Aims and scope

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

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

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

Published by the Korea Genome Organization

Contact information
Park, Taesung, Editor-in-Chief

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

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

© Copyright 2022, the Korea Genome Organization

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

Editor’s introduction to this issue (G&I 20:1, 2022)
Taesung Park

Review articles

Experimental development of the epigenomic library construction method to elucidate the epigenetic diversity and causal relationship between epigenome and transcriptome at a single-cell level
Kyunghyuk Park, Min Chul Jeon, Bokyung Kim, Bukyoung Cha, Jong-Il Kim

Whole-genome sequence analysis through online web interfaces: a review
A. W. A. C. W. R. Gunasekara, L. G. T. G. Rajapaksha, T. L. Tung

Original articles

Comparison of the copy-neutral loss of heterozygosity identified from whole-exome sequencing data using three different tools
Gang-Taik Lee, Yeun-Jun Chung

Presentation of potential genes and deleterious variants associated with non-syndromic hearing loss: a computational approach
Manisha Ray, Surya Narayan Rath, Saurav Sarkar, Mukund Namdev Sable

Hypothetical protein predicted to be tumor suppressor: a protein functional analysis

2-Methoxy-1,4-naphthoquinone (MNQ) regulates cancer key genes of MAPK, PI3K, and NF-κB pathways in Raji cells
Teck Yew Wong, Subramaniam Menaga, Chi-Ying F. Huang, Siong Hock Anthony Ho, Seng Chiew Gan, Yang Mooi Lim

Bayesian mixed models for longitudinal genetic data: theory, concepts, and simulation studies
Wonil Chung, Youngkwang Cho

Exploration of errors in variance caused by using the first-order approximation in Mendelian randomization
Hakin Kim, Kunhee Kim, Buhm Han

Ghada Tagorti, Bülent Kaya

Designing a novel mRNA vaccine against Vibrio harveyi infection in fish: an immunoinformatics approach
Sk Injamamul Islam, Moslema Jahan Mou, Saloa Sanjida, Muhammad Tariq, Saad Nasir, Sarower Mahfuj
Genome archive

Characterization of the first mitogenomes of the smallest fish in the world, *Paedocypris progenetica*, from peat swamp of Peninsular Malaysia, Selangor, and Perak
NorJasmin Hussin, Izzati Adilah Azmir, Yuzine Esa, Amirrudin Ahmad, Faezah Mohd Salleh, Puteri Nur Syahzanani Jahari, Kaviarasu Munian, Han Ming Gan

Application notes

TcellInflamedDetector: an R package to distinguish T cell inflamed tumor types from non–T cell inflamed tumor types
San-Duk Yang, Hyun-Seok Park

MPI-GWAS: a supercomputing-aided permutation approach for genome-wide association studies
Hyojung Paik, Yongseong Cho, Seong Beom Cho, Oh-Kyoung Kwon
Editor's introduction to this issue (G&amp;I 20:1, 2022)

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

In this issue, there are two review articles, eight original articles, one genome archive, and two application notes. In this editorial, I would like to focus on the two review articles, as well as two original articles and one application note on genome-wide association studies (GWAS). Recent rapid advances in single-cell RNA sequencing have made it possible to recognize a variety of previously unidentified subpopulations and rare cell states in tumors and the immune system based on single-cell gene expression profiles. Single-cell RNA sequencing is the topic of the first review article, by Dr. Jong-Il Kim’s group (Seoul National University College of Medicine, Korea). This review addresses the current development of methods for constructing single-cell epigenomic libraries, including multi-omics tools with important elements and additional requirements for the future, focusing on DNA methylation, chromatin accessibility, and histone post-translational modification. Single-cell epigenomic libraries help to understand the principles of comprehensive gene regulation that determine cell fate through transcripts alone and the resulting output of gene expression programs. The corresponding single-cell epigenome is expected to elucidate the mechanisms involved in the origin and maintenance of a comprehensive single-cell transcriptome. This review insightfully summarizes current research trends in the field of cellular differentiation and disease development at the single-cell level, moving toward the single-cell epigenome.

The second review, led by Dr. Tung (Dagon University, Myanmar), deals with recent developments in whole-genome sequencing technologies. While the analysis of whole-genome sequencing data requires highly sophisticated bioinformatics tools, many researchers do not have the bioinformatics capabilities to analyze the genomic data and are therefore unable to take maximum advantage of whole-genome sequencing. This review provides a practical guide on a set of bioinformatics tools available online to analyze whole-genome sequence data of bacterial genomes and presents a description of their web interfaces.

Now, I would like to turn to three articles about GWAS. The main goal of GWAS is the identification of causal variants associated with the phenotype of interest. All GWAS introduce appropriate statistical models to explain the phenotype and then to perform statistical tests. An important challenge in this post-GWAS era is to increase statistical power by using better statistical models and tests, and to investigate the causal effects between modifiable risk factors and the phenotypes via Mendelian randomization (MR).

The first article, the first author of which is Dr. Wonil Chung (Soongsil University, Korea), is about Bayesian mixed models for longitudinal genetic data. The authors proposed a Bayesian variable selection method for longitudinal genetic data using mixed models. Joint modeling of the main effects and interactions of all candidate genetic variants along with non-genetic factors leads to improved statistical power. The authors provided the
theoretical basis of the Bayesian method and evaluated its performance using data from the 1000 Genomes Project. By exploring various simulation settings, the authors showed that the proposed method tended to have higher statistical power than other existing methods. In particular, the proposed method was shown to detect well gene-time/environment interactions, which may account for some of the missing heritability.

The second article, written by Dr. Buhm Han (Seoul National University, Korea) and colleagues, is about improving the estimation of variance of causal effects in MR. When measuring causal effects between exposure and phenotype in GWAS, two-sample MR has been commonly used. The current two-sample MR uses a first-order approximation of standard error. Through simulation studies, the authors showed that this first-order approximation could lead to underestimation of variance of causal effects in MR. As a result, overestimation of power and an increased false-positive rate could occur. As an alternative, the authors proposed using the second-order approximation of the standard error to correct for the deviation of the first-order approximation and demonstrated its robust and accurate performance.

The third article on GWAS is about a supercomputing-aided approach (MPI-GWAS) to accelerate the permutation testing developed by Dr. Oh-Kyoung Kwon’s group (Korea Institute of Science and Technology Information, Korea). While permutation testing has the advantage of reducing inflated type 1 error rates, it suffers from high computational costs when applied to GWAS. With MPI-GWAS, it became possible to actionably compute a permutation-based GWAS in a reasonable amount of time, leveraging the power of parallel computing resources. I think that all three of these articles on GWAS will be good additions to the state-of-the-art knowledge on GWAS methodology.

**ORCID**

Taesung Park: https://orcid.org/0000-0002-8294-590X
Experimental development of the epigenomic library construction method to elucidate the epigenetic diversity and causal relationship between epigenome and transcriptome at a single-cell level

Kyunghyuk Park¹, Min Chul Jeon¹, Bokyung Kim², Bukyoung Cha¹, Jong-Il Kim¹,³,⁴*

¹Genomic Medicine Institute, Medical Research Center, Seoul National University, Seoul 03080, Korea
²Department of Obstetrics and Gynecology, Seoul National University Hospital, Seoul 03080, Korea
³Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul 03080, Korea
⁴Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine, Seoul 03080, Korea

The method of single-cell RNA sequencing has been rapidly developed, and numerous experiments have been conducted over the past decade. Their results allow us to recognize various subpopulations and rare cell states in tissues, tumors, and immune systems that are previously unidentified, and guide us to understand fundamental biological processes that determine cell identity based on single-cell gene expression profiles. However, it is still challenging to understand the principle of comprehensive gene regulation that determines the cell fate only with transcriptome, a consequential output of the gene expression program. To elucidate the mechanisms related to the origin and maintenance of comprehensive single-cell transcriptome, we require a corresponding single-cell epigenome, which is a differentiated information of each cell with an identical genome. This review deals with the current development of single-cell epigenomic library construction methods, including multi-omics tools with crucial factors and additional requirements in the future focusing on DNA methylation, chromatin accessibility, and histone post-translational modifications. The study of cellular differentiation and the disease occurrence at a single-cell level has taken the first step with single-cell transcriptome and is now taking the next step with single-cell epigenome.

Keywords: cellular heterogeneity, chromatin accessibility, DNA methylation, histone post-translational modifications (PTMs), single-cell epigenome, single-cell multiome

Introduction

Multicellular organisms constitute multiple types of tissues with identical or closely identical genomes. Those multiple tissues are originated and differentiated from the zygote with a systemic gene expression program of each tissue, which is comprised of multiple
types of cells. Current single-cell transcriptome study provides high resolution of the transcriptome map in a single tissue [1-3]. Previously unidentified subpopulation and rare-population of cells are observed with their gene expression profiles at a single-cell level. However, the study of single-cell transcriptome has a limitation to understand the principle and causality of comprehensive transcriptomic regulation on the chromatin, which is a complex of DNA and protein found in all eukaryotic cells [4]. ‘How are cell or tissue-specific expression patterns or framework specified and maintained with the same genome?’ and ‘How does cell or tissue retain the information of external signal even after no more signal exists over several divisions?’ The epigenetic field has introduced and focused on answering these questions. Particularly, epigenetic modifications of chromatin that include nucleosome density, DNA methylation, and histone modifications on identical genome give cells a higher cellular heterogeneity and specificity in a single tissue or single organism by the regulation of gene expression with their inheritable and reversible characteristics during cellular maintenance and differentiation (Fig. 1). In recent years, single-cell methods have been actively applied to the study of epigenetics and explain the causal correlation and maintenance of transcriptome at a single-cell level [5-7]. Finally, epigenetic analysis meets a suitable method. The single-cell epigenomics let us study cellular differentiation, including development, cellular heterogeneity among morphologically same cells, and disease progression with microenvironment deeper than the previous. Although single cell-specific information on epigenetic features had been notably demanded to study cell identity, it has been challenging to observe them at a single-cell level due to the absence of appropriate techniques and methods. With the rapid development of single-cell technology and methods in recent years, it has become possible to study epigenome at the single-cell level and understand associated transcriptome. Previous effort to make epigenetic encyclopedia [8,9] is now expanding to the single-cell method. The single-cell epigenetic encyclopedia is an ideal path for studying the characteristics of cellular heterogeneity [6,7,10].

Core Techniques and Methods for Single-Cell Epigenomic Library Construction

Single-cell epigenomic library construction requires more diverse techniques and methods in addition to single-cell RNA sequencing (Fig. 2). In a large category, physical cell isolation and barcod-
ing (Fig. 2A) and combinatorial cell barcoding (Fig. 2B) are used to label single cells, which are almost the same as RNA sequencing. Tagmentation by Tn5 transposase (Fig. 2C) improves genomic library construction by performing simultaneous fragmentation and tagmentation, performed separately in the past. Antibody-capturing protein-A fusion Tn5 (Fig. 2D) recognizes specific sites and performs tagmentation simultaneously [13], which enables us to perform a single-cell epigenomic study of histone post-translational modifications (PTMs) and implies the expansion of the method to any protein that binds chromatin. For a multimodal library construction, multi capture bead (Fig. 2E) and serial enzyme reaction with intact nuclei (Fig. 2F) enable multiome library construction for droplet-based and sci-seq methods, respectively. PTM, post-translational modification.

