk of infections or other immune-related disorders. Therefore, the design of targeted therapy should consider the complex interplay between these genes and their roles in multiple cellular pathways. Further research is needed to fully understand the potential risks and benefits of targeted therapy for these proteins in cancer, aging, and immune-related diseases.

**ORCID**

Nur Dyiya Zulkifli: https://orcid.org/0000-0002-1307-6819

Nurulisa Zulkifle: https://orcid.org/0000-0003-3235-8691

**Authors’ Contribution**

Conceptualization: NZ. Data curation: NDZ, NZ. Formal analysis: NDZ, NZ. Funding acquisition: NZ. Methodology: NDZ, NZ. Writing – original draft: NDZ. Writing – review & editing: NZ.

**Conflicts of Interest**

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

**Acknowledgments**

This work was supported by Ministry of Higher Education (MOHE), Malaysia through Fundamental Research Grant Scheme (FRGS/1/2019/SKK08/USM/02/17).

**References**

SIRT1 prominence in sirtuins’ PPI network


cinology 2013;140:143-152


67. Kosgei VJ, Coelho D, Gueant-Rodriguez RM, Gueant JL. Sirt1-PPARS cross-talk in complex metabolic diseases and inherited
Comparison of digital PCR platforms using the molecular marker

Cherl-Joon Lee¹, Wonseok Shin², Minsik Song³, Seung-Shick Shin³, Yujun Park⁴, Kornsorn Srikulnath⁵, Dong Hee Kim⁶*, Kyudong Han¹,⁴,⁷,⁸,⁹**

¹Department of Bio-Convergence Engineering, Dankook University, Yongin 16890, Korea
²NGS Clinical Laboratory, Dankook University Hospital, Cheonan 31116, Korea
³OPTOLANE Inc., Seongnam 13494, Korea
⁴Department of Microbiology, College of Science & Technology, Dankook University, Cheonan 31116, Korea
⁵Animal Genomics and Bioresource Research Unit (AGB Research Unit), Faculty of Science, Kasetsart University, Bangkok 10900, Thailand
⁶Department of Anesthesiology and Pain Management, Dankook University College of Medicine, Cheonan 31116, Korea
⁷HuNBiome Co., Ltd., R&D Center, Seoul 08507, Korea
⁸Bio-Medical Engineering Research Center, Dankook University, Cheonan 31116, Korea
⁹DKU-Theragen Institute for NGS Analysis (DTiNa), Cheonan 31116, Korea

Assays of clinical diagnosis and species identification using molecular markers are performed according to a quantitative method in consideration of sensitivity, cost, speed, convenience, and specificity. However, typical polymerase chain reaction (PCR) assay is difficult to quantify and have various limitations. In addition, to perform quantitative analysis with the quantitative real-time PCR (qRT-PCR) equipment, a standard curve or normalization using reference genes is essential. Within the last a decade, previous studies have reported that the digital PCR (dPCR) assay, a third-generation PCR, can be applied in various fields by overcoming the shortcomings of typical PCR and qRT-PCR assays. We selected Stilla Naica System (Stilla Technologies), Droplet Digital PCR Technology (Bio-Rad), and Lab on an Array Digital Real-Time PCR analyzer system (OPTOLANE) for comparative analysis among the various droplet digital PCR platforms currently in use commercially. Our previous study discovered a molecular marker that can distinguish Hanwoo species (Korean native cattle) using Hanwoo-specific genomic structural variation. Here, we report the pros and cons of the operation of each dPCR platform from various perspectives using this species identification marker. In conclusion, we hope that this study will help researchers to select suitable dPCR platforms according to their purpose and resources.

Keywords: digital PCR platforms, molecular marker, point-of-care testing

Introduction

Polymerase chain reaction (PCR) is a technique that can amplify target DNA and has applied to research in various fields of biology [1-3]. This technique plays an important role in clinical diagnosis and species identification using molecular markers. It can be applied to various molecular quantification methods in consideration of sensitivity, cost, speed, convenience, and specificity. However, conventional PCR assay have a disadvantage that only qualitative analysis is possible, and quantification is difficult. Therefore, a quantitative real-time PCR (qRT-PCR) that can quantify DNA amplification in real time through fluorescence measurement has been devised [4]. qRT-PCR is used as the gold standard...
in molecular diagnostics for its high sensitivity, specificity, and accuracy, as well as in forensic biology and medicine [5,6]. However, a qRT-PCR has the following limitations. First, a standard curve should be prepared for absolute quantification as a reference gene expressed at a stable level in various samples. Second, even if a standard curve has been drawn, it must be recreated if a new device or platform used. Finally, qRT-PCR depends on quantification cycle (Cq) values, and quantification is directly affected by PCR inhibitors that distort Cq values [7,8]. Therefore, digital PCR (dPCR) was developed to compensate for these limitations. dPCR is a third-generation PCR capable of real-time absolute quantification without a standard curve [9]. dPCR divides the PCR mixture into independent reactions and expresses digital signals as either a digital signal as a positive droplet “1” or a negative droplet “0” depending on whether amplification occurred or not [10]. The generated droplets can measure the target DNA concentration by the number of copies without a standard curve through Poisson distribution. dPCR has the following advantages compared to qRT-PCR. First of all, quantification is possible without a standard curve. Next, dPCR is possible to detect low copy number targets than the detection limit of qRT-PCR. Finally, dPCR is relatively less affected by PCR inhibitors than qRT-PCR [11,12].

Currently, a variety of commercially available dPCR instruments such as Thermo Fisher Quantstudio 3D, Fluidigm BioMark ddPCR 37K, Formulatrix Constellation, JN Medsys Clarity, QX200 Droplet Digital PCR System (Bio-Rad, Hercules, CA, USA), Raindance Raindrop plus, Stilla Naica, and Lab on an Array (LOAA) are on the market. Among them, we selected the Stilla Naica System (Stilla Technologies, Villejuif, France), Droplet Digital PCR Technology (Bio-Rad), and the LOAA Digital Real-Time PCR Analyzer System (OPTOLANE, Seongnam, Korea) as targets for comparative analysis.

In a previous study, we found the molecular marker that could identify cattle breeds [13]. We investigated the pros and cons of the three dPCR platforms using the Hanwoo-specific molecular marker. This study will help researchers select an appropriate platform according to their purpose and resources in studies applying dPCR.

**Methods**

**DNA isolation and PCR**

Two hundred microliters of DNA was extracted from 9 Hanwoo and 9 Holstein blood samples (200 μL) using Exgene Blood SV mini kit (GeneAll Biotechnology, Seoul, Korea). All research protocols and animal experiments in this study were reviewed and approved by the Dankook University Institutional Animal Care and Use Committee (DKU IACUC) in South Korea (DKU-22-055). The PCR composition performed to verify the molecular marker (DEL_96) in all samples is as follows; PCR amplification was carried out using 10 ng and 20 ng of template, 1 μL of DNA, 7 μL of DW, 1 μL of each oligonucleotide primer of 200 nM, and 10 μL of BioFACT Lamp Taq DNA Polymerase (BioFACT, Daejeon, Korea) in a total volume of 20 μL (Table 1). PCR amplification was performed by following process: pre-denaturation step of 5 min at 95°C, followed by 35 cycles of denaturation step 30 s at 95°C, annealing step of 40 s at 60°C, and extension step of 1 min at 72°C, followed by a final extension step of 10 min at 72°C.

**Stilla Naica System for Crystal Digital PCR**

The droplet digital PCR (ddPCR) reaction mixture consisted of 5 μL of Naica multiplex PCR Mix Buffer A (5x, cat. No. R10052, Stilla Technologies), 1 μL of Naica multiplex PCR Mix Buffer B 4% (100%, cat. No. R10052, Stilla Technologies), 1 μL of FAM probe/primer mix 25×, 1 μL of VIC probe/primer mix 25×, 50 ng DNA, and nuclease-free water up to 25 μL. The reaction mixture was loaded onto a Sapphire Chip (Stilla Technologies), and sample partitioning and thermal cycling were performed on the Naica Geode. The ddPCR was carried out in the following steps: initial denaturation step of 3 min at 9°C, 15 s at 60°C for annealing and extension, with a release step that lowers the pressure and temperature for 33 min. Each sample produced 20,000 to 30,000 droplets. The fluorescence of the Sapphire Chip, where PCR was completed, was measured using Naica Prism 3 equipment, and the fluorescence value extracted for each droplet was analyzed using Crystal Miner software (Stilla Technologies) [14].

**Bio-Rad for QX200 Droplet Digital PCR System**

The ddPCR reaction mixture contained 10 μL of 2× ddPCR supermix for probes (No dUTP) (#186-3024, Bio-Rad), the final concentration of 250 nmol/L for each of the probes, 450 nmol/L for the forward and reverse primers, 50 ng DNA, and nuclease-free water up to 20 μL. ddPCR reaction mixture was dispensed into the middle line in the DG8 cartridge (#186-4008, Bio-Rad), and 70 μL of generation oil was dispensed into the bottom wells. The sample is split into approximately 20,000 water-oil emulsion droplets using the QX200 Droplet generator. Forty microliters of the resulting water-oil emulsion droplets were transferred to a 96-well plate sealed with PX1 PCR plate Sealer (Bio-Rad). The QX200 was carried out in the following steps: enzyme reaction step of 3 min at 50°C, initial denaturation step of 15 min at 95°C, followed by 40 cycles of denaturation 10 s at 95°C, 20 s at 60°C for annealing and extension. Droplets were analyzed using QuantaSoft ver-
Table 1. Cattle gDNA quality control and dsDNA concentration