Fig. 2. Core techniques and methods of single-cell epigenomic library construction. Single-cell epigenomic sequencing requires more diverse techniques and methods than single-cell RNA sequencing. For single-cell barcoding, physical cell isolation and barcoding (A) and combinatorial cell barcoding (B) are required. Tagmentation by Tn5 (C) improves genomic library construction by performing simultaneous fragmentation and tagmentation, performed separately in the past. Antibody-capturing protein-A fusion Tn5 (D) recognizes specific sites and performs tagmentation simultaneously. Multi-capture beads (E) and the strategy of a serial enzyme reaction with intact nuclei (F) enable multiome library construction for droplet-based and sci-seq methods, respectively. PTM, post-translational modification.
Table 1. List of current methods of single-cell epigenomic library construction

<table>
<thead>
<tr>
<th>Method name</th>
<th>Individual cell isolation (or barcoding)</th>
<th>Cell throughput</th>
<th>Pros</th>
<th>Cons</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>scIMET&lt;sup&gt;cd&lt;/sup&gt;</td>
<td>Combinatorial indexing</td>
<td>100-1,000</td>
<td>Pooled cells input, whole-genome coverage</td>
<td>High cost for sequencing</td>
<td>[14]</td>
</tr>
<tr>
<td>scRRBS&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Manually or FACS</td>
<td>10-100</td>
<td>Low cost for sequencing</td>
<td>Restricted coverage</td>
<td>[15]</td>
</tr>
<tr>
<td>epi-gSCAR&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Manually or FACS</td>
<td>10-100</td>
<td>Free of bisulfite treatment</td>
<td>Restricted coverage</td>
<td>[16]</td>
</tr>
<tr>
<td>scXRRBS&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Manually or FACS</td>
<td>10-100</td>
<td>Extended genome coverage (than scRRBS)</td>
<td>-</td>
<td>[17]</td>
</tr>
<tr>
<td>scDNase-seq&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Combinatorial indexing</td>
<td>10-100</td>
<td>Pooled cells input</td>
<td>-</td>
<td>[18]</td>
</tr>
<tr>
<td>scATAC-seq&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Automatically by microfluidic device (Fluidigm)</td>
<td>100-1,000</td>
<td>Pooled cells input</td>
<td>-</td>
<td>[19]</td>
</tr>
<tr>
<td>scMNase-seq&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Manually or FACS</td>
<td>10-100</td>
<td></td>
<td></td>
<td>[20]</td>
</tr>
<tr>
<td>scCHIP-seq&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Automatically by microfluidic droplet chemistry</td>
<td>100-1,000</td>
<td>Pooled cells input</td>
<td>High loss of input</td>
<td>[21]</td>
</tr>
<tr>
<td>scChIC-seq&lt;sup&gt;h&lt;/sup&gt;</td>
<td>Manually or FACS</td>
<td>10-100</td>
<td>High sensitive enzyme (antibody fused MNEase)</td>
<td>-</td>
<td>[22]</td>
</tr>
<tr>
<td>iACT-seq&lt;sup&gt;h&lt;/sup&gt;</td>
<td>Manually or FACS</td>
<td>100-1,000</td>
<td>High sensitive enzyme (antibody fused Tn5)</td>
<td>-</td>
<td>[23]</td>
</tr>
<tr>
<td>scChIL-seq&lt;sup&gt;h&lt;/sup&gt;</td>
<td>Manually or FACS</td>
<td>100-1,000</td>
<td>High sensitive enzyme (antibody fused Tn5)</td>
<td>-</td>
<td>[24]</td>
</tr>
<tr>
<td>CoBATCH&lt;sup&gt;h&lt;/sup&gt;</td>
<td>Combinatorial indexing</td>
<td>1,000-10,000</td>
<td>Pooled cells input, High sensitive enzyme (antibody fused Tn5)</td>
<td>-</td>
<td>[25]</td>
</tr>
<tr>
<td>scCUT&amp;Tag&lt;sup&gt;h&lt;/sup&gt;</td>
<td>Automatically by microfluidic droplet chemistry</td>
<td>1,000-10,000</td>
<td>Pooled cells input, High sensitive enzyme (antibody fused Tn5)</td>
<td>-</td>
<td>[26]</td>
</tr>
</tbody>
</table>

<sup>m</sup>, DNA methylation; <sup>c</sup>, chromatin accessibility; <sup>n</sup>, histone PTMs; scIMET, single-cell combinatorial indexing for methylation analysis; scRRBS, single-cell reduced-representation bisulfite sequencing; epi-gSCAR, epigenomics and genomics of single cells analyzed by restriction; scXRRBS, single-cell extended-representation bisulfite sequencing; scDNase-seq, single-cell DNase-sequencing; scATAC-seq, Single-cell sequencing assay for transposase-accessible chromatin; scMNase-seq, single-cell micrococcal nuclease sequencing; scCHIP-seq, single-cell chromatin immunoprecipitation followed by sequencing; scChIC-seq, single-cell chromatin immunocleavage sequencing; iACT-seq, index multiplexing antibody-guided chromatin tagmentation sequencing; scChIL-seq, single-cell chromatin integration labelling sequencing; CoBATCH, combinatorial barcoding and targeted chromatin release; scCUT&Tag, single-cell cleavage under targets and tagmentation; FACS, fluorescence-activated cell sorting.

Table 2. List of current methods of single-cell multi-omics library construction

<table>
<thead>
<tr>
<th>Method name</th>
<th>Individual cell isolation (or barcoding)</th>
<th>Cell throughput</th>
<th>Pros</th>
<th>Cons</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>scM&amp;T-seq&lt;sup&gt;MR&lt;/sup&gt;</td>
<td>Manually or FACS</td>
<td>10-100</td>
<td>-</td>
<td>Need pre-separation of DNA and RNA</td>
<td>[27]</td>
</tr>
<tr>
<td>scNMT-seq&lt;sup&gt;MR&lt;/sup&gt;</td>
<td>Manually or FACS</td>
<td>10-100</td>
<td>Triple-omic library</td>
<td>Need pre-separation of DNA and RNA</td>
<td>[28]</td>
</tr>
<tr>
<td>scChRM-seq&lt;sup&gt;MC,R&lt;/sup&gt;</td>
<td>Manually or FACS</td>
<td>10-100</td>
<td>Triple-omic library</td>
<td>Need pre-separation of DNA and RNA</td>
<td>[29]</td>
</tr>
<tr>
<td>scTrio-seq&lt;sup&gt;MC,R&lt;/sup&gt;</td>
<td>Manually or FACS</td>
<td>10-100</td>
<td>Triple-omic library</td>
<td>Need pre-separation of DNA and RNA</td>
<td>[30]</td>
</tr>
<tr>
<td>sci-CAR&lt;sup&gt;R&lt;/sup&gt;</td>
<td>Combinatorial indexing</td>
<td>1,000-10,000</td>
<td>High throughput multi-omic library</td>
<td>-</td>
<td>[31]</td>
</tr>
<tr>
<td>Paired-seq&lt;sup&gt;C&lt;/sup&gt;</td>
<td>Combinatorial indexing</td>
<td>~1,000,000</td>
<td>High throughput multi-omic library</td>
<td>-</td>
<td>[32]</td>
</tr>
<tr>
<td>SNARE-seq&lt;sup&gt;R&lt;/sup&gt;</td>
<td>Automatically by microfluidic droplet chemistry</td>
<td>1,000-10,000</td>
<td>High throughput multi-omic library</td>
<td>-</td>
<td>[33]</td>
</tr>
</tbody>
</table>

<sup>m</sup>, DNA methylation; <sup>c</sup>, chromatin accessibility; <sup>n</sup>, copy number variation; <sup>R</sup>, RNA; scM&T-seq, single-cell DNA methylome and transcriptome sequencing method; scNMT-seq, single-cell nucleosome, DNA methylation, and transcription sequencing; scChRM-seq, single-cell chromatin accessibility, RNA barcoding, and DNA methylation sequencing; scTrio-seq, single-cell triple omics sequencing; sci-CAR, single-cell combinatorial indexing jointly profiles chromatin accessibility, and mRNA; Paired-seq, parallel analysis of individual cells for RNA expression and DNA accessibility by sequencing; SNARE-seq, single nucleus chromatin accessibility and RNA expression sequencing; FACS, fluorescence-activated cell sorting.

Most animals have a comprehensive system of DNA methylation. Combines these core techniques and methods (Tables 1 and 2) [14-33]. We will examine the most recent development in single-cell epigenomics in each epigenetic modification (DNA methylation, chromatin accessibility, histone PTMs, multiome), including the multimodal method. DNA Methylation, a Representative Marker for Cell Identity: Hardness and Solution of Single-Cell Library Construction Method Development

Most animals have a comprehensive system of DNA methylation.
that involves the establishment, removal, maintenance, and recognition of methyl-cytosine [34,35]. Furthermore, DNA methylation is globally reprogrammed during gamete development and embryogenesis [36,37] and is highly correlated with cellular identity, including pluripotency, age, and various diseases, particularly cancers [38-45]. Therefore, the study of single-cell methylome had been extensively demanded to observe different states of the methylome. Previous bulk sequencing methods of DNA methylation had already demonstrated cellular heterogeneity of DNA methylation. The percent level of DNA methylation per site rarely appears in 0 or 100. This infers that the methylation of the same DNA region from various cells in a single methylome can have a different methylation state. Furthermore, a comparison of tissue-specific methylome showed tissue specificity of DNA methylation [9,46-49]. These results strongly imply cellular heterogeneity of methylome. Therefore, developing the method for single-cell methylome construction is highly required. However, there are two practical obstacles. The first obstacle is the harsh chemical treatment of DNA. The gold standard of DNA methylation library construction requires chemical preprocessing of the genomic DNA, known as a bisulfite treatment, which converts cytosine to uracil by hydrolytic deamination. At the same time, methyl-cytosine remains unaffected [50]. In the following steps, uracil is amplified and sequenced as thymine. Therefore, the bisulfite library allows the discrimination of methyl-cytosine from unmethylated cytosine at a single base resolution. However, bisulfite treatment for sufficient cytosine to thymine conversion results in DNA loss, fragmentation, and biased sequencing data simultaneously [51-53]. Consequently, due to the DNA loss, whole genome amplification for single-cell methylome construction is required after bisulfite treatment [54-57]. Recently, the bisulfite-free method utilizing methylation-sensitive restriction enzyme has been developed for single-cell methylome (epigenomics and genomics of single cells analyzed by restriction [16], epigenomics, and genomics of single cells analyzed by restriction). However, the region of analysis is restricted to enzyme recognition sites. The second practical obstacle is the cost of sequencing that increases significantly in proportion as the number of cells increases. Since DNA methylation is observed in most of the genome, methylome analysis targets the whole genome, unlike the transcriptome analysis, which only targets mRNA sequences. Although the single-cell reduced-representation bisulfite sequencing (scRRBS) method [15] is optimal to overcome this obstacle, it does not efficiently examine a large number of critical regulatory elements in mammalian genomes. A recent method of extended-representation bisulfite sequencing (XRBS) is performed at a single-cell level with enriching informative methylation profile in promoters, enhancers and, CTCF binding sites [17]. In terms of single-cell methylome data, embryonic stem cells displayed cellular heterogeneity of DNA methylation [54] and the single-cell methylome with combinatorial cellular barcoding discriminated cellular identity by methylome [14]. The epi-gSCAR also showed cellular heterogeneity of DNA methylation by obtaining 506,063 CpG methylation variants from single acute myeloid leukemia-derived cells [16]. Single-cell XRBS has also sampled leukemia cells and featured methylation variability across individual cells and the highest cell-to-cell methylation variability in heterochromatic regions with the tri-methylation mark at the lysine residue of histone 3 (H3K9me3) [17]. All studies of single-cell DNA methylome show apparent cellular heterogeneity. Although none of the single-cell methylome method has resolved both obstacles of sample loss from harsh chemical treatment and high cost as of now, we can still collect single-cell methylome data with the current method (Table 1). Therefore, a data accumulation for understanding the meaning of DNA methylation heterogeneity and the development of the method should be considered together.