<table>
<thead>
<tr>
<th>Sample No.</th>
<th>Slaughter No.</th>
<th>Breed</th>
<th>Target size (bp)</th>
<th>Date of slaughter</th>
<th>DNA concentration (ng/μL)</th>
<th>A260/A280</th>
<th>A260/A230</th>
<th>Spectrometer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hanwoo_1</td>
<td>10</td>
<td>Hanwoo</td>
<td>680/310</td>
<td>13 Jul 2015</td>
<td>31.7</td>
<td>1.84</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>Hanwoo_2</td>
<td>11</td>
<td>Hanwoo</td>
<td>680/310</td>
<td>6 Jul 2015</td>
<td>39.83</td>
<td>1.84</td>
<td>1.59</td>
<td></td>
</tr>
<tr>
<td>Hanwoo_3</td>
<td>13</td>
<td>Hanwoo</td>
<td>680/310</td>
<td>6 Jul 2015</td>
<td>35.47</td>
<td>1.84</td>
<td>1.65</td>
<td></td>
</tr>
<tr>
<td>Hanwoo_4</td>
<td>16</td>
<td>Hanwoo</td>
<td>680/310</td>
<td>6 Jul 2015</td>
<td>23.66</td>
<td>1.92</td>
<td>1.61</td>
<td></td>
</tr>
<tr>
<td>Hanwoo_5</td>
<td>20</td>
<td>Hanwoo</td>
<td>680/310</td>
<td>6 Jul 2015</td>
<td>39.31</td>
<td>1.87</td>
<td>1.43</td>
<td></td>
</tr>
<tr>
<td>Hanwoo_6</td>
<td>21</td>
<td>Hanwoo</td>
<td>680/310</td>
<td>6 Jul 2015</td>
<td>22.92</td>
<td>1.87</td>
<td>1.75</td>
<td></td>
</tr>
<tr>
<td>Hanwoo_7</td>
<td>243</td>
<td>Hanwoo</td>
<td>680/310</td>
<td>6 Jul 2015</td>
<td>22.92</td>
<td>1.87</td>
<td>2.09</td>
<td></td>
</tr>
<tr>
<td>Hanwoo_8</td>
<td>256</td>
<td>Hanwoo</td>
<td>680/310</td>
<td>6 Jul 2015</td>
<td>41.78</td>
<td>1.88</td>
<td>2.22</td>
<td></td>
</tr>
<tr>
<td>Holstein_1</td>
<td>264</td>
<td>Holstein</td>
<td>680 (only)</td>
<td>6 Jul 2015</td>
<td>33.25</td>
<td>1.81</td>
<td>1.83</td>
<td></td>
</tr>
<tr>
<td>Holstein_2</td>
<td>274</td>
<td>Holstein</td>
<td>680 (only)</td>
<td>6 Jul 2015</td>
<td>23.55</td>
<td>1.84</td>
<td>1.76</td>
<td></td>
</tr>
<tr>
<td>Holstein_3</td>
<td>280</td>
<td>Holstein</td>
<td>680 (only)</td>
<td>6 Jul 2015</td>
<td>19.53</td>
<td>1.86</td>
<td>1.71</td>
<td></td>
</tr>
<tr>
<td>Holstein_4</td>
<td>292</td>
<td>Holstein</td>
<td>680 (only)</td>
<td>6 Jul 2015</td>
<td>67.14</td>
<td>1.86</td>
<td>2.21</td>
<td></td>
</tr>
<tr>
<td>Holstein_5</td>
<td>296</td>
<td>Holstein</td>
<td>680 (only)</td>
<td>6 Jul 2015</td>
<td>26.86</td>
<td>1.73</td>
<td>1.11</td>
<td></td>
</tr>
<tr>
<td>Holstein_6</td>
<td>306</td>
<td>Holstein</td>
<td>680 (only)</td>
<td>6 Jul 2015</td>
<td>35.22</td>
<td>1.82</td>
<td>2.21</td>
<td></td>
</tr>
<tr>
<td>Holstein_7</td>
<td>317</td>
<td>Holstein</td>
<td>680 (only)</td>
<td>6 Jul 2015</td>
<td>19.2</td>
<td>1.67</td>
<td>1.14</td>
<td></td>
</tr>
<tr>
<td>Holstein_8</td>
<td>320</td>
<td>Holstein</td>
<td>680 (only)</td>
<td>6 Jul 2015</td>
<td>27.57</td>
<td>1.75</td>
<td>1.11</td>
<td></td>
</tr>
<tr>
<td>Holstein_9</td>
<td>332</td>
<td>Holstein</td>
<td>680 (only)</td>
<td>6 Jul 2015</td>
<td>33.57</td>
<td>1.77</td>
<td>1.04</td>
<td></td>
</tr>
</tbody>
</table>

sion 1.7.4.0917 (Bio-Rad) [15,16].

**OPTOLANE for LOAA Digital PCR System**

The reaction mixture consisted of 10 μL 3× Dr. PCR Master Mix (OPTOLANE), 10 μL Primer & Probe mix (final concentration of 20 pmol for the forward and reverse primer, 10 pmol for each of the probes), and 10 μL of DNA containing 50 ng. The reaction mixture was placed in a semiconductor-based cartridge and spread evenly on the chip using a sample loader, LOAA POSTMAN. The LOAA dPCR system was carried out in the following steps: enzyme reaction step of 3 min at 50°C, initial denaturation step of 15 min at 95°C, followed by 40 cycles of denaturation 10 s at 95°C, 20 s at 62°C for annealing and extension. The PCR-completed chip was automatically analyzed with LOAA Dr. PCR software 3.0.0 (OPTOLANE) [17,18].

**Results and Discussion**

A previous study reported that the Hanwoo-specific molecular marker was found only in the Hanwoo genome by nonallelic homologous end-joining [19]. Therefore, it can be used as a strong molecular marker to differentiate between Hanwoo and Holstein. As shown in Fig. 1, the Hanwoo strain has a polymorphic structure in the Del_96 region by a TE-association deletion event. Among various commercial dPCR equipment, we selected the Stilla Naica System (Stilla Technologies), Droplet Digital PCR Technology (Bio-Rad), and the LOAA Digital Real-Time PCR Analyzer System (OPTOLANE) as targets for comparative analysis. As shown in Table 2, Crystal Digital PCR takes about 2.5 h, which is relatively faster than QX200, because droplet formation and PCR proceed in Geode equipment. In the QX200, it took about 6 h, the longest time, from droplet formation in the Droplet Generator to amplification in the PX1 PCR Cycler to finally obtain the result. However, it has the advantage of being able to perform dPCR for more samples at once than Crystal Digital PCR and LOAA dPCR [20]. Unlike two equipments, LOAA dPCR can obtain experimental results in the shortest time as both PCR and data analysis are performed in a single equipment in a chip format. In addition, OPTOLANE’s unique technology can be applied to the chip to reduce the weight and size of the equipment, so it is highly applied to point-of-care testing. However, in the case of LOAA dPCR equipment, there is a disadvantage in that dPCR can be performed on only one sample per cartridge.

For the dPCR analysis of the three instruments, 9 Hanwoo and 9 Holstein blood samples were used as DNA templates, and three repetitions were performed with the FAM probe (Thermo Fisher Scientific, Waltham, MA, USA), VIC probe (Thermo Fisher Scientific), Cy5 probe (SFC Probes, Cheongju, Korea) customized.
Fig. 1. Polymorphic patterns of the Del_96 region in 9 Hanwoo (A) and Holstein (B) genomes. As a result of gel chromatography, heterozygous alleles at 680 bp and 310 bp were identified in the 9 Hanwoo samples, but no deleted alleles were confirmed in the Holstein sample.

Table 2. Comparison of three different dPCR platforms

<table>
<thead>
<tr>
<th></th>
<th>Stilla Crystal Digital PCR</th>
<th>Bio-Rad QX200 Droplet Digital PCR</th>
<th>OPTOLANE Lab on an Array digital PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>dPCR type</td>
<td>Droplet</td>
<td>Droplet</td>
<td>Chip</td>
</tr>
<tr>
<td>Detection mode</td>
<td>End point</td>
<td>End point</td>
<td>Real-time</td>
</tr>
<tr>
<td>Partitions</td>
<td>30,000</td>
<td>20,000</td>
<td>20,000</td>
</tr>
<tr>
<td>Duration</td>
<td>~2.5 h</td>
<td>~6 h</td>
<td>~1.5 h</td>
</tr>
<tr>
<td>Detection</td>
<td>3 colors</td>
<td>2 colors</td>
<td>2 colors</td>
</tr>
<tr>
<td>Price</td>
<td>× 1.1</td>
<td>× 1</td>
<td>× 0.2</td>
</tr>
</tbody>
</table>

dPCR, digital polymerase chain reaction.

for each manufacturer (Fig. 2, Supplementary Tables 1, 2). The Crystal Digital PCR formed 20,000 and 30,000 droplets, producing about 23,243 on average. FAM dye was detected in all Hanwoo and Holstein genomes, and VIC dye was only detected in Hanwoo samples (Fig. 3). The signal of the VIC dye was detected as an average 367.25 channel concentration (copy/µL) in the genome of Hanwoo. The QX200 can form 20,000 droplets and produce about 13,409 droplets on average. FAM dye was used as a Hanwoo-specific probe, and VIC dye was designed to fit the QX200 instrument to be detected in all genomes of Hanwoo and Holstein (Fig. 3). The signal of the VIC dye was detected as an average 367.25 channel concentration (copy/µL) in the genome of Hanwoo. The QX200 can form 20,000 droplets and produce about 13,409 droplets on average. FAM dye was used as a Hanwoo-specific probe, and VIC dye was designed to fit the QX200 instrument to be detected in all genomes of Hanwoo and Holstein (Fig. 3). The signal of the VIC dye was detected as an average 367.25 channel concentration (copy/µL) in the genome of Hanwoo. FAM dye was detected in all samples like other instruments, and FAM dye was detected only in Hanwoo samples. The FAM dye signal was detected as an average 681.73 concentration (copy/µL) in the genome of Hanwoo. As a result, Hanwoo-specific probes in all dPCR instruments showed significant detection only in Hanwoo samples, suggesting that all Hanwoo samples contained a specific deletion region (Fig. 4). However, in Holstein samples, Crystal Digital PCR, QX200, and LOAA dPCR using a Hanwoo-specific probe showed channel concentrations (copy/µL) of 0.66, 0.19, and 0.35, respectively. This result is consistent with previous studies: (1) Probes designed in the TE region can detect non-specific signals in structure variation with high similarity. (2) Abnormally high fluorescence intensity measured in dPCR analysis can be mistakenly recognized as a positive well. Nevertheless, the Hanwoo-specific probe was statistically sufficient to discriminate between Hanwoo and Holstein.

Since the outbreak of the coronavirus disease-19 pandemic, the
Fig. 2. Probe designs for each equipment. (A) The FAM dye was designed to detect all bovine genomes through digital PCR analysis, and the VIC dye was designed at the Del_96 boundary to detect only Hanwoo. (B) The VIC dye was designed to detect all cattle genomes, and the FAM dye was designed to detect only Hanwoo at the Del_96 boundary. (C) The Cy5 dye was designed to detect all cattle genomes, and the FAM dye was designed to detect only Hanwoo at the Del_96 boundary of Hanwoo.

Fig. 3. Comparison of each positive droplet rate through three repeated experiments. (A) X-axis and Y-axis show the name of each sample and the number of positive droplets formed with FAM dye and VIC dye using the Stilla Naica System. (B) X-axis and Y-axis indicate the name of each sample and the number of positive droplets formed with FAM dye and VIC dye using Bio-Rad equipment. (C) X-axis and Y-axis indicate the name of each sample and the number of positive droplets formed with FAM dye and Cy5 dye using OPTOLANE equipment.
need for equipment that complements conventional PCR, which can only perform qualitative analysis, and qRT-PCR, which is greatly affected by PCR inhibitors, has emerged. Species identification experiment results using the Stilla Naica System, Droplet Digital PCR Technology, and LOAA Digital Real-Time PCR Analyzer System among commercially available dPCRs confirmed that all equipment has the potential as a platform for species identification. In particular, compared to other dPCR equipment, the LOAA Digital Real-Time PCR Analyzer System has been made smaller and lighter with the manufacturer’s proprietary technology. dPCR has operational pros and cons depending on each platform. In addition, consumables of dPCR in common are expensive compared to qRT-PCR. However, dPCR can contribute to developing genetically modified organism testing and drug resistance research in addition to species identification with high accuracy and sensitivity. Therefore, this study is expected to help researchers select a suitable dPCR platform according to their purpose and resources.

**ORCID**

Cherl-Joon Lee: https://orcid.org/0000-0001-8282-2216  
Wonseok Shin: https://orcid.org/0000-0001-5964-1425  
Minsik Song: https://orcid.org/0000-0002-0658-5666  
Seung-Shick Shin: https://orcid.org/0000-0002-8536-2056  
Yujun Park: https://orcid.org/0000-0001-9431-0824  
Kornsorn Srikulnath: https://orcid.org/0000-0002-5985-7258  
Dong Hee Kim: https://orcid.org/0000-0002-5056-7406  
Kyudong Han: https://orcid.org/0000-0001-6791-2408

**Authors’ Contribution**

Conceptualization: KH. Data curation: CJL, WS, KH, DHK. Formal analysis: CJL, WS, MS, KS. Funding acquisition: KH. Methodology: CJL, WS, SSS, YP. Writing – original draft: CJL, WS. Writing – review & editing: CJL, WS, KH, DHK.
**Conflicts of Interest**

This study was supported by IL-YANG Pharmaceutical Co., Ltd.