Chromatin Accessibility Providing Binding Sites for Transcription Factors: Single-Cell Library Construction Methods and Their Research Outputs

Nucleosomes comprising histones are found in the nuclei of all eukaryotic cells. Interestingly, chromatin structure with nucleosome shall consist of two distinct structural states: the first one is heterochromatin, which is highly compacted and less accessible to DNA binding proteins than other chromatin regions, and the other one is euchromatin, which is loosely packed and less intense than heterochromatin. The chromatin accessibility of cis-regulatory regions, such as enhancers and promoters around the transcription start site, is crucial to gene regulation, which regulates the binding of various proteins and interacts with other epigenetic markers, including DNA methylation, histone modifications, and non-coding RNA. Two methods of DNase I hypersensitive site sequencing (DNase-seq) [58] and assay for transposase-accessible chromatin sequencing (ATAC-seq) (Fig. 2) [59] have been performed for numerous studies with bulk samples over the past decade and have now being extended to single-cell experiments. Early studies of one single-cell DNase-seq (scDNase-seq) and two of the single-cell ATAC-seq (scATAC-seq) [18,19,60] used physical compartmentalization into each well, combinatorial cellular indexing, and microfluidics for barcoding single cells, respectively (Fig. 2). They clearly showed the cellular variation of chromatin accessi-
Histone PTMs, Markers for the Active and Repressive Transcriptional Status of the Genes: The Latest Development of the Single-Cell Library Construction Methods and Their Research Outputs

The nucleosome consists of 147 base pairs of DNA wound around histone octamers, which is a fundamental subunit of chromatin inside the nucleus. Two copies of each histone protein H2A, H2B, H3, and H4 compose a single nucleosome. Numerous studies have confirmed that chemical modifications of the amino-terminal tails of histone proteins influence transcription and show a correlation with chromatin accessibility. This regulation is also involved in a complex interplay with DNA methylation [63]. DNA methylation generally shows a higher correlation with various H3 methylation states than the DNA sequence [64]. Each histone modification displayed distinct interactions with DNA methylation. Among multiple modifications, methylation of lysine 4, 9, and 27 of H3 (H3K4me, H3K9me, H3K27me) and acetylation of lysine 27 of H3 (H3K27ac) are extensively studied due to their strong correlation with transcriptional states and inheritable characteristics during cell division. Tri-methylation of H3K4 (H3K4me3) is a hallmark of a transcriptionally permissive state enriched in promoter regions of active genes. In contrast, tri-methylation of H3K9 (H3K9me3) and H3K27 (H3K27me3) are representative repressive histone PTMs. H3K27 acetylation (H3K27ac) is an active enhancer mark enriched in the transcription start site’s proximal and distal regions. The method of chromatin immunoprecipitation followed by sequencing (ChiP-seq) had been the gold standard to study extensive PTMs-DNA interactions [65-68] with the bulk sample. This method has been expanded to single-cell omics to explore cellular heterogeneity of histone PTMs of a mixture of mouse embryonic stem cells, fibroblasts, and hematopoietic progenitors [21]. This study applied drop fluidics to label each DNA sequence of single cells at the beginning of the protocol. Then those single cells were immunoprecipitated with specific antibodies in the presence of cell barcode. Recently, preprocessing of cells with tagmentation containing antibody reaction was introduced. Various methods that do not rely on immunoprecipitation procedure by using fused MNase with protein A (single-cell chromatin immunocleavage sequencing [scChIC-seq] [22]) or fused Tn5 transposome with protein A (index multiplexing antibody-guided chromatin tagmentation sequencing [23], single-cell chromatin integration labelling sequencing [24], combinatorial barcoding and targeted chromatin release [CoBATCH] [25], single-cell cleavage under targets and tagmentation (scCUT&Tag) [26,69]) improve DNA fragment recovery and reads per cell. Notably, the scCUT&Tag method was used for histone PTMs and transcription factors at the single-cell level [69]. Likewise, CoBATCH method was also used for polymerase alongside histone PTMs [25]. Analysis of single-cell regulatory elements, including binding sites of transcription factors and polymerase with histone PTMs, further enhances our understanding of transcriptional regulation regarding cellular heterogeneity. Furthermore, those methods can be adapted for any DNA and chromatin binding proteins at a single-cell level. Notably, DNA methylation analysis can also be done with those methods by utilizing antibody for methyl-cytosine. The entire study demonstrated above observed cellular heterogeneity of histone PTMs of targeted tissues. For a concrete example, the recent scCUT&Tag profile sufficiently determined cell identity by histone PTMs and showed the regulatory feature of promoter bivalency of active (H3K4me3) and repressive (H3K27me3) marks, spreading of H3K4me3 and promoter- enhancer connectivity of the mouse central nervous system [69].
Multiome, Single-Cell Epigenome Library with Transcriptome: Current Development of the Methods and Their New Research Findings with a Future Direction

Numerous studies have demonstrated that gene expression is maintained and changed based on epigenetic information [7,63,70]. Therefore, multi-omics library construction of parallel epigenome and transcriptome (epi-RNA multiome) had been highly demanded because hierarchy and correlation between them can be observed directly through a multi-omics analysis. It particularly requires the method and technique of single-cell owing to their cellular variability. Optimization of demanding methods and techniques for single-cell multi-omics library construction is quite challenging. It requires either the preemptive step of physically separating DNA and RNA molecules from single cells or serial (or dual) enzyme reactions in identical single cells. For example, the single-cell DNA methylome and transcriptome sequencing method (scM&T-seq) physically separated RNA molecules by bead-captured oligo dT primer from DNA of a respective single cell [27]. Single-cell nucleosome, DNA methylation, and transcription sequencing (scNMN-seq), triple multiome also use the same physical compartmentalization of DNA and RNA molecules [28]. Both methods sampled mouse embryonic stem cells and showed links between DNA methylation and transcription, all three molecular layers, and dynamics coupling during differentiation. In particular, multiome analysis of mouse gastrulation utilizing scNMN-seq indicated important change and correlation by the temporal sampling of embryos [71]. Strikingly, mesoderm and endoderm showed global epigenetic change at enhancer regions driven by ten-eleven translocation-mediated demethylation and concomitant increase of accessibility. In contrast, ectoderm’s methylation and global chromatin accessibility are already established in the early epiblast [71]. Furthermore, this study featured regulatory elements associated with different states of primed or remodeled of three primary germ layers with subsequent gene expression profiles in detail, which is an ideal example of what a multiome study is aiming for. Another triple multiome, scChRM-seq (single-cell chromatin accessibility, RNA barcoding, and DNA methylation sequencing), provided a detailed map of the methylome, chromatin accessibility, and transcriptome in growing human oocytes [29]. They observed a global de novo DNA methylation setting that correlates with chromatin accessibility during human oocyte growth. The scTrío-seq (single-cell triple omics sequencing) performed triple multiome of the genome, methylome, and transcriptome and indicated that the copy number variations (CNVs) cause proportional changes in transcription. In contrast, CNVs do not affect DNA methylation in the same regions with an individual mammalian cell [30] and subpopulation within human hepatocellular carcinomas. Despite brilliant methodology and outputs of sc-M&T-seq, scNMN-seq, scChRM-seq, and scTrío-seq, their cell throughput is limited from tens to hundreds due to the requirement of laborious manual separation of each cell and physical separation of DNA and RNA molecules before each enzyme reaction. Single-cell combinatorial indexing jointly profiles chromatin accessibility, and mRNA (sci-CAR [31]) increased the cell throughput to thousands of levels, and the method aimed at simultaneous RNA- and ATAC-seq. The sci-CAR is the variant method of single-cell combinatorial indexing RNA sequencing [3,72,73]. All sci-named methods use smart combinatorial cell barcoding, enabling millions of cell throughput and making cell compartmentalization unnecessary. The sci-CAR method generated the multiome data of thousands of cells, which was a substantially higher cell throughput than other multiome methods stated above. The interesting feature of the sci-CAR protocol is the serial enzyme treatment with fresh or fixed nuclei. It implies that the different enzymes can be incubated with nuclei serially, and different pools of DNA and RNA molecules are simultaneously utilized to be a final single multiome library. Regarding the data, sci-CAR reconstructed the chromatin accessibility profiles of mouse kidney cell types with the transcriptome. Although sparsity of resulting data, particularly concerning chromatin accessibility of sci-CAR, was still required for improvement, it provided researchers the idea to develop a following multiome that needs multiple enzyme reactions with high cell throughput. Indeed, the parallel analysis of individual cells for RNA expression and DNA accessibility by sequencing (Paired-seq) was developed based on the sci-CAR methodology to increase cell throughput [32]. Interestingly, tagmentation reaction preceded reverse transcription (RT) in the Paired-seq protocol, whereas RT precedes tagmentation in the sci-CAR protocol. In addition, Paired-seq includes three more rounds of combinatorial barcoding than sci-CAR so that the cell throughput increased to millions of cells. Another method of droplet-based single nucleus chromatin accessibility and RNA expression sequencing (SNARE-seq) [33]) utilizes dual capture beads for tagmented DNA and mRNA [12]. Both methods demonstrated the transcriptome and chromatin accessibility of major and rare cell populations and pinpointed lineage-specific accessible sites of rare cells during mouse neurogenesis. Since multiome directly can show the relationship between epigenetic molecules and transcripts, we can construct the map of gene expression profiles with their causal landscape of the genome at a single-cell level. However, only high cell through-
put (> 1,000) epi-RNA multiome has been developed for chromatin accessibility (Table 2). The high cell throughput epi-RNA multiome for DNA methylation and histone PTMs are also highly demanded to understand comprehensive regulation of gene expression programs.

**Conclusion**

The major epigenetic features such as chromatin accessibility, DNA methylation, and histone modifications show clear cross-relationships and provide cellular identity controlling gene expression landscape. Epigenetic features above have to be observed in each cell owing to their strong characteristics of cellular heterogeneity. The development of single-cell genomics has evolved rapidly over the past decade with technological diversity. However, the development of single-cell epigenomics is slower due to the need for appropriate techniques or optimized methods for each epigenetic modification. Recently, optimization and methods for viewing various single-cell epigenetic changes have been developed in multiple ways and suggest relevant ideas for the new method. Alongside the single-cell epigenomic method, the spatial epigenomic method has also been developed recently [74]. These two methods share the same purpose of understanding the complex interrelationships within a single organism, tissue, tumor, and biological system. In addition, the data of two methods merged in the case of the transcriptome in that spatial transcriptome guides where the cell populations are, and single-cell transcriptome adds high resolution of that information at the single-cell level. This merging analysis strategy can be applied equally to the epigenome analysis. Single-cell RNA genomics has identified cell subpopulations within numerous tissues, and single-cell epigenomics will show their regulatory landscape of the genome at the single-cell level. This integrated analysis leads us to answer the question of ‘How are cell or tissue-specific expression patterns or framework specified and maintained with the same genome?’ with critical epigenetic information for cellular differentiation and the occurrence of various diseases.

**Authors’ Contribution**

Conceptualization: KP. Data curation: KP, BK, BC. Formal analysis: KP. Funding acquisition: JIK. Methodology: KP. Writing - original draft: KP, MCJ, BK. Writing - review & editing: JIK, MCJ, KP.

**Conflicts of Interest**

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

**Acknowledgments**

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2020R1A6A1A03047972)

**References**


5:5619.


Whole-genome sequence analysis through online web interfaces: a review

A. W. A. C. W. R. Gunasekara¹, L. G. T. G. Rajapaksha¹, T. L. Tung²*

¹Veterinary Medical Center and College of Veterinary Medicine, Jeonbuk National University, Jeonju 54596, Korea
²Department of Botany, Dagon University, 11422 Yangon, Myanmar

The recent development of whole-genome sequencing technologies paved the way for understanding the genomes of microorganisms. Every whole-genome sequencing (WGS) project requires a considerable cost and a massive effort to address the questions at hand. The final step of WGS is data analysis. The analysis of whole-genome sequence is dependent on highly sophisticated bioinformatics tools that the research personal have to buy. However, many laboratories and research institutions do not have the bioinformatics capabilities to analyze the genomic data and therefore, are unable to take maximum advantage of whole-genome sequencing. In this aspect, this study provides a guide for research persons on a set of bioinformatics tools available online that can be used to analyze whole-genome sequence data of bacterial genomes. The web interfaces described here have many advantages and, in most cases exempting the need for costly analysis tools and intensive computing resources.