**Acknowledgments**

The authors gratefully acknowledge the Bio-Medical Engineering Research Center at Dankook University. The research institute has been supported by the Department of Microbiology was supported through the Research-Focused Department Promotion & Interdisciplinary Convergence Research Project as a part of the Support Program for University Development for Dankook University in 2022.

**Supplementary Materials**

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

**References**

Introduction

Pathogenic mycobacteria are significant sources of illness in humans and animals. Despite the accessibility of antibiotics and chemotherapeutic candidates that are efficient against some mycobacteria, the emergence of drug-resistant strains necessitates the development of new active molecules and intervention strategies [1-3]. Within more than 130 years since discovering Mycobacterium tuberculosis (MTB) as the causative microorganism responsible for human tuberculosis (TB) by Robert Koch, numerous scientific advances have been made to help cope with this significant pathogen. However, despite this progress, MTB still holds many unresolved secrets. Much work remains to be done to translate the essential findings from recent research into novel strategies against the pathogen [4-7]. Tuberculosis, one of the ancient recorded human diseases, continues to be one of the leading causes of death, claiming two million lives annually. TB affects bone, the central nervous system, and many other physiological systems. However, it is essentially a pulmonary illness caused by the precipitation of aerosolized MTB onto lung alveolar surfaces. From this point, the causes of the illness are contingent on the immunological reactivity of the host to varying degrees [8-10].
MTB has an irregular, highly recurrent life cycle that encompasses a varied and heterogeneous spectrum of habitats and physiological states, many of which are unique from other infections. It is not unanticipated that MTB has conceived a specific set of metabolic capacities to facilitate its adaption to and movement across hosts, given its peculiar, if not distinctive, environment [11-13].

Numerous advancements in computational biology have made it possible to construct diverse technologies and strategies for predicting protein structure and the identification of sequence commonalities for active investigation and the analysis of active site residue relationships [14-16]. A bioinformatics examination of the proteins enables one to assess their three-dimensional architectural structure, categorize novel features, investigate specific processes to understand our biological lineage, uncover different clusters, and assign the proteins’ function. The obtained information can also convey reasonable pharmacological strategies and assets in developing promising anti-disease medications [17-19]. The hypothetical protein from MTB consists of disulfide oxidoreductase involved in the catalyzyation of dithiol oxidation and/or disulfide reduction of target sites in MTB.

Methods

Sequence retrieval

The amino acid sequence of the protein was obtained from the NCBI database in FASTA format. The physicochemical properties were determined using ProtParam (ExPASy) and SMS v.2.0 programs. Afterwards, the subcellular location of the selected protein was determined. This study also anticipated the protein family, superfamily, domain, coil, and folding pattern of the protein. The STRING program was used for protein protein interaction determination. Moreover, secondary structural documentation was performed using the SOPMA, DISOPRED (v. 3.0), and SPIPRED (v. 4.0) programs. The tertiary structure was predicted using the Modeller program with the HHpred database and validated by the PROCHECK, Verify3D, and ProSA-web tools. Furthermore, the CASTp server was used for active site determination of the selected protein present in MTB. Additionally, the antigenicity, allergenicity, and toxicity properties of the protein were determined.

Physicochemical properties

The physicochemical parameters of the protein were evaluated by the ProtParam assessment tool of the ExPASy server program [20] and the SMS v.2.0 program (https://www.bioinformatics.org/sms2/index.html, accessed on September 10, 2022).

Subcellular localization identification

The CELLO (v.2.5) [21,22], PSORTb (v3.0) [23], HMMPOT (v2.0) [24,25], and TMHMM (v2.0) [26,27] programs are used to detect the subcellular localization and protein topology analysis.

Prediction of the protein family, superfamily, domain, coil, and folding pattern

The NCBI CD tool was used to anticipate the conserved domain [28]. The GenomeNet [29], Pfam program [30], SuperFamily program [31], and ScanProsite tool [32] used for the evolutionary relationships determination.

Protein-protein interaction

The STRING program (v.11.5) [33] was used to determine the protein-protein (pr-pr) interaction.

Secondary structural assessment

The SOPMA program was used following the default parameters similarity threshold (8), window width (17), and the number of states (4) [34]. DISOPRED (v.3.0) [35] and the SPIPRED (v.4.0) [36] used for the determination of further secondary characteristics and protein topology.

Structure prediction and validation

The tertiary structure of the protein is generated by using the Modeller program [37]. The HHpred tool selected the most suitable template for protein structure anticipation [38-40]. The PROCHECK and Verify3D programs of the SAVES (v.6.0) tool were used for the structural validation of the protein [41]. Additionally, the ProSA-web program was used to calculate the Z-score and validate the modeled 3D structure of the protein [42].

Active sites determination

The CASTp program was used to determine the active sites in the protein [43].

Antigenicity, allergenicity, and toxicity

The VaxiJen (v2.0) program [44] was used to determine the protein’s antigenicity. Moreover, the AllerTOP (v. 2.0) program was used to predict the allergenicity of the protein [45]. The Toxin-Pred program [46] was used to demonstrate the toxicity of the protein.
Results and Discussion

**Sequence retrieval**
The protein's amino acid sequence was retrieved from the NCBI database in FASTA format. The protein contains 173 amino acids (Table 1).

**Physicochemical parameters determination**
By examining the properties of each amino acid in the protein, it is possible to comprehend how its physicochemical properties were characterized. The ProtParam program estimated the physicochemical characteristics. The protein comprises 173 amino acids, whereas Ala (n = 30, 17.3%) is the most abundant amino acid in the protein sequence (Table 2, Fig. 1). There is no His (H) in the protein sequence. The half-life of a protein is defined as the time required for the radio-labeled focal protein concentration to fall by 50% relative to the quantity at the beginning of the chasing [47]. The estimated half-life for the protein of about 30 h (mammalian reticulocytes, *in vitro*), >20 h (yeast, *in vivo*), and >10 h (*Escherichia coli, in vivo*). The demonstrated isoelectric point (pI), the total number of atoms, and molecular weight as of 5.19, (4.98*), 2,572, and 18,382.98 Dalton (Table 2).

Moreover, the total number of positively (Arg + Lys) and negatively charged residues (Asp + Glu) are 10 and 11 in the protein. The instability index (29.40) demonstrates protein stability, whereas the aliphatic index (86.42) denotes protein balance over a broad temperature scale. The grand average of hydropathicity (GRAVY, 0.334) determines the enhancement of thermostability [48].

**Subcellular location identification and protein topology prediction**
The computerized estimation of the subcellular location of bacterial proteins is essential for proteome categorization and for selecting novel therapeutic targets and vaccination candidates. Various subcellular localization predictors have been created in recent years, including both generic localized and feature-based predictors [49-52]. The CELLO (v.2.5) and PSORTb (v3.0) predicted the subcellular location of the protein as extracellular (Table 3). Moreover, transmembrane helix identification in integral membrane proteins is an essential bioinformatics component. In addition to predicting individual transmembrane helices, the most ef-

### Table 1. Protein retrieval

<table>
<thead>
<tr>
<th>Protein individuality</th>
<th>Protein information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus</td>
<td>OHO19689</td>
</tr>
<tr>
<td>Amino acid</td>
<td>173 aa</td>
</tr>
<tr>
<td>Definition</td>
<td>Hypothetical protein BBW91_12415 [Mycobacterium tuberculosis]</td>
</tr>
<tr>
<td>Accession</td>
<td>OHO19689</td>
</tr>
<tr>
<td>Version</td>
<td>OHO19689.1</td>
</tr>
<tr>
<td>GenBank ID</td>
<td>OHO19689.1</td>
</tr>
<tr>
<td>Source</td>
<td>Mycobacterium tuberculosis (Mycobacterium tuberculosis variant tuberculosis)</td>
</tr>
<tr>
<td>FASTA sequence</td>
<td>&gt;OHO19689.1 hypothetical protein BBW91_12415 [Mycobacterium tuberculosis]</td>
</tr>
<tr>
<td>FASTA sequence</td>
<td>MSLRLVSPIKADGIVAVAIAVLMFLANTPRAVAADELRQFTATTLSGAPFDGASLGKPAVLWFTPWCPEFCNAAPLSQAVAAANPAVTFVGIATRADVGAMQSEPVSYNLNLFTNLNADGVWARYNVFWQAPAVFVFYRADGISTFVNNPTAAOSQDELSGRVAA</td>
</tr>
</tbody>
</table>

### Table 2. Physicochemical parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular weight</td>
<td>18,382.98</td>
</tr>
<tr>
<td>Formula</td>
<td>C_{835}H_{1274}N_{218}O_{239}S_{6}</td>
</tr>
<tr>
<td>Theoretical pl</td>
<td>5.19, 4.98*</td>
</tr>
<tr>
<td>Total number of atoms</td>
<td>2,572</td>
</tr>
<tr>
<td>Total number of positively charged residues (Arg + Lys)</td>
<td>10</td>
</tr>
<tr>
<td>Total number of negatively charged residues (Asp + Glu)</td>
<td>11</td>
</tr>
<tr>
<td>The estimated half-life</td>
<td>a) 30 h (mammalian reticulocytes, <em>in vitro</em>)</td>
</tr>
<tr>
<td></td>
<td>b) &gt; 20 h (yeast, <em>in vivo</em>)</td>
</tr>
<tr>
<td></td>
<td>c) &gt; 10 h (<em>Escherichia coli, in vivo</em>)</td>
</tr>
<tr>
<td>Aliphatic index</td>
<td>86.42</td>
</tr>
<tr>
<td>Instability index (II)</td>
<td>29.40</td>
</tr>
<tr>
<td>Grand average of hydropathicity (GRAVY)</td>
<td>0.334</td>
</tr>
</tbody>
</table>

*pI calculated by the SMS v.2.0.*
**Amino Acid Composition**

Fig. 1. Amino acid composition. The protein contains Ala (30, 17.3%), Arg (7, 4.0%), Asn (10, 5.8%), Asp (8, 4.6%), Cys (2, 1.2%), Gln (6, 3.5%), Glu (3, 1.7%), Gly (10, 5.8%), Ile (5, 2.9%), Leu (13, 7.5%), Lys (3, 1.7%), Met (4, 2.3%), Phe (12, 6.9%), Pro (11, 6.4%), Ser (12, 6.9%), Thr (12, 6.9%), Trp (5, 2.9%), Tyr (3, 1.7%), and Val (17, 9.8%).

Table 3. Subcellular localization and protein topology analysis results

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELLO (v2.5)</td>
<td>Extracellular</td>
</tr>
<tr>
<td>PSORTb (v3.0)</td>
<td>Extracellular</td>
</tr>
<tr>
<td>HMMTOP (v2.0)</td>
<td>1 Transmembrane helix</td>
</tr>
<tr>
<td>TMHMM (v2.0)</td>
<td>1 Transmembrane helix</td>
</tr>
</tbody>
</table>

Effective approaches to date strive to predict the complete topology of the protein, including the entire count of transmembrane helices and their direction related to the membrane [27,53]. The TMHMM (v2.0) and HMMTOP (v2.0) programs anticipated the protein has a single transmembrane helix (at 12–30 region in the protein residue). Most membrane proteins’ transmembrane portions have helices as their secondary structures. When a membrane protein is conducting its task, whether sending messages throughout the membrane or assisting an ion channel in unlocking or closing, the transmembrane helices frequently move together [54-57].

**Prediction of the protein family, superfamily, domain, coil, and folding pattern**

The Conserved Domain Database (CDD) intends to annotate biomolecular sequences with the evolutionarily conserved protein domain placement. A repository of pre-computed domain identification is kept for NCBI’s Entrez database-tracked proteins, and real-time search facilities are provided. CDD also facilitates comparative analysis of protein families employing conserved domain architectures, and a new curation effort focuses on giving functional categorization of various subfamily structures [58-60]. The CDD tool classified the protein as protein disulfide oxidoreductase (domain architecture ID 10122406, accession ID cd03011) associated with the catalyzation of dithiol oxidation or disulfide reduction of target proteins [61].

The GenomeNet program identified six different motifs, including AhpC-TSA (position between 43–144, independent E-value $6.8 \times 10^{-15}$), Redoxin (position between 44–137, independent E-value $1.60 \times 10^{-19}$), thioredoxin (position between 52–114, independent E-value $1.0 \times 10^{-6}$), thioredoxin-2 (position between 59–164, independent E-value $4.30 \times 10^{-8}$), thioredoxin-8 (position between 61–146, independent E-value $3.6 \times 10^{-8}$), and thioredoxin-9 (position between 57–108, independent E-value $8.3 \times 10^{-7}$). The Pfam program [62] and the ScanProsite tool also validated...
the six motifs, including AhpC-TSA (accession ID PF00578), Redoxin (accession ID PF08534), thioredoxin (accession ID PF00085), thioredoxin-2 (accession ID PF13098), thioredoxin-8 (accession ID PF13905), and thioredoxin-9 (accession ID PF14595). Moreover, the SuperFamily program anticipated the protein as a member of the thioredoxin-like superfamily (accession ID 52833, E-value 2.1 × 10^{-30}). Thioredoxins are small proteins composed of around one hundred amino acid residues that participate in numerous redox processes. Thioredoxins operate by reversibly oxidizing an active center disulfide bond. An intramolecular disulfide bond connects the two cysteine residues in reduced or oxidized forms [63-68].

**Protein-protein interaction**

The cellular machinery is supported by proteins as well as their functional connections. For a comprehensive comprehension of biological events, their connection network must be considered, yet the existing knowledge on protein-protein relationships is inadequate and of different annotation granularity and trustworthiness. The STRING database attempts to gather, assess, and integrate all publicly accessible sources of knowledge on protein-protein interactions and supplement them with computational predictions. It seeks to establish a worldwide network that is complete and objective, incorporating both primary (physical) and secondary (functional) linkages [69-72]. The STRING program (v.11.5) was performed to determine the protein-protein (pr-pr) interaction (Fig. 2). The string program demonstrated the functional follows with the scores as of trxC (0.795), Rv2876 (0.734), dipZ (0.884), Rv1929c (Rv1929c), mpt63 (0.731), Rv1676 (0.818), Rv2968c (0.798), Rv1929c (0.795), Rv2969c (Rv2969c), and Rv1138c (0.763).

The trxC, Rv2876, dipZ, Rv1929c, mpt63, Rv1676, Rv2968c, Rv1929c, Rv2969c, and Rv1138c are the thioredoxin that participates in multiple redox reactions through the reversible oxidation of its active center dithiol to a disulfide and catalyzes dithiol-disulfide exchange reactions, possible conserved transmembrane protein, cytochrome c-type biogenesis protein, conserved hypothetical protein, immunogenic protein mpt63 (antigen mpt63/mpb63), uncharacterized protein, probable conserved integral membrane protein, conserved hypothetical protein (uncharacterized), possible conserved membrane or exported protein, and possible conserved membrane or exported protein, respectively [73-76]. The mtp53

![Fig. 2. The STRING network determines the protein-protein (pr-pr) interactions. For node content: colored nodes–query proteins and first shell of interactors, white nodes–the second shell of interactors, empty nodes–proteins of unknown 3D structure, filled nodes–some 3D structure is known or predicted.](https://doi.org/10.5808/gi.23001)
is a soluble secreted antigen mpt53 precursor, disulfide oxidoreductase, that speeds up the oxidation of diminished, unfolded secreted proteins to make disulfide bonds [77].

**Secondary structural assessment**
Static high-resolution structures have contributed significantly to our protein structure and molecular activity knowledge. As structural biology has progressed, it has become evident that high-resolution structures alone cannot adequately represent the molecular basis for the structure and the action of proteins in solution [78-80]. The secondary structural components, such as helix, sheet, coil, and turn, strongly correlate with proteins’ operation, architecture, and interaction [81-84]. The SOPMA program identified the alpha-helix (n = 66, 38.15%), extended strand (n = 39, 22.54%), beta-turn (n = 13, 7.52%), and random coil (n = 55, 31.79%) (Fig. 3).

The SPIPRED (v:4.0) and the DISOPRED (v:3.0) programs were used to determine the secondary structure, sequence plot, and transmembrane topology (Fig. 4).

**Structure prediction and validation**
Homology modeling is a technique for constructing the three-dimensional structures of proteins based on their primary sequence and using existing information from structural matches to other proteins. Sequence/structure compatibility is improved in the homology modeling procedure, a framework is constructed, and side chains are appended [85,86]. The HHpred is an accessible, collaborative web service for protein bioinformatics analysis. Experts and non-experts have access to a vast array of integrated outside-generated, cutting-edge bioinformatics tools [87].

The HHpred is a robust technology for remote homology determination and structure prediction, first constructed as hidden Markov models as well as popularized by the first pairwise comparison study of homologous protein patterns. It permits several repositories, such as PDB, CDD, Pfam, SMART, SCOP, and COG [38]. It accepts a single query array or many lineups as an entry and provides the results via a user-friendly layout similar to PSI-
BLAST. Local or worldwide integration and the discovery of secondary systems are among the screening capabilities. HHpred can construct multiple inquiry prototypes, various model alignments with numerous schemes, and three-dimensional representations calculated with the Modeller program from these combinations [39]. The most suitable template (HHpred ID: 1LU4_A, PDB ID: 1LU4) was selected with the probability (99.92%), E-value $1.7 \times 10^{-22}$, and target length of 136.

Moreover, the PROCHECK program of the SAVES (v.6.0) tool was used for the Ramachandran plot assessment (Fig. 5). The amino acid sequences in the most favored regions, residues in additional allowed regions, the number of glycine residues (shown as triangles), and the number of proline residues is 9 are 106 (91.4%), 10 (8.6%), 8, and 9, respectively. Likewise, the Verify3D program demonstrated as 100% of the residues averaged a 3D-1D score (≥0.2), whereas at least 80% of the amino acids scored (≥0.2) in the 3D/1D profile to pass [88]. Identifying flaws in theoretical and experimental representations of protein architecture is a fundamental challenge in structural biology. ProSA is a well-known application with a broad user base commonly used to improve and assess empirical protein architectures and structure projection and analysis. Protein structural investigation is often a demanding and laborious process [42]. In addition, the ProSA-web calculated the Z-score as -6.53.

**Antigenicity, allergenicity, and toxicity**

Vaccine development in the post-genomic age often commences with the *in silico* assessment of genome data, with the most likely defensive antigens anticipated instead of the cultivation of pathogenic bacteria. Despite the apparent benefits of this method, such as speed and cost-effectiveness, its success is contingent on the precision of antigen prediction. Antigens are identified using sequence alignment in most cases [89-93]. This situation is hazardous for several reasons. Specific proteins may share comparable structures and biological activities despite visible sequence similarity. The antigenicity of a sequence may be encoded in a subtle and convoluted manner, making straightforward detection by sequence alignment impossible [94-96]. Considering the protein’s physical and chemical attributes, the VaxiJen program projected that it was antigenic, with the baseline threshold of 0.4 used as the
Fig. 6. Active site determination. (A) Active sites of the protein. The "red sphere" indicates the active sites of the protein. (B) The amino acid residues in the active site (blue color).

antigenicity parameter. The overall anticipated antigenicity score was measured as 0.5936.

Allergy overreaches the immune function to a formerly exposed, normally innocuous chemical, leading to skin rash, mucous membrane swelling, sneezing or wheezing, or other aberrant symptoms. The rising prevalence of altered proteins in food, commercial items, laundry detergent, medical therapies, and diagnostics renders anticipating and detecting possible allergies a significant social concern. Using bioinformatics, allergen prediction has been extensively studied, and several tools have been created over the past decade; many are accessible on the complimentary internet [97,98]. Furthermore, the AllerTOP (v. 2.0) anticipated the protein as of probable non-allergen protein. Over the last several decades, scientific study has focused on developing peptide/protein-based treatments for various ailments. With various benefits over small molecules, including high selectivity, significant penetration, and simplicity of production, peptides have emerged as prospective therapeutic agents against various disorders. However, the toxicity of peptide- and protein-based therapies is one of their limitations. To forecast the toxicity of peptides and proteins, we built in silico models in this work [99-101]. The ToxinPred program predicted the protein as nontoxic.

Adaptation between pathogens and their innholders has resulted in several metabolic strategies employed by intracellular infections to cope with the defense responses and nutritional insufficiencies throughout infection. Comprehending how proteins act is essential for explaining how they operate, and this protein contains disulfide oxidoreductase, a crucial enzyme associated with the oxidation of dithiol and/or the reduction of disulfide in target sites. This study reveals the fundamental characteristics of the protein of MTB. Moreover, the protein-protein interactions, active amino acid residues, allergenicity, antigenicity, and toxicity uncover the protein potentiality of MTB infection.

ORCID
Abu Saim Mohammad Saikat: https://orcid.org/0000-0002-6850-1245

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

References


A bioinformatic approach to identify pathogenic variants for Stevens-Johnson syndrome

Muhammad Ma’ruf¹, Justitia Cahyani Fadli¹, Muhammad Reza Mahendra¹, Lalu Muhammad Irham¹²,³, Nanik Sulistyani¹, Wirawan Adikusuma¹²,³, Rockie Chong⁵, Abdi Wira Septama²

¹Faculty of Pharmacy, Universitas Ahmad Dahlan, Yogyakarta 55164, Indonesia
²Research Centre for Pharmaceutical Ingredients and Traditional Medicine, National Research and Innovation Agency (BRIN), South Tangerang 15314, Indonesia
³Departemen of Pharmacy, University of Muhammadiyah Mataram, Mataram 83127, Indonesia
⁴Center for Vaccine and Drugs, Research Organization for Health, National Research and Innovation Agency (BRIN), South Tangerang 15314, Indonesia
⁵Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095, USA

Stevens-Johnson syndrome (SJS) produces a severe hypersensitivity reaction caused by Herpes simplex virus or mycoplasmal infection, vaccination, systemic disease, or other agents. Several studies have investigated the genetic susceptibility involved in SJS. To provide further genetic insights into the pathogenesis of SJS, this study prioritized high-impact, SJS-associated pathogenic variants through integrating bioinformatic and population genetic data. First, we identified SJS-associated single nucleotide polymorphisms from the genome-wide association studies catalog, followed by genome annotation with HaploReg and variant validation with Ensembl. Subsequently, expression quantitative trait locus (eQTL) from GTEx identified human genetic variants with differential gene expression across human tissues. Our results indicate that two variants, namely rs2074494 and rs5010528, which are encoded by the HLA-C (human leukocyte antigen C) gene, were found to be differentially expressed in skin. The allele frequencies for rs2074494 and rs5010528 also appear to significantly differ across continents. We highlight the utility of these population-specific HLA-C genetic variants for genetic association studies, and aid in early prognosis and disease treatment of SJS.