Keywords: average nucleotide identities, online web servers, single nucleotide polymorphisms, virulence factors, whole-genome sequencing

Introduction

The development of DNA sequencing has revolutionized the idea of a genome and the knowledge of genes. These technologies have a dynamic history, which occurred within the last few decades. In brief, whole-genome shotgun techniques were first applied in 1979 for small size genomes ranging from 4,000–7,000 bp in experimental levels followed by a well-established DNA sequencing technique called “Sanger sequencing” which developed in the 1980s [1,2]. Rapid developments of the DNA sequencing techniques made it possible for automated sequencing in the 1990s, which allowed the first fully sequenced genome, Haemophilus influenzae in 1995 [3]. Later, around 2003 the sequencing of the entire human genome was completed [4,5]. Since then, numerous sequencing methods have been developed and they have evolved into a commercial platform called NGS or next-generation sequencing. Among many NGS technologies available, whole-genome sequencing (WGS) is involved with the determination of the entire DNA sequence from an organism’s genome at a single time [6]. It involves identifying the nucleotide arrangement of a complete genome of an organism, which is supported by automatic DNA sequencing methods and computational techniques that facilitates the assembly of millions of small DNA fragments [7]. Today, the advances and extensive use of NGS techniques have greatly affected the progress of the scientific research field.
Early WGS methods were expensive, difficult to perform, and time-consuming, especially in the developmental era of genomic data [8]. A decade ago, high-quality reference genome sequences were only available for a model or well-studied organisms. Today, the implementation of WGS facilitates a better understanding of the genomic functions in an organism and its expression mechanisms. Moreover, WGS provides much more comprehensive information on various genes by sequencing the noncoding DNA regions, which captures 95%–99% of the genome. The information gained through WGS has proven to be very useful in terms of understanding the origins of pathogenic microorganisms, their transmission routes, and in public health management [9,10]. Genome-wide approaches enhance the power and resolution for the above-mentioned applications and improve the reliability of conclusions.

There is no doubt that every WGS project needs a considerable cost and effort to address the questions at hand. However, the analysis of WGS data highly depends on sophisticated bioinformatics tools. Many laboratories and research institutions do not have the bioinformatics capabilities to analyze the large amount of genomic data generated through sequencing and therefore are unable to take maximum advantage of WGS [11]. The goal of this study is to provide a guide for research personnel on bioinformatics tools available online that are needed to interpret WGS data and, how these online web interfaces can be applied to bacterial genome analysis settings easily, affordably, and, in most cases, without the need for intensive computing resources and infrastructure. Moreover, in this article, we discuss how to utilize genomic annotation servers, classical multilocus sequence typing (MLST), whole-genome MLST (wgMLST), single nucleotide polymorphisms (SNPs), average nucleotide identity (ANIs), prophages, cluster of orthologous groups (COG), virulence factors, and, genomic mapping tools for bacterial WGS data analysis. There is still much work that needs to be done for the development of online web interfaces to improve data quality and its applications in WGS. Consequently, it is necessary to develop more advanced and efficient data analysis pipelines for processing and analyzing whole genomes.

**General Workflow of WGS**

Several steps are involved in a bacterial WGS project. First, a biological sample (bacteria) is collected and cultured on appropriate media. The DNA is extracted by using commercial DNA extraction kits and/or by manual DNA extraction methods. The DNA quality is usually measured through the qubit meter. Following this, a DNA library is prepared. Once the DNA library is prepared, sequencing can be performed in any WGS machine (such as Illumina/ion torrent) as the researcher’s requirements. Millions of short sequence reads are produced as the final result, typically a few hundred nucleotides long or less. After sequencing, raw reads will be trimmed to remove adapter and low-quality reads. By using these reads, the novel genome can be reconstructed with or without using a reference sequence. In reference-based reconstruction, the short reads are aligned to a closely related reference genome, which has a complete genomic representation. It is important to note that all the reads will not align with the reference genome (there can be some novel regions in the genome of interest that are absent in the reference genome). Sites with problematic nucleotide compositions also can be filtered out. As an alternative for reference mapping, *de novo* assembly can be performed. Here, all the short reads are aligned to each other (known as contigs) without the use of a reference sequence. The number of contigs produced depends on the total number of short-read DNA sequences in hand. Following reconstruction, the novel genomes can be analyzed through online web interfaces as described below (Fig. 1).

**The Genomic Annotation**

Once the assembly of a bacterial genome is completed, the next important step is genomic annotation. Simply it refers to the identification of functional/non-functional genomic segments and/or open reading frames and matching them to other reference genome sequences in an existing database [12]. A typical genomic annotation must include biological information such as gene models and gene functions and their protein products [13]. The annotation of a genome is depending on a set of rules guided by the annotation pipeline. Hence, the quality of the annotation always relies on the quality of the genome assembly [14]. Apart from the NCBI prokaryotic genome annotation server (PGAP), rapid subsystem annotation using subsystem technology or RAST annotation (http://rast.theseed.org/) is the most common pipeline available online for bacterial genome annotation [15]. Aside from subsystem statistics, the RAST annotation server is capable of providing metabolic construction along with functional, sequence, and KEGG database pathways (Kyoto Encyclopedia of Genes and Genomes database) through the annotation of a respective genome. Depending on the job load, annotation time for a genome can be varying. Final output data is available in various types of file formats which is very important for further analysis of genomes (Fig. 2).
Classical MLST and SNP Calling

Correct, standardized identification is a basic need for any researcher working with bacteria, whether it’s a pathogen, commensals, or used for industrial purposes. For a long time, MLST has been considered as the “gold standard” for bacterial classification, and has been used widely for molecular studies [16]. Classical MLST or multilocus sequence typing is a technique that usually depends on seven housekeeping genes that reside in the bacterial genome [17]. The unique sequences of housekeeping genes in bacteria are assigned to a random integer number, in order to assign a unique genome profile (also known as allelic profile) which specifies its sequence type (ST). Since the ST is universal, the data collected through MLST has proven to be useful in characterizing bacterial isolates of different epidemiological origins [18]. To date, the PubMLST server is considered the most popular database on the internet related to MLST [19]. Finding a housekeeping gene sequence from bacterial WGS data can be time-consuming. Apart from PubMLST, various easy-to-handle online servers are available with the capability to identify classical MLST genes directly from a whole-genome sequence.

In a bacterial genome, analyzing SNP is considered as an important step in terms of understanding genomic relationships. The SNPs are the mirror showing how far your genome is divergent from other reference strains. In a typical bacterial genome, the presence of a small number of SNPs indicates that they are genetically similar and can be originated from a common ancestor [20]. Sometimes when isolates are distant in time or geographical origin, a large number of SNPs are present in between the respective genomes, indicating that they did not originate from the same source and/or they have been gone through evolution for a longer period [16]. Hence, the SNP base similarities and differences allow researchers to trace the transmission patterns of pathogenic organisms worldwide [21].

The center for genomic epidemiology (https://cge.cbs.dtu.dk/services/) provides both classical MLST and SNP analysis of WGS. This server offers a comprehensible researcher friendly platform. The MLST scheme in the CGE server is associated with the PubMLST database [22]. Once the bacterial genomic data is uploaded to the server, each allelic number/their sequences representing housekeeping genes and ST can be obtained within a few minutes. The SNP analysis in the CGE server depends on a set of parameters selected by the user [23]. To analyze the SNP variations, it is necessary to upload the reference genome along with the genomes of interest. Most importantly this server is capable of producing an SNP base phylogenetic tree with evolutionary distances, and it is available in several file formats. As a result, users can modify the phylogenetic trees according to their requirements (Fig. 3).

Fig. 1. General overview of high throughput sequencing workflow of a bacterial genome. Following genome assembly, online web interfaces can be utilized for the purpose of analyzing WGS. MLST, multi locus sequence typing; WGS, whole-genome sequencing.
Many researchers suggested, previously mentioned classical MLST scheme doesn’t provide a higher resolution of bacterial genomes when compared to the large number of DNA sequences available in hand [10,24]. On this aspect extended versions of the classical MLST scheme have been developed. Besides, many researchers focus on identifying differences in genes present in bacterial genomes. Studying differences of genes is a key determinant to understanding virulence and pathogenicity among different bacterial strains [25-27]. The newly developed whole-genome MLST or wgMLST tools enable the recognition of genetic variations among bacterial pathogens with high accuracy [28,29]. The online web interface called cano-wgMLST (http://baccompare.imst.nsysu.edu.tw/index.php) can be used as a primary tool to identify the differences between genes and/or similarities among genomes (Fig. 4). This server provides a phylogenetic tree, heat map as well as the percentage of gene occurrence among respective genomes. The phylogenetic tree is constructed based on the core genome and highly discriminatory genes [29].

ANI or average nucleotide identity refers to the measurement of nucleotide level similarity between two or more genomes [30].

Fig. 2. General subsystem features and KEGG pathway of drug metabolism of Vibrio parahaemolyticus 3HP_AHPND genome through RAST server (Different colors in the subsystem category distribution indicates different subsystem features whereas KEGG pathway indicates the functions for V. parahaemolyticus 3HP_AHPND genome). KEGG, Kyoto Encyclopedia of Genes and Genomes.
The ANIs exhibit genetic relatedness among bacterial strains. In the early days of genomic research, DNA-DNA hybridization is considered as the gold standard to compare nucleotide identities of bacterial genomes [31]. In parallel to the evaluation of genomic technologies, various software’s have been developed to assess the ANIs among bacterial genomes. The simplest tool that can use to calculate nucleotide level similarities is the JSpeciesWS online web server (http://jspecies.ribohost.com/jspeciesws/). The server measures the probability of multiple genomes belonging to the same species by pairwise comparisons of ANIs (Fig. 4). It is suggested that closely related bacterial species share a high rate of nucleotide similarities [32]. On researchers point, it is an important aspect since it provides capabilities to track epidemiological outbreaks [30].

**Virulence Factors, Prophages, and COGs**

Virulence factors are the properties of an organism that provide capabilities to establish itself on or within a particular host species and prompt the potential cause of the disease [26]. They are the driven forces of pathogenicity acquired by microorganisms, as a result of the long-term evaluation process. Common virulence factors of bacterial pathogens include adherence, anti-phagocytosis, chemotaxis and mortality, enzyme, iron uptake, quorum sensing, secretion systems, toxin, and immune evasion. Virulence factor database or VFDB is the most popular online server for bacterial genome-related virulence factor analysis (http://www.mgc.ac.cn/VFs/). This server allows the identification of virulence factors with structural features, mechanisms, and functions [33]. Further-
more, it is possible to analyze virulence factors in species level as well as the genus level through this server (Fig. 5A).

Prophages are the genetic materials that are inserted and integrated into bacterial chromosomes or plasmids without causing any disruption to the bacterial cell [34]. One key function of prophages is to increase the virulence potential of bacteria by horizontal gene transfer [35]. In terms of survival, prophages can give bacteria both resistance mechanisms and metabolic advantages [36]. The latest version of PHASTER (https://phaster.ca/) is an efficient, fast, and user-friendly online server in terms of prophage analysis [37]. The server provides graphical illustrations of prophages with their respective phage features. (Fig. 5B).

COG or cluster of orthologous groups is a set of proteins encoded by genomes of certain organisms related to direct evolution that are referred to be orthologous [38]. Studying COG in the recent past had a significant impact on the phylogenetic classification of proteins from microbial genomes [39]. The WebMGA (http://weizhong.lab.ucsd.edu/webMGA/) is one web interface that predicts the COGs of bacterial genomes. The data will be available as a text file based on different COG classes. Following analysis, researchers can build graphical illustrations of COG as their requirements (Fig. 5C).

Graphical Illustration of Genomes (Genome Mapping)

In general, genome mapping refers to the assignment of genes into their respective positions of the genomes [40]. To date, the majority of genomic mapping is conducted through highly sophisticated software. Difficulties of operating and high costs associated with the software lead many researchers to think twice when doing WGS projects. Several online servers are providing graphical illustrations of genomes. Representing genomic features is very important since they are the landmarks in the genome of an organism. It can effectively convey information that helps to understand the biological properties of microorganisms [41]. Also, unique information related to specific genes can be displayed in genomic maps. Furthermore, genomic maps can display sequence differences concerning a reference genome, gene expression, the positions of contigs for incomplete genomes, and the sequence coverage information. Among the limited number of online web servers developed so far, the CGview (http://stothard.afns.ualberta.ca/cgview_server/) and GView (https://server.gview.ca/) servers are widely used for graphical illustration of bacterial genomes. In these servers, parameters for a certain genomic map need to be set by the user. The CGview server provides a genomic map with distinct genomic features and through the GView server it is possible to
Fig. 5. (A) Virulence factor analysis workflow of VFDB server. (B) Prophage analysis workflow of PHASTER server. (C) Cluster of orthologous group (COG) analysis workflow of WebMGA server. *Vibrio parahaemolyticus* genome 3HPAHNPND was used as a reference for all the applications.
analyze multiple genomes at once and generate a comparative genomic map (Fig. 6) [41].

Apart from the graphical illustrations of genomes, many researchers tend to use WGS based phylogenetic maps. The use of large-scale genomic data to generate a phylogenetic tree is impossible without analysis software and/or operating system. The WGS base phylogenetic trees lead researchers to understand evolutionary history and relationships among microorganisms [42]. There is a finite number of online servers available on this aspect. The CVTree3 (http://tlife.fudan.edu.cn/cvtree3/) is one such server that can be utilized for bacterial genomes in terms of phylogenetic tree mapping [43]. This web interface uses FAA or FFN files to produce phylogenetic trees. Annotation through the RAST server became very useful at this point. Because RAST server annotation provides FAA file as final output. The file generated through the RAST server is 100% compatible with the CVTree3. In this server, analysis of the genome highly depends on user-defined criteria. The phylogenetic trees generated through the CVTree3 server can be downloaded in various formats, which allows the researchers to modify them according to their requirements.

**Conclusion**

Combined analysis of a respective genome along with ANI, SNPs, MLST, wgMLST, virulence, prophages, and COGs through these online web servers will motivate any researcher to move forward in bacterial WGS analysis without depending on other sophisticated genomic analysis tools. These web interfaces are deemed to be fast and accurate and can be used as a confirmation guide along with epidemiological analysis, research, and surveillance.

**ORCID**

A. W. A. C. W. R. Gunasekara: https://orcid.org/0000-0002-1323-5242
L. G. T. G. Rajapaksha: https://orcid.org/0000-0002-8299-9107
T. L. Tung: https://orcid.org/0000-0002-1181-5921

**Authors’ Contribution**

Conceptualization: AWACWRG. Data curation: AWACWRG, Formal analysis: AWACWRG. Methodology: LGTGR. Writing: original draft: LGTGR. Writing - review & editing: TLT.

**Conflicts of Interest**

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

**Acknowledgments**

The authors would like to convey their thanks to software developers, who originally developed the WGS related bioinformatics tools.

**References**

4. Austin CP. The impact of the completed human genome sequence on the development of novel therapeutics for human dis-


25. Holmes RK, Jobling MG. Genetics. In: Medical Microbiology (Baron S, ed.). Galveston: University of Texas Medical Branch at Galveston, 1996.


27. Peterson JW. Bacterial pathogenesis. In: Medical Microbiology (Baron S, ed.). Galveston: University of Texas Medical Branch at Galveston, 1996.


29. Holmes RK, Jobling MG. Genetics. In: Medical Microbiology (Baron S, ed.). Galveston: University of Texas Medical Branch at Galveston, 1996.


34. Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin EV. Microbial genome analysis: the COG approach. Brief Bioin-
Comparison of the copy-neutral loss of heterozygosity identified from whole-exome sequencing data using three different tools

Gang-Taik Lee¹, Yeun-Jun Chung²,³*  

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

Loss of heterozygosity (LOH) is a genomic aberration. In some cases, LOH can be generated without changing the copy number, which is called copy-neutral LOH (CN-LOH). CN-LOH frequently occurs in various human diseases, including cancer. However, the biological and clinical implications of CN-LOH for human diseases have not been well studied. In this study, we compared the performance of CN-LOH determination using three commonly used tools. For an objective comparison, we analyzed CN-LOH profiles from single-nucleotide polymorphism array data from 10 colon adenocarcinoma patients, which were used as the reference for comparison with the CN-LOHs obtained through whole-exome sequencing (WES) data of the same patients using three different analysis tools (FACETS, Nexus, and Sequenza). The majority of the CN-LOHs identified from the WES data were consistent with the reference data. However, some of the CN-LOHs identified from the WES data were not consistent between the three tools, and the consistency with the reference CN-LOH profile was also different. The Jaccard index of the CN-LOHs using FACETS (0.84 ± 0.29; mean value, 0.73) was significantly higher than that of Nexus (0.55 ± 0.29; mean value, 0.50; p = 0.02) or Sequenza (0 ± 0.41; mean value, 0.34; p = 0.04). FACETS showed the highest area under the curve value. Taken together, of the three CN-LOH analysis tools, FACETS showed the best performance in identifying CN-LOHs from The Cancer Genome Atlas colon adenocarcinoma WES data. Our results will be helpful in exploring the biological or clinical implications of CN-LOH for human diseases.

Keywords: CN-LOH, loss of heterozygosity, SNP array, whole-exome sequencing

Introduction

Loss of heterozygosity (LOH) is a large-scale genomic aberration, in which one parental allele is lost, resulting in the loss of the heterozygous state. LOH can sometimes be accompanied by an alteration of the copy number; however, most LOH events occur without changing the copy number. This is called copy-neutral LOH (CN-LOH), and it is caused by diverse genetic events such as uniparental disomy or gene conversion [1-3]. CN-LOH has effects on many diseases, such as cardiovascular disease [4], Prader-Willi
syndrome [5], and congenital adrenal hyperplasia [6]. In particular, many studies have shown that CN-LOH is a widespread event in diverse cancers, with regular patterns appearing on specific chromosomes [7,8]. The pathogenesis of CN-LOH in carcinogenesis is primarily derived from the biallelic inactivation of a tumor suppressor gene (TSG) [9]. For example, when a TSG in a chromosomal region harbors an inactivating mutation at birth in one allele, this TSG can be completely inactivated by CN-LOH as a second hit, which may induce malignant disease [10,11]. Loss of the wild-type allele of the KIT gene and duplication of gain of function mutations in oncogenes due to CN-LOH may also cause tumorigenesis [12]. Despite its potential importance, the biological and clinical implications of CN-LOH on human diseases have not been well studied.

Next-generation sequencing (NGS) technology and recent efforts to collect information on genomic alterations in cancers, such as The Cancer Genome Atlas (TCGA) and COSMIC databases, have facilitated the analysis of CN-LOH in addition to somatic mutations and alteration of gene expression [13]. Several analytical tools have been developed to detect CN-LOH events, using NGS data, such as Sequenza [14], Nexus Copy Number, and FACETS [15]. However, no study has compared the performance of these tools for identifying CN-LOH from whole-exome sequencing (WES) data, which are the most common NGS data for various diseases, including cancer. In this study, we compared the performance of CN-LOH determination by the three commonly used tools.

Methods

Data
We used 443 colon adenocarcinoma WES data entries from the TCGA database. Before CN-LOH analysis, the purity and ploidy of each tumor sample were checked using FACETS. To ensure a reliable analysis of CN-LOH, we selected samples with an estimated ploidy of 1.5–2.6 and an estimated purity above 0.8. Of the tumor samples that fit these criteria, we randomly selected 10 samples for the CN-LOH analysis in this study (Supplementary Tables 1 and 2). We also used Affymetrix SNP Array 6.0 data (CEL files) of the same samples, which were obtained from the GDC Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive/) (Supplementary Table 2). Tumor cell purity and ploidy information of the 10 samples are available in Supplementary Table 3.

Analysis of CN-LOH using the SNP array data
We first defined the CN-LOH profiles of the 10 colon adenocarcinoma samples using the SNP array data as a reference for the CN-LOH status of the 10 samples. Nexus Copy Number (Biodiscovery Inc., El Segundo, CA, USA) was used to define the LOH and copy number alteration profiles of each sample. The SNP-FASST2 segmentation algorithm was used for copy number estimation with the following parameter values: the minimum number of probes per segment was defined as 5, and the segments showing copy number differences of more than ±0.18 in the log2 scale were defined as copy gain and loss, respectively, as described elsewhere [16]. To obtain CN-LOH, we excluded instances of LOH due to copy number alteration or allelic imbalance. The CN-LOH reference data were converted from hg19 to hg38 using the Lift-Over tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) to match the genome reference with the WES data.

Analysis of CN-LOH using WES data with different tools
We defined the CN-LOH profiles of the 10 colon adenocarcinoma samples using their WES data (.bam files). Three CN-LOH analysis tools were used: Sequenza, Nexus Copy Number (hereinafter called Nexus), and FACETS. For Sequenza, we analyzed the data by referring to the default setting values following the manufacturer’s instructions. For Nexus, B-allele frequency (BAF) values were defined using the BAM (ngCGH) algorithm, and copy number estimation was performed using the SNPRank segmentation algorithm following the manufacturer’s instructions. All other conditions were implemented according to the default values provided by the manufacturer. CN-LOH analysis using FACETS was performed using the default values. Of the LOHs identified from the WES analysis, we removed the regions of allelic imbalance and copy number aberrations by using the Nexus filtering option to ensure the identification of CN-LOHs. To compare the CN-LOH calls between the reference and the three tools, we used a log-R ratio (LRR) plot and BAF plot using “Karyoploter” [17]. The probe intensity values and SNP intensity values of the reference, which were required for the LRR and BAF plots, were calculated using Nexus. In detail, the accuracy of CN-LOH calls using the three tools was estimated based on the overlapping of CN-LOH calls between each tool and the reference. If a CN-LOH region deduced from the tool overlapped with the CN-LOH region in the reference set, we defined this call as consistent with the reference regardless of the length of the overlapping region.

Statistical analysis
To estimate the similarity between the reference data and the analysis data using the three tools, we used the Jaccard index. The equation is as follows:
We also used the "ROCR" [18] R package to calculate the area under the curve (AUC), sensitivity, and specificity for each tool.

**Results**

We analyzed the CN-LOH profiles using the SNP array data of the 10 colon adenocarcinoma patients. In total, 25 CN-LOH events were identified from the 10 patients according to the reference data (Table 1, Fig. 1). The average number of the reference CN-LOH events identified from the SNP array data was 2.5 per sample (range, 0 to 6), the average length was 44.5 Mb (range, 1.5 to 136.4 Mb), and the median length was 30 Mb.

We next analyzed the CN-LOH profiles of the WES data from the 10 colon adenocarcinomas using the three analysis tools (Nexus, FACETS, and Sequenza) (Table 1, Fig. 1). The average number of CN-LOH events identified from the TCGA WES data was 4.7 (range, 0 to 17), 10.3 (range, 0 to 25), and 11.1 (range, 1 to 41) for FACETS, Nexus, and Sequenza, respectively. The average size of the CN-LOH events was 35.3 Mb, 10.9 Mb, and 18.9 Mb according to FACETS, Nexus, and Sequenza, respectively. The details of the CN-LOH profiles of the 10 samples using the WES data with the three analysis tools are available in Supplementary Table 4.

Most CN-LOHs identified from TCGA WES data using the three tools were consistent with the reference data identified from SNP array data. Of the 25 reference CN-LOHs, 24 (96%) were consistently detected by FACETS. The consistency of the CN-LOHs identified by Nexus and Sequenza with the reference profile was lower than that of FACETS: 76% (19/25) for Nexus and 28% (7/25) for Sequenza, respectively. Examples of CN-LOHs consistently or inconsistently defined between the reference data and the three tools are illustrated in Fig. 2.