Keywords: bioinformatics, genetic variation, genomic, pathogenic variants, Stevens-Johnson syndrome

Introduction

Stevens-Johnson syndrome (SJS) and toxic epidermal necrolysis (TEN) are potentially life-threatening diseases [1]. In particular, SJS is a syndrome resulting from severe hypersensitivity reactions caused by infection with the herpes simplex or mycoplasmal viruses, vaccinations, systemic diseases, certain agents, food, and drugs [2]. SJS occurs in the skin and mucous membranes in the orifices and eyes in mild to severe conditions, with abnormalities in the skin in the form of an erythema, vesicles, or bullae accompanied by purpura [3].

SJS/TEN is widespread in the mucocutaneous immune region, causing exfoliation of the skin on the mucosal surface [4]. The incidence of SJS, SJS/TEN, and TEN in the United States reported 9.2, 1.6, and 1.9 cases from 2009–2012 [5]. The incidence of SJS...
cases that occur in Indonesia is around 12 cases per year, with different causes [6]. The incidence in the United Kingdom from 1995–2013 there were 5,766 cases of SJS/TEN per million people per year [7]. In Korea, it is reported that the incidence rate of SJS/TEN from 2009–2013 was 3.96–5.03 and 0.94–1.45 per million people per year [8]. Events caused by SJS, SJS-TEN, and TEN have an average of 5.3, 0.8, and 0.4 cases per million children each year [9]. The mortality rate of 4.8–9% in SJS, 19.4–29% in the SJS/TEN case, and 14.8–48% in TEN [10]. SJS can appear with non-specific fever symptoms that cause malaise, headache, cough, and rhinorrhea. On the skin, patients suffering from SJS can have polymorphic lesions and mucous membranes with marked skin blisters and erosion [11].

Therefore, primary prevention is the best mitigation for SJS. SJS is categorized as a severe cutaneous adverse reaction (SCAR), and several drugs have been implicated in disease pathogenesis. Non-steroidal anti-inflammatory drugs and other multi-ingredient formulations are widely used to relieve the symptoms. Several studies are reporting adverse skin drug reactions that are often SJS-associated with severe ocular complications [12]. Prevention is possible if patients who are susceptible to this SCAR when prescribed certain drugs are identified. Besides, the genomic variants known to have an important role in SJS progression. However, little information revealed the specific variant as a biological risk variant [13]. Even though several studies have been exploited previously, a limited number of variants summarized the variants associated with expression in tissue-related SJS. This study aims to investigate the variants associated with SJS through a bioinformatic-based approach and further prioritize the biological risk variants. Besides, the pattern of gene expression profiles and population allele frequencies of genetic variants were assessed using various databases. Here, the results will enable future studies to assess whether these variants may be associated with various infectious risks for SJS/TEN, as well as SJS progression and disease susceptibility.

### Methods

Genomic information not only can be leveraged to identify the variant-associated disease, but it can also be translated into actionable knowledge for the disease. SJS is one of the severe skin reaction due to genomic risk factors. In this study, we used a bioinformatic-based approach to prioritize the pathogenic variants that potentially trigger the SJS. Detailed information regarding the study design has been depicted in Fig. 1. We used the keyword “Stevens-Johnson syndrome (SJS)” to derive SJS associated from the genome-wide association studies (GWAS) National Human Genome Research Institute (NHGRI) Catalog database (http://www.ebi.ac.uk/gwas; accessed December 19, 2022). SJS associated with 74 single nucleotide polymorphisms (SNPs) were obtained and further analysis was carried out using HaploReg (version 4.1). Further analysis yielded a total of 41 SNPs with a significance value of $p$-value < $10^{-8}$. This value is used to account for several tests in the GWAS catalog. These values are widely used to identify associations between variants and shared genetic traits with adjacent gene expression [14]. Furthermore, an evaluation was carried out between the relationships of various genetic variants and gene expression profiles using expression quantitative trait locus (eQTL) using the GTEx Portal database (http://www.gtexportal.org/home/; accessed December 19, 2022), which was found by gene expression from various networks. The HLA-C (human leukocyte antigen C) genetic variant is present in human skin and tissue (lower extremities) exposed to sunlight obtained from the GTEx Portal database. Then confirm the variant using the Ensembl Genome Browser (https://wwwensembl.org/index.html; accessed December 19, 2022). Furthermore, the allele frequencies of variants associated SJS were evaluated in different populations including African, American, East Asian, European, and Southeast Asian populations. Samples from each region consisted of 331 individuals (Africa), 199 individuals (America), 247 individuals (East Asia), and 136 individuals (Southeast Asia). Then, to find out the function of the various gene variants, an evaluation was carried out using the SNP Nexus database (https://www.snp-nexus.org; accessed December 19, 2022).

### Results and Discussion

#### Identification of genomic variants for SJS

We first identified SNPs associated with SJS from the GWAS database, resulting in 74 SNPs associated with SJS. We identified 41 unique SNPs associated with SJS after removing all SNPs duplication (Table 1). Based on the number of SNPs obtained, candidate SNPs were further constrained and prioritized using HaploReg version 4.1, with a $p$-value of < $10^{-8}$. Based on the findings presented in Table 2, we focused on two genomic variants from the same gene that qualify as the biological risk SNPs for SJS from this study.

Through our integrative bioinformatics approach, two variants with a missense mutation (rs2074494, rs5010528) that encoded...
Fig. 1. Bioinformatics workflow for the identification of genetic variants associated with Stevens-Johnson syndrome. GWAS, genome-wide association studies; SNP, single nucleotide polymorphism.
Table 1. SNPs from the GWAS catalog with a significance of p-value < 10^{-8}

<table>
<thead>
<tr>
<th>Variant and risk allele</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs6457109</td>
<td>3 × 10^{-16}</td>
</tr>
<tr>
<td>rs7760545</td>
<td>4 × 10^{-16}</td>
</tr>
<tr>
<td>rs35835721</td>
<td>8 × 10^{-15}</td>
</tr>
<tr>
<td>rs137899365</td>
<td>3 × 10^{-14}</td>
</tr>
<tr>
<td>rs60581484</td>
<td>2 × 10^{-13}</td>
</tr>
<tr>
<td>rs1131151</td>
<td>4 × 10^{-12}</td>
</tr>
<tr>
<td>rs2074494</td>
<td>5 × 10^{-12}</td>
</tr>
<tr>
<td>rs199755581</td>
<td>8 × 10^{-11}</td>
</tr>
<tr>
<td>rs1562468327</td>
<td>6 × 10^{-10}</td>
</tr>
<tr>
<td>rs199755581</td>
<td>1 × 10^{-9}</td>
</tr>
<tr>
<td>rs536142737</td>
<td>2 × 10^{-9}</td>
</tr>
<tr>
<td>rs16957893</td>
<td>2 × 10^{-8}</td>
</tr>
<tr>
<td>rs2734583</td>
<td>2 × 10^{-8}</td>
</tr>
<tr>
<td>rs150289893</td>
<td>2 × 10^{-8}</td>
</tr>
<tr>
<td>rs126845082</td>
<td>2 × 10^{-8}</td>
</tr>
<tr>
<td>rs48089948</td>
<td>2 × 10^{-8}</td>
</tr>
<tr>
<td>rs3094188</td>
<td>3 × 10^{-8}</td>
</tr>
<tr>
<td>rs55765602</td>
<td>3 × 10^{-8}</td>
</tr>
<tr>
<td>rs77542827</td>
<td>3 × 10^{-8}</td>
</tr>
<tr>
<td>rs778096762</td>
<td>3 × 10^{-8}</td>
</tr>
<tr>
<td>rs1597607761</td>
<td>3 × 10^{-8}</td>
</tr>
<tr>
<td>rs1391213386</td>
<td>3 × 10^{-8}</td>
</tr>
<tr>
<td>rs374138762</td>
<td>4 × 10^{-8}</td>
</tr>
<tr>
<td>rs879665274</td>
<td>4 × 10^{-8}</td>
</tr>
<tr>
<td>rs1211926109</td>
<td>4 × 10^{-8}</td>
</tr>
<tr>
<td>rs116953913</td>
<td>5 × 10^{-8}</td>
</tr>
<tr>
<td>rs1263106470</td>
<td>6 × 10^{-8}</td>
</tr>
<tr>
<td>rs5010528</td>
<td>8 × 10^{-8}</td>
</tr>
</tbody>
</table>

SNP, single nucleotide polymorphism; GWAS, genome-wide association studies.

the HLA-C genes were prioritized as the biological risk SNPs for SJS. SJS is characterized by necrosis and shedding of the epidermis, known as the triad of disorders of the vesiculobullous skin, orifice mucosa, and eyes, accompanied by severe general symptoms [6]. It was also reported that the HLA-C gene has an important role in protecting against cancer and viruses. However, the HLA-C gene may also be involved in allograft rejection, the state of preeclampsia, and is also present in autoimmune diseases [15]. The diagnosis of SJS/TEN is a blistering autoimmune disease, which is included in linear IgA dermatosis and paraneoplastic pemphigus but is also present in pemphigus vulgaris and bullous pemphigoid, acute generalized exanthematous pustulosis, and later disseminated drug persistently erupting bullosa and staphylococcal scalded skin syndrome [16].

**HLA-C gene expression across 16 human tissues**

To evaluate HLA-C gene expression in human tissues, we used the GTEx Portal database (http://www.gtexportal.org/), which contains gene expression levels in various tissues. eQTL annotation comprises the most apparent functional consequences of genetic variation. Whole blood, spleen, lung, and lymphocyte cells demonstrate the highest HLA-C gene expression across the 16 human tissues analyzed from GTEx in Fig. 2. Furthermore, we have found that the ID SNPs rs2074494 and rs5010528 have similar gene expression variation in the Sun-Exposed Skin (Lower Leg).

**HLA-C gene expression in the sun-exposed skin**

Human leukocyte antigen (HLA) class I genes, including HLA-A, HLA-B, and HLA-C, have been reported as the loci most strongly associated with susceptibility to all types of SJS and TEN, including cold medicine-related (CM-SJS) and TEN with severe ocular complications (SOC). Although non-synonymous substitutions affecting peptide binding or HLA molecular conformation have been considered significant factors in the pathogenesis of immunological diseases, indeed, different HLA-C expression levels have been reported for the different alleles, with higher HLA-C expression leading to increased Tc (Trypanosoma cruzi) responses and adverse effects in Crohn's disease. It was also reported that genetic variation in HLA-A and other autosomal genes has been identified as a risk factor for SJS/TEN associated with flu drugs with SOC such as CM-SJS or TEN with SOC [18].

https://doi.org/10.5808/gi.23010
When individuals with a genetic background containing SJS/TEN with SOC susceptibility factors are infected by some viral or microbial infection, they develop abnormal immune responses [19]. It is reasonable to presume that there is an interaction between HLA multiplication and susceptibility genes such as HLA-A and TLR3 [20], HLA-A and REC14-32, and HLA-A and PTGER3 [21]. Several susceptibility genes for SJS/TEN CM associated with SOC may be involved in the formation of functional networks. An imbalance in this gene can trigger the mucocutaneous inflammation seen in patients with SJS/TEN associated with CM. SJS/TEN with SOC in the acute stage shows inflammation of the skin and ocular surface and oral mucosal erosions and paronychia [12].