Some of the CN-LOHs identified from the TCGA WES data were not consistent between the three tools, and the consistency with the reference CN-LOH profile was also different. In the TCGA-SS-A7HO-01A sample (Fig. 2B), there was a copy loss in the q-arm of chromosome 9, and a small (91,605,233–93,165,795 bp) CN-LOH event occurred inside this copy loss area. This CN-LOH event identified by SNP array data was consistently detected by FACETS, but was not identified by Nexus or Sequenza. In the TCGA-AA-3854-01A sample (Fig. 2D), the reference and FACETS detected two CN-LOH events on chromosome 22, and Sequenza also detected these CN-LOHs; however, Sequenza could not discriminate the small homozygous

### Table 1. CN-LOHs identified from the 10 colon adenocarcinoma patients

<table>
<thead>
<tr>
<th>Samples</th>
<th>Reference</th>
<th>FACETS</th>
<th>Nexus</th>
<th>Sequenza</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCGA-4N-A93T-01A</td>
<td>2</td>
<td>3</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>TCGA-A6-2677-01A</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>TCGA-A6-6652-01A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>TCGA-AA-3655-01A</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>TCGA-AA-3848-01A</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>TCGA-AA-3854-01A</td>
<td>5</td>
<td>7</td>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>TCGA-CM-5862-01A</td>
<td>3</td>
<td>4</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>TCGA-CK-6746-01A</td>
<td>1</td>
<td>17</td>
<td>22</td>
<td>41</td>
</tr>
<tr>
<td>TCGA-QG-A91Y-01A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>TCGA-SS-A7HO-01A</td>
<td>6</td>
<td>6</td>
<td>25</td>
<td>14</td>
</tr>
<tr>
<td>Total No. of CN-LOH</td>
<td>25</td>
<td>47</td>
<td>103</td>
<td>111</td>
</tr>
<tr>
<td>Average No. of CN-LOH</td>
<td>2.5</td>
<td>4.7</td>
<td>10.3</td>
<td>11.1</td>
</tr>
<tr>
<td>Minimal length (bp)</td>
<td>1,560,563</td>
<td>704,899</td>
<td>2,028,231</td>
<td>7,655</td>
</tr>
<tr>
<td>Maximal length (bp)</td>
<td>136,468,365</td>
<td>181,230,348</td>
<td>71,438,575</td>
<td>137,587,418</td>
</tr>
<tr>
<td>Mean length (bp)</td>
<td>44,481,811</td>
<td>35,328,191</td>
<td>10,864,090</td>
<td>18,883,527</td>
</tr>
<tr>
<td>Median length (bp)</td>
<td>30,037,847</td>
<td>15,273,966</td>
<td>7,929,436</td>
<td>4,336,034</td>
</tr>
</tbody>
</table>


https://doi.org/10.5808/gi.21066
Fig. 1. CN-LOHs identified from the 10 colon adenocarcinoma patients. The upper, middle, and lower plots of each sample represent the LRR, BAF plot, and the CN-LOHs identified from each tool, respectively. In the lower plot, reference CN-LOHs were defined from the SNP array data of the 10 colon adenocarcinoma patients. The CN-LOHs of the three tools were defined from the TCGA WES data of the same patients. The orange regions represent the CN-LOH area. CN-LOH, copy-neutral loss of heterogeneity; LRR, log-R ratio plot; BAF, B-allele frequency; SNP, single-nucleotide polymorphism; TCGA, The Cancer Genome Atlas; WES, whole-exome sequencing.
deletion in this region (25,662,836–25,920,518 bp). Nexus could not identify the CN-LOHs in this case.

For an objective comparison of the CN-LOH analysis tools, we analyzed the Jaccard index, which represents the similarity of CN-LOH calls between the tools and the reference data, as described in the Methods section. The median value of the Jaccard index of FACETS (0.84 ± 0.29; mean value, 0.73) was significantly higher than those of Nexus (0.55 ± 0.29; mean value, 0.50; p = 0.02) and Sequenza (0 ± 0.41; mean value, 0.34; p = 0.04) (Fig. 3A and 3B). Nexus showed a better Jaccard index than Sequenza, but the difference was not statistically significant. As another method of verification, we performed receiver operating characteristic analyses for the three tools. In the receiver operating characteristic analysis, FACETS showed the best identification power (AUC, 0.83; sensitivity, 0.7714; specificity, 0.8854) and Sequenza (AUC, 0.55; sensitivity, 0.4300; specificity, 0.6785) (Fig. 3C). Collectively, these results suggest that FACETS may be the most suitable tool for identifying CN-LOHs from WES data.

We next analyzed the CN-LOH profiles of the 398 TCGA colon adenocarcinoma WES data entries using FACETS under the conditions used in this study to identify CN-LOH. Chromosome 17p (17p13.1) showed the highest frequency of CN-LOHs (28.1%), followed by chromosomes 18 and 5q (Fig. 4).

Discussion

It is well known that copy number alterations can induce over-activation or inactivation of cancer-related genes, which consequently contribute to tumorigenesis or the progression of cancer. Howev-
Fig. 3. Comparison of the CN-LOH analysis tools. (A) Box plot of the Jaccard index for the three tools, with p-values are provided above the brackets. (B) Jaccard index table for each sample analyzed using the three tools. (C) ROC analyses for the three tools. CN-LOH, copy-neutral loss of heterogeneity; ROC, receiver operating characteristic.

Fig. 4. The genome-wide frequencies of CN-LOHs in the TCGA colon adenocarcinomas. TCGA colon adenocarcinoma WES data (n = 398) were analyzed using FACETS under the conditions used in this study to identify CN-LOHs. CN-LOH, copy-neutral loss of heterogeneity; TCGA, The Cancer Genome Atlas; WES, whole-exome sequencing.

The biological implications of CN-LOH for tumorigenesis have not been well studied. A reason for this is the difficulties in identifying CN-LOH, especially with WES data. Although several algorithms have been developed, there is no global standard method for the CN-LOH analysis using WES data. To detect genome-wide CN-LOHs, WES data would be relatively disadvantageous because the exonic region is a small part of the whole genome. However, despite this disadvantage, CN-LOH identification from WES data would be valuable for understanding the pathogenesis of diseases, because WES is a common source of data for diverse diseases, including cancer.

For an objective comparison of the CN-LOH identification performance of the three tools, we used TCGA colon adenocarcinoma WES data because Affymetrix SNP Array 6.0 data of the same
samples were available in the GDC Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive/). To verify instances of CN-LOH, we selected samples with an estimated ploidy of 1.5–2.6, which suggests that the majority of the genome would have a copy-neutral status. Furthermore, to minimize the effect of normal cell contamination, we selected samples with an estimated purity > 0.8. For an objective comparison of the tools, we used CN-LOHs from the SNP array data of the same colon adenocarcinomas as reference data.

FACETS showed greater consistency of CN-LOH calls with the reference data than Nexus and Sequenza. Accordingly, the Jaccard index of FACETS was significantly higher than those of Nexus and Sequenza. The Jaccard index is a coefficient that shows the degree of similarity between different datasets [19]; therefore, a higher Jaccard index (converging to 1) indicates greater similarity between different groups. Taken together, our data suggest that FACETS would be suitable to identify CN-LOHs from WES data.

When we applied this condition to the TCGA colon adenocarcinoma WES data, chromosomes 17p, 18, and 5q showed frequent CN-LOH events, which is largely consistent with previous reports of the CN-LOH profiles in colon adenocarcinomas using 63 sets of whole-genome sequencing data [20].

There are several limitations of this study. First, we used the default settings of the tools, which might not have been the best conditions to obtain maximally reliable CN-LOH profiles. Second, we did not explore the reasons for the inconsistency of CN-LOH calls between the three tools in this study. Third, the quality of the WES data was most likely different among the 10 samples, which may have affected the reliability of CN-LOH calling.

In conclusion, of the three CN-LOH analysis tools, FACETS showed the best performance in identifying CN-LOHs from TCGA colon adenocarcinoma WES data. Our results may be helpful for exploring the biological or clinical implications of CN-LOH for human diseases.

**ORCID**

Gang-Taik Lee: https://orcid.org/0000-0003-4391-8850
Yeun-Jun Chung: https://orcid.org/0000-0002-6943-5948

**Authors’ Contribution**

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

**Conflicts of Interest**

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

**Acknowledgments**

This work was supported by a grant from the National Research Foundation of Korea (2017M3C9A6047615 and 2019R1 A5A2027588). We thank KREONET (Korea Research Environment Open NETwork) and KISTI (Korea Institute of Science and Technology Information) for allowing us to use their network infrastructure.

**Supplementary Materials**

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

**References**


Introduction

Hearing impairment is the most common neurosensory disorder [1] in humans. As of global statistics, approximately 360 million people have been suffered from hearing loss (HL) [2]. As a result, HL is reported as the fourth leading cause of disability in humans worldwide [3]. Meanwhile, congenital HL appears as the most prevalent chronic condition and defect in children [4,5]. The growing cases of hearing impairment in newborn are also reported by the universal newborn hearing screening board [5]. It is also identified that 70% of congenital HL accounts for non-syndromic hearing loss (NSHL) [6] whereas the rest 30% comes under syndromic [7]. As per literature, molecular and genetic aetiology is the prime cause of congenital HL. But, due to genetic and clinical heteroge-
neity in NSHL, the disease pathogenesis [6] is not clearly understooed. Therefore, there is a subsequent delay in the discovery of disease pathways and identification of molecular targets since 1990 [8]. Again, complexity in the identification of potential target genes and their associated variants at an accurate genomic location has created lots of diagnostic complications in NSHL [9].

The association of several important genes such as GJB2, GJB3, GJB6, SLC26A4, KCNQ4, DFNAS5, SLC26A5, MYO1A, MYO7A, MYH11A, and CDH23 has been identified with respect to the progression of congenital NSHL. Among these, GJB2 is established as an imperative target of NSHL [9]. In addition, the successful adoption of genome-wide association studies in biomedical research has discovered the genetic architecture of novel variants and their disease associations [10,11] which encourages researchers to focus on the identification of a potential biomarker for NSHL.

Generally, identification of potential biomarkers through the integration of multiple complementary approaches [12] is better than using any single method. As of evidence, many studies have been reported using more than one in silico methods such as pathway enrichment and functional annotation or protein-protein interaction (PPI) network to screen and/or discover potential target genes from a large scale gene pool in diseases like atherosclerosis, nasopharyngeal carcinoma, systemic sclerosis, Parkinson disease, papillary thyroid carcinoma, neurodegeneration, etc. [13-19]. So, to identify any putative target genes, different aspects including the study of different regulatory pathways, miRNA-based gene regulation, annotation of protein functional network, and variant analysis are need to be addressed. Particularly, the discovery of important biological pathways involved in signal transductions [20] has a great aid and value. In this connection, the study of functionally important PPI provides a meaningful direction towards understanding the involvement of disease targets in different biological processes [20]. Similarly, the discovery of miRNAs based gene regulatory networks in the post-transcriptional process has a certain importance in biomarker identification [21]. As proof, experimental studies have also shown the involvement of a few miRNAs (miR-182, miR-183, and miR-96) in human inner ear development [22-25]. In addition to this, detection of deleterious single nucleotide polymorphisms (SNPs) has also been implemented in biomarker identification to ascertain disease susceptible genes. Especially, non-synonymous SNPs in the coding region have a vital role to produce various damaging effects on protein structure, stability, charge, etc., which leads to functional dysfunction of several disease targets [26-28]. Therefore, the identification of non-synonymous SNPs within the proper coding architecture of disease targets is an important strategy of biomarker identification. The present study is focused on the discovery of potential NSHL targets and their functional variants using multiple in silico complementary approaches.

Methods

Data mining
A systematic review of literature was carried out by following the guidelines of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). The scientific articles on NSHL were searched in PubMed (https://pubmed.ncbi.nlm.nih.gov/), Science Direct (https://www.sciencedirect.com/), Cochrane Library (www.cochranelibrary.com/), and JSTOR (www.jstor.org/) databases using MeSH terminologies such as “Non syndromic hearing loss” (congenital non syndromic hearing loss* (MeSH), non-syndromic hearing loss* (MeSH), non-syndromic hearing loss ((MeSH) and “genetics” (gene* (MeSH), genetics* (MeSH)), AND “epidemiology & pathogenesis” (epidemiology* (MeSH), pathogenesis* (MeSH)). The articles were collected from the last decade up to December 2020. Following inclusion and exclusion criteria were followed to select the appropriate literature for reference.