### Table 2. Stevens-Johnson syndrome variant and risk allele that codes for prioritized SNPs

<table>
<thead>
<tr>
<th>Variant and risk allele</th>
<th>Variants near risk allele (r² &gt; 0.8)</th>
<th>p-value</th>
<th>Gencode</th>
<th>Allele type</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2074494</td>
<td>rs1050276</td>
<td>5 × 10⁻¹²</td>
<td>HLA-C</td>
<td>Missense</td>
</tr>
<tr>
<td>rs5010528</td>
<td>rs1050409</td>
<td>8 × 10⁻⁶</td>
<td>HLA-C</td>
<td>Missense</td>
</tr>
</tbody>
</table>

SNP, single nucleotide polymorphism; HLA-C, human leukocyte antigen C.

#### Fig. 2. Human leukocyte antigen C (HLA-C) gene expression–associated with Stevens-Johnson syndrome in several human tissues from the GTEx Portal.

When individuals with a genetic background containing SJS/TEN with SOC susceptibility factors are infected by some viral or microbial infection, they develop abnormal immune responses [19]. It is reasonable to presume that there is an interaction between HLA multiplication and susceptibility genes such as HLA-A and TLR3 [20], HLA-A and REC14-32, and HLA-A and PTGER3 [21]. Several susceptibility genes for SJS/TEN CM associated with SOC may be involved in the formation of functional networks. An imbalance in this gene can trigger the mucocutaneous inflammation seen in patients with SJS/TEN associated with CM. SJS/TEN with SOC in the acute stage shows inflammation of the skin and ocular surface and oral mucosal erosions and paronychia [12].

### Relationship between HLA-C gene and eQTLs from the GTEx database

Gene expression of SJS in human tissue was evaluated via the GTEx Portal database. This aims to determine the gene expression level in various tissues including in the skin tissue. We identified genomic variations from HLA-C gene expression using the GWAS catalog database and found 74 SNPs. From these analyses, we determined top 10 SNPs with the highest p-values. We were further processed with the variant annotation tool SNPnexus to determine the annotation of prioritized SNP variations. After these analyses, two statistically significant SNPs were obtained and prioritized. In this case, we prioritized 2 SNPs at risk for SJS based on an analysis of the number of SNPs expanded using HaploReg version 4.1 and a p-value of < 10⁻⁸ to determine the functional annotation of the SNPs. The results of genetic variation are shown in Table 3.

Table 3 shows the two identified variants (rs2074494 and rs5010528) encoded the HLA-C gene in differential tissue expression in the human skin. By using the GTEx portal (http://www.gtexportal.org/home/), we further emphasized that the variants of rs2074494 and rs5010528 encoded the HLA-C genes were a higher expression in the skin tissue to Table 3 and Fig. 3.

### Allele frequencies of SJS candidate variants across continents

Once we identified the candidate HLA-C expression–associated variants, we set out to determine the allele frequencies across transcontinental populations as shown in Table 4. The allele fre-
<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Gencode ID (ENSG00000-)</th>
<th>Gene symbol</th>
<th>p-value</th>
<th>Effect size</th>
<th>Tissue</th>
<th>Expression level</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2074494</td>
<td>204525.16</td>
<td>HLA-C</td>
<td>7.2 x 10^{-10}</td>
<td>-0.48</td>
<td>Artery-Tibial</td>
<td>CC &gt; CT &gt; TT</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>13 x 10^{-6}</td>
<td>-0.50</td>
<td>Muscle-Skeletal</td>
<td>CC &gt; CT &gt; TT</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>15 x 10^{-5}</td>
<td>-0.78</td>
<td>Brain-Nucleus accumbens (basal ganglia)</td>
<td>CC &gt; CT &gt; TT</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>17 x 10^{-6}</td>
<td>-0.51</td>
<td>Heart-Left ventricle</td>
<td>CC &gt; CT &gt; TT</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>46 x 10^{-6}</td>
<td>-0.58</td>
<td>Testis</td>
<td>CC &gt; CT &gt; TT</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>36 x 10^{-5}</td>
<td>-0.39</td>
<td>Skin-Sun exposed (lower leg)</td>
<td>CC &gt; CT &gt; TT</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>5.4 x 10^{-26}</td>
<td>0.56</td>
<td>Adipose-Subcutaneous</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>1 x 10^{-21}</td>
<td>0.41</td>
<td>Whole blood</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>1.4 x 10^{-14}</td>
<td>0.51</td>
<td>Adipose-Visceral</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>4.1 x 10^{-14}</td>
<td>0.49</td>
<td>Lung</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>1.1 x 10^{-10}</td>
<td>0.68</td>
<td>Spleen</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>2.1 x 10^{-9}</td>
<td>0.51</td>
<td>Heart-Atrial appendage</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>5.7 x 10^{-5}</td>
<td>0.28</td>
<td>Colon-Transverse</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>9.9 x 10^{-5}</td>
<td>0.38</td>
<td>Skin-Not sun exposed (suprapublic)</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>5.1 x 10^{-7}</td>
<td>0.72</td>
<td>Cells-EBV-transformed lymphocytes</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>7.0 x 10^{-7}</td>
<td>-0.25</td>
<td>Nerve-Tibial</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>7.7 x 10^{-7}</td>
<td>0.31</td>
<td>Nerve-Tibial</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>1.6 x 10^{-6}</td>
<td>0.34</td>
<td>Breast-Mammary tissue</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
<tr>
<td></td>
<td>204525.16</td>
<td>HLA-C</td>
<td>1.8 x 10^{-6}</td>
<td>0.32</td>
<td>Skin-Sun exposed (lower leg)</td>
<td>AA &gt; AG &gt; GG</td>
</tr>
</tbody>
</table>

Source: Expression Quantitative Trait Loci (eQTL) obtained from the GTEx Portal. *HLA-C*, human leukocyte antigen C; eQTL, expression quantitative trait loci; SNP, single nucleotide polymorphism.

**Fig. 3.** Human leukocyte antigen C (*HLA-C*) gene expression for each genotype of the single nucleotide polymorphisms: (A) rs2074494 and (B) rs5010528.

Frequencies for both variants were evaluated in different populations including African, American, East Asian, European, and Southeast Asian populations. Samples from each region consisted of 331 individuals (Africa), 199 individuals (America), 247 individuals (East Asia), and 136 individuals (Southeast Asia). We extracted the allele frequencies in Africa, America, East Asia, Europe, and
Southeast Asia from the Ensemble Genome Browser (http://www.ensembl.org). Allele frequencies across populations differ for each HLA-C variant. Table 4 and Fig. 4 show gene expression levels at higher frequencies of the rs5010528 associated (G) allele and rs2074494 associated (T) allele. At population frequencies of the rs5010528 (G) allele, the Asian population (East Asia and Southeast Asia) is expressed at a much lower level than that from the populations of Africa, America, and Europe.

The allele frequency of the “rs2074494” T allele in the African population is expressed at a much lower level than that of the populations of America, Europe, and Southeast Asia. Taken together, the allele frequencies of the variants “rs2074494” and “rs5010528” indicated the contribution of differential variant prevalence for HLA-C gene expression.

Another study revealed that HLA-S gene which were reported to be potential associated in chickenpox disease [22]. In this study, we investigated the skin tissue expression of the HLA-C gene, which has been linked to SJS and can lead to SCAR infection. Notably, the variants associated with HLA-C expression have not been reported for SJS. Basic research on the genetics of SJS and TEN to date has focused on HLA, the system associated with the presence of specific receptors, cytotoxic proteins, and the part of immunocytes during disease pathogenesis [23]. Considering the global impact of SJS, examining the distribution of HLA-C variants

Table 4. Analysis of allele frequencies for the HLA-C gene from SNP nexus variant annotation

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Position (hg38)</th>
<th>Gene symbol</th>
<th>Location</th>
<th>Allele</th>
<th>Allele frequency (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2074494</td>
<td>Chr6:31271956</td>
<td>HLA-C</td>
<td>Missense</td>
<td>C T</td>
<td>T: 0.006 (10) T: 0.099 (69) T: 0.189 (191) T: 0.037 (37) T: 0.040 (39)</td>
</tr>
<tr>
<td>rs5010528</td>
<td>Chr6:31273255</td>
<td>HLA-C</td>
<td>Missense</td>
<td>A G</td>
<td>G: 0.243 (321) G: 0.187 (130) G: 0.056 (56) G: 0.138 (139) G: 0.099 (97)</td>
</tr>
</tbody>
</table>

HLA-C, human leukocyte antigen C; SNP, single nucleotide polymorphism; Ref, Reference; Alt, alternative; AFR, Africa; AMR, America; EAS, East Asia; EUR, Europe; SAS, Southeast Asia.

![Fig. 4. Summary of allele frequency analysis on human leukocyte antigen C (HLA-C) gene expression in Africa, America, East Asia, Europe, and Southeast Asia.](https://doi.org/10.5808/gi.23010)
may be an essential quest that allows further understanding of global disease susceptibility.

The HLA gene encodes several molecules that are crucial to the immune system. With that in mind, a strong relationship between HLA genes and autoimmune diseases has been demonstrated for more than half a century \[24\]. Findings that most patients with carbamazepine-induced SJS and CBZ-SJS/TEN toxic epidermal necrolyses have an associated HLA-B*15:02 in an Asian population. In contrast, the association with HLA-A*31:01 was only reported in Japan and Europe. and has a novel association between HLA-A*31:01 and CBZ-SJS/TEN in Indians \[25\]. The association with various HLA genes can then be analyzed using publicly available databases. We used publicly available databases like the GTeX Portal, SNPnexus, and Ensembl. We identified genetic variants associated with HLA-C expression in skin tissue, the leading site of SCAR infection in SJS disease. It has previously been reported that specific HLA genotypes have been associated with the occurrence of severe skin disease due to drug-induced side effects (SCARs), which cases are included in SJS/TEN \[26\].

The allele frequencies in all populations differ for each SNP, as shown in Fig. 4. In general, it is known that the G and A allele frequencies for rs2074494 and rs5010528 were also seen to have a lower frequency in Southeast Asia (rs2074494, 4% and rs5010528, 14%), Europe (rs2074494, 4% and rs5010528, 16%), and East Asia (rs2074494, 19% and rs5010528, 6%), compared with American (rs2074494, 10% and rs5010528, 24%) and African (rs5010528, 24%) populations. In conclusion, by leveraging a bioinformatic-based approach it is revealed the pathogenic variants that are potentially associated with SJS. We propose that these variants could be used for further study to identify the SJS diagnostic biomarker as well as for prognosis. However, we acknowledged that there are limitations to the bioinformatic-based approach used to investigate the genetic variants associated with SJS. One of the main limitations is that not all the variants necessarily have genes that encode them (i.e., non-coding variants), and even if they do, these genes or genetic variants may not be suitable drug targets. Nonetheless, clinical validation is recommended as a next step to confirm our findings and gain a better understanding of the underlying etiology and functional effect of the SJS disease.

Conclusion
In this study, we conducted a comprehensive bioinformatic analysis of SJS from genomic databases, revealing differential tissue expression of the HLA-C gene across 16 human tissues. Even though HLA-C is highly expressed in whole blood, spleen, lung, and lymphocyte cells, the relevant disease variants (rs2074494 and rs5010528) are differentially expressed in the skin tissue. Overall, alleles for rs2074494 and rs5010528 have lower frequencies in Southeast Asia (rs2074494, 4% and rs5010528, 14%), Europe (rs2074494, 4% and rs5010528, 16%), and East Asia (rs2074494, 19% and rs5010528, 6%), as compared to the American population (rs2074494, 10% and rs5010528, 24%) and the African (rs5010528, 24%) population.

ORCID
Muhammad Ma’ruf: https://orcid.org/0009-0006-5088-8089
Justitia Cahyani Fadli: https://orcid.org/0009-0005-0072-5823
Muhammad Reza Mahendra: https://orcid.org/0009-0002-9754-1198
Lalu Muhammad Irham: https://orcid.org/0000-0002-0091-4887
Nanik Sulistyani: https://orcid.org/0000-0002-6541-8198
Wirawan Adikusuma: https://orcid.org/0000-0001-9165-690X
Rockie Chong: https://orcid.org/0000-0001-6736-9687
Abdi Wira Septama: https://orcid.org/0000-0002-6471-7735

Authors’ Contribution
Conceptualization: MM, JCF, MRM, LMI. Data curation: MM, JCF, MRM, LMI, WA, RC. Formal analysis: LMI, WA, RC. Methodology: MM, JCF, MRM. Writing – original draft: MM, JCF, MRM, LMI. Writing – review & editing: MM, JCF, MRM, LMI, NS, WA, RC, AWS.

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

Acknowledgments
Thank you to the Faculty of Pharmacy at Universitas Ahmad Dahlan for allowing us to guide bioinformatics lessons.

References


https://doi.org/10.5808/gi.23010
Assessing the impact of recombination on the estimation of isolation–with–migration models using genomic data: a simulation study

Yujin Chung*

Department of Applied Statistics, Kyonggi University, Suwon 16227, Korea

Recombination events complicate the evolutionary history of populations and species and have a significant impact on the inference of isolation–with–migration (IM) models. However, several existing methods have been developed, assuming no recombination within a locus and free recombination between loci. In this study, we investigated the effect of recombination on the estimation of IM models using genomic data. We conducted a simulation study to evaluate the consistency of the parameter estimators with up to 1,000 loci and analyze true gene trees to examine the sources of errors in estimating the IM model parameters. The results showed that the presence of recombination led to biased estimates of the IM model parameters, with population sizes being more overestimated and migration rates being more underestimated as the number of loci increased. The magnitude of the biases tended to increase with the recombination rates when using 100 or more loci. On the other hand, the estimation of splitting times remained consistent as the number of loci increased. In the absence of recombination, the estimators of the IM model parameters remained consistent.

Keywords: coalescent, gene tree, isolation–with–migration models, recombination, simulation

Introduction

The estimation of divergence between populations and species is a central topic in population genetics and evolution. The difficulty in determining divergence from genetic data is due to the presence of opposing evolutionary processes. For instance, genetic drift increases divergence, and gene flow reduces it [1,2]. The isolation–with–migration (IM) model is a widely adopted demographic model that seeks to reconcile conflicting signals. It models the divergence of two populations from a common ancestral population at a specific time in the past, and accounts for the exchange of migrants between the two populations [1–4].

DNA sequence alignments are commonly used data in the study of IM models [2,5]. IM models typically assume that alignments are shaped exclusively by neutral evolutionary processes and do not take into account the influence of selection, with an absence of recombination within a locus and free recombination between loci [5]. It is deemed essential to confirm that the DNA sequence alignments are orthologs, because the relationship between homologous DNA sequences is believed to be result from past branching processes [6]. To minimize the potential for recombination within a locus, filtering using
a four-gamete test [7] should be employed. With the increasing availability of data from the nuclear portions of genomes, it has become possible to obtain data with a history that may include recombination events [6]. However, the four-gamete test has limitations; it may not detect all recombination events [6,7], and methods for filtering genomic data based on four-gamete test results have not been thoroughly evaluated.

Hey and Wang [6] conducted a simulation study on the effect of four-gamete filtering on IM model inference using the IMa3 program [8]. To assess the impact of the four-gamete test, they compared three methods of non-recombined block sampling: the longest interval, overlapping interval, and random non-overlapping interval sampling. The results showed that the distributions of the migration rate parameters were flatter with recombination, and the distribution of the maximum a posteriori (MAP) values shifted to the right. This implies that using four-gamete filtering reduces the statistical power for detecting non-zero migration rates. They also found that random intervals performed better, although using the longest interval led to higher-resolution results because of using more data. These results can be applied to genealogy-sampling-based methods for IM models or subsets of IM models.

The study conducted by Hey and Wang [6] provides a significant opportunity for further investigation into the impact of recombination and the filtering method. First, it is important to examine the sources of estimation errors present in their findings. These errors arise from estimating two levels of uncertainty: the distribution of DNA sequences given a genealogy and that of genealogy given an IM model [2,5]. Additionally, errors in detecting recombination breakpoints can also contribute to the overall errors. Hence, it is necessary to understand the source of errors in Hey and Wang’s [6] study and to differentiate between the estimation errors arising from the two levels of uncertainty in the IM model inference and from recombination detection. This information is crucial for improving the accuracy of further studies and fully understanding the impact of the recombination and filtering methods.

Secondly, it is important to extend Hey and Wang’s study [6] to encompass genomic data. As the availability of genomic data increases, software such as MIST [1] has been developed to analyze large amounts of genomic data [2]. However, the study conducted by Hey and Wang [6] focuses on data containing 10–50 loci. Hence, it is imperative to extend the investigation to include an analysis of a larger number of loci and to examine whether the same results obtained by Hey and Wang [6] are observed with a large number of loci. Therefore, it is necessary to evaluate the consistency of the parameter estimators of IM models with a large number of loci.

This study aims to extend the study of Hey and Wang [6] and conduct a simulation-based investigation that examines the impact of recombination on the estimation of IM models using the MIST program. First, this study aims to assess the consistency of the parameter estimators, by analyzing up to 1,000 loci. Second, in order to separate the sources of the errors, we focused on errors arising from the level of uncertainty in the distribution of genealogies given an IM model. To accomplish this, we assumed that recombination breakpoints were known and analyzed the true gene trees of loci determined by the random interval sampling. This study disregarded estimation errors of gene trees and recombination breakpoint detection, focusing solely on the effect of recombination.

Methods

The impact of recombination on gene trees

This study examined how recombination affects the inference of an IM model. Genetic drift, one of the fundamental mechanisms of evolution, results in changes in the allele frequencies in a population by random chance [5]. The evolutionary paths of gene copies by genetic drift can be represented as a gene tree, and the coalescent theory [9] is commonly employed to describe the distribution of the gene tree of a locus. However, the presence of recombination events can complicate evolutionary history (Fig. 1A) and, therefore, affect the inference of an IM model. To isolate the impact of recombination on the IM model inference via gene trees, our study focused on analyzing the true gene trees of loci, which were determined using the true recombination breakpoints.

Recombination events can result in the formation of complex network-like structures. For example, Fig. 1A illustrates the complex evolutionary history of DNA sequences A–G with three recombination events. The first recombination occurred between the white and dashed DNA sequences. The second recombination occurred between the grey and white DNA sequences. Finally, the third recombination occurred between descendants of previously recombined sequences. The three recombination breakpoints in alignment A–G (Fig. 1B) produced four non-recombined blocks, each with a different gene tree as a result of recombination events.

In this study, we simulated complex evolutionary histories with and without recombination. Given the true recombination events, we were able to determine the non-recombined genetic blocks and their corresponding true gene trees. We applied random interval samplings based on the results of the study by Hey and Wang [6]. For instance, using the random interval sampling, we analyzed the gene tree corresponding to a randomly selected block from the four blocks defined by the recombination breakpoints in Fig. 1C.
IM model inference from gene trees

In a 2-population IM model, two populations of effective sizes, $N_1$ and $N_2$, have diverged from a common ancestral population of effective size $N_3$ at generation $t$ in the past. The IM model also takes into account migration between the two populations. While $M_1$ is the proportion of population 1 that is replaced by migrants from population 2 per generation, $M_2$ is the proportion of population 2 that is replaced by migrants from population 1 in each generation. The IM model provides a comprehensive framework for studying population divergence with migration dynamics between two populations over time. The calculation of the probability of alignments under an IM model is achieved through the integration of two levels of uncertainty: (1) calculating the probability distribution of an alignment given a genealogy using a mutation or substitution model \[10-13\] and (2) calculating the probability distribution of a genealogy given a demographic model with parameters $\Psi$ using a stochastic process such as coalescent processes \[9,14\]. Integration over possible genealogies is typically performed using a Markov chain Monte Carlo (MCMC) simulation \[1-4,8,15\].

In this study, we utilize the MIST program to estimate the demographic parameters of an IM using gene trees. The MIST program implements a two-step analysis to infer the IM model from DNA alignments \[1\]. In the first step, gene trees of loci without migrations are simulated through an MCMC simulation, without needing any prior information about the demographic model, thereby mitigating the issue of slow mixing. In the second step, the joint posterior density of the demographic parameters in the IM model is estimated from the sampled gene trees, and the MAP estimations of all demographic parameters are obtained. Rather than sampling migrations and the underlying demographic model parameters in the first step, integration over potential migrations is carried out in the second step when the posterior distribution of the IM model is calculated. The MIST program can be used to infer an IM model from true or estimated gene trees \[16\].

By utilizing the second stage of the MIST program, we calculated the posterior distribution of an IM model from gene trees, with uniform priors for demographic parameters, which served as the likelihood of the IM model in this study. The MAP estimates of the parameters obtained in stage 2 were equivalent to the maximum likelihood estimates of the demographic parameters.

Simulation setting

In this study, we adopted the simulation setting in Hey and Wang \[6\]. Similar to Hey and Wang \[6\], we set the neutral mutation rate per generation per base pair to $\mu = 10^{-8}$, and hence the mutation rate for the entire locus of length 5,000 base pairs to $u = 10^{-8} \times 5,000$. We denoted $r$ as the probability of a cross-over per generation between the ends of the locus. For clarity and simplicity in notation \[15\], we use demographic parameters scaled by the mutation rate throughout the remainder of this manuscript (Table 1). Similar to the simulation setting in Hey and Wang \[6\], the scaled population sizes are assumed to be equal to $\theta_i = 4N_iu = 10$ for $i = 1, 2, 3$ in the entire simulation. Splitting time $\tau = tu$ changes as 0.5, 2.5, 10 in terms of the number of mutations. The migration rates...
Table 1. Notations and simulation setting

<table>
<thead>
<tr>
<th>Demographic parameters</th>
<th>Scaled parameters</th>
<th>Description</th>
<th>True values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_i$</td>
<td>$\theta_i = 4N_iu$</td>
<td>Population sizes</td>
<td>$\theta_1 = \theta_2 = \theta_3 = 10$</td>
</tr>
<tr>
<td>$t$</td>
<td>$\tau = tu$</td>
<td>Splitting time</td>
<td>$\tau = 0.5, 2.5, 10$</td>
</tr>
<tr>
<td>$M$</td>
<td>$m = M/u$</td>
<td>Migration rates</td>
<td>$m_1 = 0, 0.1; m_2 = 0$</td>
</tr>
<tr>
<td>$r$</td>
<td>$\rho = r/u$</td>
<td>Recombination rates</td>
<td>$\rho = 0, 0.2, 1, 5$</td>
</tr>
</tbody>
</table>

$m_i = M/u$ for $i = 1, 2$ are expressed as the number of migrations per mutation. Following Hey and Wang [6], we examine two scenarios: no migration with $m_1 = m_2 = 0$ and unidirectional migration with $m_1 = 0.1$ and $m_2 = 0$ when the splitting time of an IM model is $10$ ($\tau = 10$). The recombination rate per mutation $\rho = r/u$ is varied as $0, 0.2, 1, \text{ and } 5$. It is important to note that $\rho = 0$ indicates no recombination within a locus, which is a typical assumption for the IM model inference.