Inclusion criteria
The following criteria were followed for selecting the articles:
1. Literatures providing information about target genes involved in NSHL.
2. The articles published in the English language.

All available full-length original research articles and case reports were included.

Exclusion criteria
The rest of the literatures were excluded based on the following exclusion criteria:
1. The review articles, abstracts, articles written in other languages, letters to the editor, short reports, and correspondences.
2. The articles are based on syndromic HL and the adult population.
3. Articles without any genetic information.

Dataset preparation
The NSHL target genes were selected from shortlisted literatures and subjected for removal of duplicate genes. The unique NSHL target gene list was prepared for further analysis.
Discovery of NSHL pathway
All of the selected genes were employed to discover their association in different biochemical and signaling pathways involved in the development and progression of NSHL using KEGG (Kyoto Encyclopedia of Genes and Genomes) Mapper web server (https://www.genome.jp/kegg/mapper.html).

Functional network analysis
The pathway associated NSHL target proteins were analyzed using STRING (https://string-db.org/) web server in order to represent a functional network between them.

Discovery of regulatory miRNAs-gene interaction in NSHL progression
Functionally associated target genes were subjected to miRTarBase (http://miRTarBase.cuhk.edu.cn/) web server to analyze their regulatory interaction with three previously reported miRNAs (miR-182, miR-183, and miR-96). The identified miRNA-gene interaction hubs were validated and graphically represented through miRNet (https://www.mirnet.ca/) tool.

Variant analysis
Effective SNPs were predicted from NSHL target genes to identify their functional variants. SNPs were searched from dbSNP database (https://www.ncbi.nlm.nih.gov/snp/). Validation of SNPs were performed using SIFT (https://sift.bii.a-star.edu.sg/) [29], PredictSNP1 (https://loschmidt.chemi.muni.cz/predictsnp1/), and PredictSNP2 (https://loschmidt.chemi.muni.cz/predictsnp2/) algorithms [30].

The schematic representation of overall procedures is depicted in Fig. 1.

Results

Selection of target genes
Total 14,215 numbers of articles were selected on the basis of PRISMA guidelines from different scientific repositories such as PubMed (3,172), JSTOR (209), Science Direct (10,827), and Cochrane (7). Among these, a total of 787 studies were satisfied with all of the inclusion criteria and considered for this study (Fig. 2), whereas the remaining studies were excluded as per the exclusion criteria manually. Further, a total of 2,707 NSHL target genes were collected from these 787 studies, out of which 422 genes were identified as unique. Again, from these 422 unique target genes, mitochondrial RNA (5), DNA (1), and microRNAs (7) were discarded, and the rest of 381 genes (Fig. 2) were considered as the final data set of NSHL targets.
Ray M et al. • Potential target genes of NSHL

Fig. 2. PRISMA flow diagram, presenting the systematic review and gene selection procedure. KEGG, Kyoto Encyclopedia of Genes and Genomes.
Analysis of NSHL target genes

The association of 381 selected genes was searched against different biological pathways to discover potential NSHL targets. It was interpreted, a total of 23 genes (Fig. 2) are significantly involved in six different pathways such as Notch signaling pathway (hsa04330), Wnt signaling pathway (hsa04310), gap junction (hsa04540), tight junction (hsa04530), JAK-STAT signaling pathway (hsa04630), and adherens junction (hsa04520) which are associated with NSHL development (Table 1). Therefore, these 23 genes such as NOTCH1, RBPJ, LRP5, SMAD4, RAF1, ADCY1, GJA1, SNAI2, ACTB, ACTG1, FGFR1, MET, TJP2, CLDN14, CLDN9, MARVELD2, MYH9, NEDD4, NF2, RDX, IFNLR1, IL13, and PTPN11 were anticipated as potential NSHL targets. Afterward, functional association of these 23 NSHL targets was studied using the STRING tool with the high level of confidence parameter (score 0.007). The resulting protein-protein network showed a strong interaction between 16 nodes (genes) (Fig. 3) with a significant PPI enrichment p-value of 0.00107. Particularly, strong functional interactions have been observed between six groups of targets i.e., RBPJ, NOTCH1, SNAI2, and SMAD4; RDX, ACTB, ACTG1, and MYH9; CLDN9-CLDN14; TJP2-MARVELD2; PTPN11-MET; NEDD4-GJA1 in the resulted PPI network. Further, the involvement of these 16 NSHL target genes was also confirmed in a few hearing associated biological processes such as inner ear development, inner ear auditory receptor cell differentiation, homeostatic process, chemical homeostasis, signal transduction, regulation of response to external stimulus, related to hearing development and impairment (Table 2).

Further study was performed to analyze the gene regulatory network between these 16 NSHL target genes with three important miRNAs such as miR-182, miR-183, and miR-96 which are sig-

Table 1. Associated genes in signaling pathways of non-syndromic hearing loss development

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Pathway IDs</th>
<th>NSHL target genes</th>
<th>No. of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notch signaling</td>
<td>hsa04330</td>
<td>NOTCH1, RBPJ</td>
<td>2</td>
</tr>
<tr>
<td>Wnt signaling</td>
<td>hsa04310</td>
<td>LRP5, SMAD4</td>
<td>2</td>
</tr>
<tr>
<td>Gap junction</td>
<td>hsa04540</td>
<td>RAF1, ADCY1, GJA1</td>
<td>3</td>
</tr>
<tr>
<td>Adherens junction</td>
<td>hsa04520</td>
<td>SMAD4, SNAI2, ACTB, ACTG1, FGFR1, MET</td>
<td>6</td>
</tr>
<tr>
<td>Tight junction</td>
<td>hsa04530</td>
<td>TJP2, ACTB, ACTG1, CLDN14, CLDN9, MARVELD2, MYH9, NEDD4, NF2, RDX</td>
<td>10</td>
</tr>
<tr>
<td>JAK-STAT pathway</td>
<td>hsa04630</td>
<td>IFNLR1, IL13, PTPN11, RAF1</td>
<td>4</td>
</tr>
</tbody>
</table>

NSHL, non-syndromic hearing loss.

Table 2. NSHL targets involved in different biological processes obtained from STRING functional network analysis

<table>
<thead>
<tr>
<th>Biological processes</th>
<th>GO IDs</th>
<th>NSHL associated gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inner ear development</td>
<td>GO:0048839</td>
<td>NOTCH1, PTPN11</td>
</tr>
<tr>
<td>Homeostatic process</td>
<td>GO:0042592</td>
<td>RAF1, NOTCH1, MET, PTPN11, SMAD4</td>
</tr>
<tr>
<td>Chemical homeostasis</td>
<td>GO:0048878</td>
<td>RAF1, MET, PTPN11, SMAD4</td>
</tr>
<tr>
<td>Signal transduction</td>
<td>GO:0007165</td>
<td>RAF1, NOTCH1, MET, PTPN11, SMAD4, SNAI2, TJP2</td>
</tr>
<tr>
<td>Regulation of response to external stimulus</td>
<td>GO:0032101</td>
<td>RAF1, NOTCH1, MET, PTPN11, SNAI2</td>
</tr>
</tbody>
</table>

NSHL, non-syndromic hearing loss.
significantly expressed in the human inner ear and associated with NSHL. Among these, the strongly validated interaction was found between the target genes SMAD4 and SNAI2 with miR-182 and miR-183, whereas miR-96 regulates only SNAI2 (Fig. 4).

**Identification of deleterious variants in NSHL target genes**

Non-synonymous SNPs (nsSNPs) were predicted for two important NSHL targets (SMAD4, SNAI2) confirmed from previous analysis. The prediction was resulted total 178 rsIDs of pathogenic variants for SMAD4 gene with clinical significance from dbSNP (Supplementary Fig. 1). At the same time no pathogenic variant was obtained for SNAI2 gene. Initial validation using SIFT algorithm (<0.05) was identified 18 rsIDs (out of 178 rsIDs) in SMAD4 with deleterious effect (Supplementary Fig. 2). Afterwards, cross validation using PredictSNP1 was resulted 17 rsIDs in SMAD4 (Supplementary Fig. 3) as deleterious variants. Finally, theses variants were further validated using PredictSNP2 algorithms which was resulted 13 numbers of deleterious nsSNPs (Table 3, Fig. 5) assigned with 10 rsIDs i.e., rs377767345 (G352E), rs121912581 (G352R), rs377767347 (R361H, R361L), rs377767348 (C363R), rs121912580 (G386D), rs377767367 (G395V, G491V), rs377767369 (W509R), rs377767371 (G510V), rs377767382 (L533P, L533R), rs377767381 (L533V).

Furthermore, functional domain regions of SMAD4 were analyzed from the respective protein structure. Interestingly, all of these 13 deleterious variants (Table 3) were found to occur within its functional domain i.e., MH2 domain (323-552 AA) of SMAD4 gene (Fig. 6). Therefore, the present observation has added remarkable evidence to this study for putative NSHL target identification.

**Discussion**

Heterogeneity in NSHL is still under the puzzled cube which has encouraged researchers to understand its accurate genetics. In the present study, authors have attempted to discover potential NSHL target genes and their functional variants using a computational approach. Initial prediction has identified the involvement of 23 genes in the inner ear development and hearing impairment path-
ways such as Notch signaling pathway, Wnt signaling pathway, gap junction pathway, adherens junction pathway, tight junction pathway, and JAK-STAT pathway. As per the literature evidence, all of these six pathways are appeared to have a crucial role in the NSHL progression. Several studies have been reported about the differential expression pattern of different functional genes in Notch signaling pathway which has multiple roles in the development of inner ear including regulation of hair cell, determination of neurons, sensory regions, and non-sensory regions [31]. It is also disclosed that, Wnt signaling pathway has a dominant role in the dorsal structure formation of the inner ear [32,33]. In addition, a critical role in hearing is directed through gap junction pathway and mutations in connected genes have been reported to cause a high incidence of human deafness [34]. Similarly, several physiological processes such as cochlear development, growth of auditory neurons, immune mediation, and planar cell alignment are maintained through Adherens junction pathway [35]. Again, in the inner ear, tight junction is important for the separation of endolymphatic and perilymphatic space to maintain the concentration gradient between endo and perilymph and the endocochlear potential [35,36]. Mutations in genes and/or proteins associated with this pathway are reported to cause hereditary HL [35,36]. Likewise, JAK-STAT signaling pathway balances the noise-induced damage to cochlear tissue and loss of hearing sensitivity. The imbalanced expression of STAT3 is caused by loud sound which leads to cochlear tissue damage and loss of hearing sensitivity [38]. All of this evidence stood in support of 23 genes as potential NSHL targets identified from the present study.

Afterward, participation of 16 genes (out of 23) in different biological processes such as inner ear development, homeostatic process, chemical homeostasis, signal transduction, inner ear auditory receptor cell differentiation, and regulation of response to external stimulus related to hearing impairment and NSHL progression was elucidated through study of their PPI network [39-41]. Therefore, all of these biological processes are providing the major key points to consider the involved genes as potential NSHL targets. The predicted strong functional association between these genes also supported the above hypothesis. From subsequent analysis, a strong regulatory interaction between two important NSHL target genes (SMAD4 and SNAI2) with three established miRNAs (miR-183, miR-182, and miR-96) was obtained which have added a worthy value to this current study. The expressions of these three miRNAs in the human inner ear and their involvement in NSHL progression [15-18] have been studied well. According to the previous findings, both of these miRNAs (miR-182 and miR-183) down-regulate the SMAD4 expression in human bladder, prostate cancer, and ovarian cancer [42-44]. Apart from these several others such as miR-183 regulates ATP2B4, BTG1, EZRIN, GNG5, KCN]14, NCSI, NEFL, PEX19, PPP2CA, SLC6A6, TREK-1, ZCCHC3 [18] and miR-182 regulates CaV1.2, MITF, RDX [15,18]. Similarly, miR-96 is known to regulate several genes including MYRIP, AQP5, CaBP1, CaV1, CELSR1, CELSR2, GRID1, KCC2, MITF, RDX, RYK, and TFCP2L3 [17,18]. At the same time, a study in the mice model has confirmed the regulation of SNAI2 gene by miR-96 [45].