As this study aimed to assess the consistency of demographic parameters in an IM model, we consider $10, 100,$ and $1,000$ loci for each case. We applied the $ms$ program [17] to simulate the gene trees and recombination events. In all cases, gene trees with four tips were simulated as two sequences were sampled from each population at each locus. For each scenario, 100 replicates were generated. When the MIST program was used to analyze the simulated gene trees, we set the uniform priors with upper bounds of $50$ for population sizes, $0.2$ for migration rates, and $20$ for the splitting time for all cases.

**Results**

We evaluated the effect of recombination on the distribution of gene trees and IM model inferences. Therefore, we used the true recombination breakpoints and analyzed true gene trees of randomly selected non-recombined loci using the MIST program.

First, we examined the consistency of the estimators of the IM model parameters in the absence of recombination. (i.e., $\rho = 0$)

When there are no migrations ($m_1 = m_2 = 0$), the estimates of the six parameters converged to their respective true value as the number of loci increased, as shown in Figs. 2 and 3 by grey lines with empty circles. The same consistency was observed when $m_1 = 0.1$, as depicted by the grey lines with empty circles in Fig. 4. Moreover, in all cases of $\rho = 0$, the standard errors decrease with the number of loci.

In the presence of recombination, the estimates of the IM model parameters were more biased than those in the absence of recombination, regardless of the absence of migration ($m_1 = 0, m_2 = 0$) or the presence of migration ($m_1 = 0.1$). For the case of no migrations ($m_1 = 0, m_2 = 0$), population sizes tended to be overestimated as the recombination rate increased when using $100$ loci or more (Fig. 2). In specific, $\hat{\theta}_1$ and $\hat{\theta}_2$ were significantly increased as $\rho$ increased ($p < 0.001$ for the slope of the regression line) at both intermediate and high splitting times, but the increase was only significant ($p < 0.001$) at the low splitting time ($\tau = 0.5$), as depicted in Fig. 2A and B. In all cases of splitting times and the number of loci (Fig. 2C), $\hat{\theta}_3$ was significantly increased with an increasing recombination rate ($p < 0.001$). In addition, the estimate of the ancestral common population size $\theta_1$ (Fig. 2C) exhibited greater bias than estimates of the other two population sizes, $\theta_1$ and $\theta_2$ (Fig. 2A and 2B). At intermediate splitting time $\tau = 2.5$ and $1,000$ loci, the bias of $\hat{\theta}_3$ was 0.6138, 1.7131, and 2.3964 as $\rho$ increased from 0.2 to 5. At high splitting time $\tau = 10$, the bias of $\hat{\theta}_3$ was alleviated somewhat as 0.7184, 1.515, and 1.86. The estimates of the migration rates and splitting time appeared to approach the true values with an increasing number of loci, as depicted in Fig. 3.

The standard errors of all parameter estimators showed a substantial reduction with an increasing number of loci in the presence of recombination (Figs. 2 and 3).

For the case of $m_1 = 0.1$, population sizes were significantly increased as $\rho$ increased ($p < 0.001$) when using $100$ or more loci (Fig. 4A–4C). Moreover, the biases in population size estimations were more severe than those in the case of no migration. For example, at high splitting time $\tau = 10$, the bias of $\hat{\theta}_3$ was 1.59039, 1.6164, and 1.9518. However, the estimate of the non-zero migration rate $\hat{m}_1$ was underestimated as the recombination rate increased, despite having a large number of loci (Fig. 4D). The estimated splitting time converged toward the true value (Fig. 4E). The standard errors of all parameter estimators showed a substantial reduction with an increasing number of loci in the presence of recombination (Fig. 4).

**Discussion**

This study extends the investigation of Hey and Wang [6]. It conducts a simulation-based investigation to examine the impact of recombination on estimating IM models using the MIST program.
In particular, we aimed to assess the consistency of the demographic parameter estimators by analyzing up to 1,000 loci and focus on errors arising from the uncertainty from the distribution of genealogies given an IM model. By assuming that recombination breakpoints are known and analyzing the true gene trees of loci randomly selected from blocks determined by the true recombination breakpoints, we isolated the effect of recombination only.

This study examines the consistency of the estimators of the IM model parameters in the absence and presence of recombination. In the absence of recombination, the estimates of the six parameters converged to their true values as the number of loci increased, and the standard errors decreased with the number of loci. However, in the presence of recombination, the estimates of the IM model parameters were more biased, with population sizes overestimated as the recombination rate increased. The splitting time appeared to approach the true value with an increasing number of loci in all cases.

The present investigation yielded findings that are congruent with those of the previous research conducted by Hey and Wang [6], indicating that certain migration rates and population sizes are subject to bias when recombination is present. However, the current study also revealed that the magnitude of such biases tended to increase with an increase in the number of loci. It is worth noting that Hey and Wang [6] reported biases in the estimate of the splitting time using up to 50 loci. In contrast, the current study found that the estimation of the splitting time since divergence was consistent and accurate with a greater number of loci.

Further investigations are required to understand the sources of errors when inferring IM models. In particular, it would be intriguing to assess the accumulated errors that arise from estimating recombination breakpoints from DNA alignments and inferring IM models from DNA alignment analyses, and to compare these results with those obtained in this study. A simulation study could...
Fig. 3. Simulation results illustrating the impact of recombination on migration rates, $m_1$ (A), $m_2$ (B), and splitting time (C) when $m_1 = m_2 = 0$ and $\theta_1 = \theta_2 = \theta_3 = 10$. Please refer to the Fig. 2 caption for the detailed labels.

Fig. 4. Simulation results illustrating the impact of recombination on isolation-with-migration model parameters. (A–F) The true parameter values of $\theta_1 = \theta_2 = \theta_3 = 10$, $m_1 = 0.1$, $m_2 = 0$ and $\tau = 10$ are indicated by gray horizontal lines. Each plot compares the results in the absence of recombination ($\rho = 0$) with those from low to high recombination rates ($\rho = 0.2$, 1, 5). Bars indicate standard errors, and the x-axis for the numbers of loci is on a log scale. Points with bars were horizontally scattered around the corresponding number of loci to minimize overlap.
provide insight into whether biases in population sizes and migration rates persist and whether the estimate of the splitting time remains consistent.

This study highlights the importance of considering the effect of recombination in IM model inference and avenues for improving the methodology. The identified biases in population sizes and migration rates serve as valuable information for developing correction methods or adjustment techniques that mitigate the bias introduced by recombination. Moreover, the study underscores the need for future models to explicitly incorporate recombination processes. By explicitly accounting for recombination in the modeling framework, researchers can capture the complexities and nuances of genome evolution more realistically. Furthermore, investigating additional factors, such as the estimation of recombination breakpoints, and integrating them into the model can lead to a more comprehensive and realistic representation of evolutionary processes. By developing correction methods, explicitly incorporating recombination, and exploring additional factors, researchers can construct better genome evolution models that capture the complexities of real-world evolutionary processes.

**ORCID**

Yujin Chung: https://orcid.org/0000-0002-3801-3961

**Conflicts of Interest**

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

**Acknowledgments**

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1C1C1011250).

**References**

Instructions for authors

Enacted January 2003
Recently revised January 9, 2019

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

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

Research and publication ethics

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

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

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

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

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

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

**Conflict of Interest Statement**

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

**Originality and Duplicate Publication**

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

**Secondary Publication**

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

- The authors have received approval from the editors of both journals (the editor concerned with the secondary publication must have access to the primary version).
- The priority for the primary publication is respected by a publication interval negotiated by editors of both journals and the authors.
- The paper for secondary publication is intended for a different group of readers; an abbreviated version could be sufficient.
- The secondary version faithfully reflects the data and interpretations of the primary version.
- The secondary version informs readers, peers, and documenting agencies that the paper has been published in whole or in part elsewhere—for example, with a note that might read, “This article is based on a study first reported in the [journal title, with full reference]”—and the secondary version cites the primary reference.
- The title of the secondary publication should indicate that it is a secondary publication (complete or abridged republication or translation) of a primary publication.

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
and reprint order form will be sent to the corresponding author. The corresponding author is responsible for communicating with the other authors about revisions and final approval of the proofs. The first proofreading is the author's responsibility, and the proof should be returned within three days from the date of receipt.

Copyrights, open access policy and open data policy

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

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

Archiving policy
It is accessible without barrier from Korea Citation Index (https://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 1. Examples of data sharing statements that fulfill ICMJE requirements**

<table>
<thead>
<tr>
<th>Element</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
<th>Example 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will individual participant data be available (including data dictionaries)?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>What data in particular will be shared?</td>
<td>All of the individual participant data collected during the trial, after deidentification.</td>
<td>Individual participant data that underlie the results reported in this article, after deidentification (text, tables, figures, and appendices).</td>
<td>Individual participant data that underlie the results reported in this article, after deidentification (text, tables, figures, and appendices).</td>
<td>Not available</td>
</tr>
<tr>
<td>What other documents will be available?</td>
<td>Study protocol, statistical analysis plan, informed consent form, clinical study report, analytic code</td>
<td>Study protocol, statistical analysis plan, analytic code</td>
<td>Study protocol</td>
<td>Not available</td>
</tr>
<tr>
<td>When will data be available (start and end dates)?</td>
<td>Immediately following publication. No end date.</td>
<td>Beginning 3 months and ending 5 years following article publication.</td>
<td>Beginning 9 months and ending 36 months following article publication.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>With whom?</td>
<td>Anyone who wishes to access the data.</td>
<td>Researchers who provide a methodologically sound proposal.</td>
<td>Investigators whose proposed use of the data has been approved by an independent review committee (&quot;learned intermediary&quot;) identified for this purpose.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>For what types of analyses?</td>
<td>Any purpose</td>
<td>To achieve aims in the approved proposal.</td>
<td>For individual participant data meta-analysis.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>By what mechanism will data be made available?</td>
<td>Data are available indefinitely at (link to be included).</td>
<td>Proposals should be directed to 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>

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

ICMJE, International Committee of Medical Journal Editors.

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

---

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

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

**Contact address**

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

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

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

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

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

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

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

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.

Research and publication misconduct

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,

Ethics Committee

1. The Ethics Committee of Genomics & Informatics shall include the Editor-in-Chief and the Ethics Editor. The other committee members shall be appointed by the Ethics Editor among the associate editors of Genomics & Informatics.
2. When the Journal faces suspected cases of research and publication misconduct, the Ethics Committee shall carefully examine the case of ethical violation by all means deemed necessary.
3. For the resolving process, the Ethics Committee shall follow a flowchart provided by the Committee on Publication Ethics (http://publicationethics.org/resources/flowcharts). The Ethics Committee shall take appropriate disciplinary action.
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.