As, the regulation of SMAD4 and SNAI2 through miR-183, miR-182, and miR-96 in human NSHL is still not clear, further investigation may provide proper direction into it.

SMAD4

Fig. 6. Predicted non-synonymous single nucleotide polymorphisms within the functional MH2 domain regions (323-552 AA) of putative non-syndromic hearing loss target (SMAD4) are highlighted.
Here, total of 13 deleterious nsSNPs (G352E, G352R, R361H, R361L, C363R, G386D, G395V, G491V, W509R, G510V, L533P, L533R, and L533V) were obtained within the functional region i.e., MH2 domain of the SMAD4 gene. SMAD4 is a multifunctional modulator, which regulates cell proliferation, differentiation, autophagy, invasion, and metastasis. Deregulation of SMAD4 is also associated with embryonic developmental diseases [46]. It has a vital role in the activation of Wnt signaling pathway [47]. The SMAD4 variants G352E [48], G352R [49], R361H, R361L [48], C363R [50], G386D [49], W509R, G510V [50], L533P, L533R [48], and L533V [51] are previously identified in juvenile polyposis/hereditary hemorrhagic telangiectasia syndrome, whereas G491V is reported in colorectal cancer [51]. The appearance of most of these deleterious mutations in the MH2 domain of SMAD4 protein has emphasized their importance for further study. MH2 domain is a multifunctional region that mediates differential interaction with other proteins, transcription factors, co-activators, and co-repressors. These interactions provide specificity and selectivity to SMAD4 functions [52, 53].

Interestingly, the association of a deleterious variant i.e., rs377767367 (G491V) in NSHL target SMAD4 was discovered and is not reported in the NCBI ClinVar database for any disease conditions. Thus, present findings will not only provide insights into the genetics of NSHL but also help in further genetic counselling of NSHL after subsequent investigation.

NSHL is a critical hereditary disorder. Complexity in disease diagnosis increases due to its clinical and genetic heterogeneity. The present study has reported SMAD4 as the most potential target gene in the HL development pathways. The discovery of important miRNAs, miR-183, and miR-182 in association with the regulation of SMAD4 has also supported the above hypothesis. Identification of pathogenic and deleterious nsSNPs within its functional domain is also added an additional clue to this study. Outcomes of the present investigation may be useful in understanding the genetics of NSHL and also throw light on the diagnosis modalities of NSHL. All of these findings may be experimented for further validation.

**Authors’ Contribution**

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

**Conflicts of Interest**

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

**Acknowledgments**

We are thankful to Indian Council of Medical Research (ICMR), New Delhi for providing the Senior Research fellowship.

**Supplementary Materials**

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

**References**


27. Das SS, Chakravorty N. Identification of deleterious SNPs and their effects on BCL11A, the master regulator of fetal hemoglobin expression. Genomics 2020;112:397-403.


Hypothetical protein predicted to be tumor suppressor: a protein functional analysis

Md. Abdul Kader¹, Akash Ahammed¹, Md. Sharif Khan¹, Sheikh Abdullah Al Ashik¹, Md. Shariful Islam², Mohammad Uzzal Hossain³*

¹Department of Biotechnology and Genetic Engineering, Mawlana Bhashani Science and Technology University, Tangail 1902, Bangladesh
²University of Kentucky, Lexington, KY 40506, USA
³Bioinformatics Division, National Institute of Biotechnology, Dhaka 1349, Bangladesh

Litorilitus sediminis is a Gram-negative, aerobic, novel bacterium under the family of Colwelliaceae, has a stunning hypothetical protein containing domain called von Hippel-Lindau that has significant tumor suppressor activity. Therefore, this study was designed to elucidate the structure and function of the biologically important hypothetical protein EMK97_00595 (QBG34344.1) using several bioinformatics tools. The functional annotation exposed that the hypothetical protein is an extracellular secretory soluble signal peptide and contains the von Hippel-Lindau (VHL; VHL beta) domain that has a significant role in tumor suppression. This domain is conserved throughout evolution, as its homologs are available in various types of the organism like mammals, insects, and nematode. The gene product of VHL has a critical regulatory activity in the ubiquitous oxygen-sensing pathway. This domain has a significant role in inhibiting cell proliferation, angiogenesis progression, kidney cancer, breast cancer, and colon cancer. At last, the current study depicts that the annotated hypothetical protein is linked with tumor suppressor activity which might be of great interest to future research in the higher organism.

Keywords: functional annotation, hypothetical protein, novel bacterium, tumor suppressor, VHL domain

Introduction

Bacteria possess tremendous compatibility that can be used to the necessity of human welfare and Litorilitus sediminis can be one of them. L. sediminis is a Gram-negative, aerobic, curved-rod shaped, non-spore-forming, catalase, and oxidase-positive bacterium with the polar or sub-polar flagellum. It was isolated from a sediment sample that was collected from the coastal region of Qingdao, China [1]. This organism grew optimally at 37°C, pH 8–9. This type of bacterium was novel among the other genera under the family of Colwelliaceae. The characteristics like phenotypic, chemotaxonomic, and well-confirmed phylogenetic evidence of Litorilitus belonging to the family Colwelliaceae was distinctive that implied as a novel genus. This novel bacterium has a prominent concentration of cellular constituents compared with other genera and these are C16:0 and C16:1 ω7c fatty acids, phosphatidylethanolamine, phosphatidylglycerol, aminophospholipid, and two amino lipids (AL1, AL2) as well as isoprenoid quinone 8 [1]. Along with bacterial cellular compo-
nents, a profuse number of proteins exist where approximately 2% of the genes code for proteins as well as the remaining are non-coding or still functionally unknown [2].

The number of genes having unknown functions referred to as hypothetical proteins is present in each organism's genome [3] and these are a category of the protein whose existence is not confirmed by any experimental evidence but can be predicted to be expressed from an open reading frame [4]. The hypothetical proteins can be classified as uncharacterized protein families which are experimentally verified to exist but have not been identified or linked to a known gene, and the other type is the domain of unknown functions [5] that is experimentally characterized proteins in the absence of known functional or structural domains [6,7]. Despite the lack of functional characterization, they play a significant role in understanding biochemical and physiological pathways like exploring new structures and functions [8], pharmacological targets and markers [9], and early detection and benefits for proteomic and genomic research [10]. With the advancement of Computational Biology, it has become easier to analyze hypothetical proteins using bioinformatics tools that provide various advantages like the determination of 3D structural conformation, identification of new domains and motifs, assessment of new cascades and pathways, phylogenetic profiling, and functional annotation [11]. A recent study showed that the annotated hypothetical protein is linked with hydrolase activity which might be of great interest to further research in bacterial genetics [12].

However, due to novel genera under the family of Colwelliaceae, this study intended to characterize the protein EMK97_00595 (Litorilituus sediminis), a family of von Hippel-Lindau (VHL) that have an overwhelming function as a tumor suppressor in higher organisms. The main feature of VHL is that it is a critical regulator of the ubiquitous oxygen-sensing pathway and can act as a substrate recognition component of the E3 ubiquitin ligase complex [13], which leads to efficient ubiquitylation and proteasomal degradation of HIFα protein. On the other hand, in hypoxia condition HIFα is not prolyl hydroxylated and may escape pVHL recognition, resulting in accumulation of HIFα and formation of a complex with HIF1β, goes into the nucleus and activates a transcriptional program to cope with the short-term, long-term effects of oxygen deprivation, several signaling pathways as well as angiogenesis factor for leading cell proliferation or tumor [14,15]. So the function of the hypothetical protein that exists in the L. sediminis is considerable.

Therefore, this study manifests a reliable interpretation of this hypothetical protein EMK97_00595 (QBG34344.1) by adopting an integrated workflow that can be a potential research interest in the field of tumor suppression study.

Methods

Sequence retrieval and similarity identification

The hypothetical protein EMK97_00595 (Litorilituus sediminis) was chosen by exploring the NCBI database which can act as a significant research interest in numerous cancer research fields in the near future (Supplementary Table 1). The sequence of the hypothetical protein (GenBank accession: QBG34344.1 and NCBI reference sequence: WP_130598461.1) that may contain a tumor suppressor domain was retrieved and collected as a FASTA format and submitted to several prediction servers for the in-silico characterization. Initially, a similarity search was performed using the NCBI BLASTp program [16] against the non-redundant and Swissprot database [17], for predicting the function of the hypothetical protein.

Multiple sequence alignment and phylogeny analysis

A multiple sequence alignment is a tool used to explore closely related genes or proteins to find the evolutionary relationships between genes and to identify shared patterns among functionally or structurally related genes. Sequence alignment was performed by the MUSCLE server of EBI [18], and an evolutionary relationship was accomplished by Jalview 2.11 software [19], between the hypothetical protein EMK97_00595 and the proteins that had structural similarity with the protein of interest.

Analysis of physicochemical properties

ProtParam [5] is a tool that computes various physical and chemical parameters of protein sequences. The physicochemical properties of the hypothetical protein were predicted using the ProtParam tool in the ExPASy server [20], which predicts all the relative properties including molecular weight, theoretical pI, amino acid composition, the total number of positive and negative residues, instability index, aliphatic index and grand average of hydropathicity (GRAVY) [21-23].
Analysis of the secondary structure

The servers that were utilized to predict protein secondary structure were SOPMA [24] and PSIPRED [25]. SOPMA is a general secondary structure prediction tool, on the other hand, PSIPRED is a server for comprehensive analysis of protein. The server SOPMA was initially employed to predict the secondary structure and then the result derived from the SOPMA server was validated by exploiting PSIPRED.

3D structure modeling and quality assessment

HHpred server [26] that works based on the pairwise comparison profile of hidden Markov models, was used to build the 3-dimensional structure using the best scoring template. The confidence of the predicted structure was also visualized by SWISS-MODEL [27]. Several quality assessment tools of the SAVES and ProFunc [28] server were applied to estimate the reliability of the predicted 3D structure model of the hypothetical protein. The Ramachandran plot for the model was built using the PROCHECK program [29] to visualize the backbone dihedral angles of amino acid residues. The quality of the protein 3D structure was assessed with the help of the ERRAT server [30] and Varify 3D server was used to determine the compatibility of an atomic model (3D) with its amino acid sequence as well as comparing the results to standard structures [31,32].

Active site determination

Computed Atlas of Surface Topography (CASTp) is an online active site determination server [33] that calculates the location, delineation, and concave surface regions on 3D structures of proteins. CASTp predicted the active site of the selected hypothetical protein that showed the binding sites, amino acid binding regions with area and volume.

Identification of protein subcellular localization and topology

The subcellular location of the following protein was predicted by using the BUSCA web server [34]. BUSCA amalgamates different tools—DeepSig, TPpred3, PredGPI, BetAware, ENSEMBLE3.0, BaCelLo, MemLoci, and SChloro to predict protein features related to localization. The result was further checked by Cello [35], PsortB [36], Gneg-mPLoc [37], SOSUIGramN [38], and PSLpred [39]. Prediction of signal peptide was done by using PrediSi [40] and SignalP-5.0 Server [41]. The solubility of the hypothetical protein was evaluated by Protein-sol [42] and SOSUI [43] webserver. Protein transmembrane helices were assessed by HMMTOP [44], TMHMM [45], and Sable [46] webserver. The topology of hypothetical protein was predicted by the ProFunc server [14].

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

Domain/superfamily/family of the following hypothetical protein was analyzed by using the servers—CDD (conserved domain database) from NCBI [47], Pfam [48], SMART [49], Interpro [50], SCOP [51,52], Supfam [53], Motif, ProFunc [28], Phyre [54], and CATH-Gene3D [55]. Among them, CDD, Pfam, SMART, Interpro, SCOP, Supfam, MotifFinder were employed to predict function from the sequence of the hypothetical protein, and ProFunc, Phyre 2, and CATH-Gene3D servers were used to predict the function from the 3-dimensional structure of the hypothetical protein. Only the lowest e-value was considered to determine protein classification, which indicates good similarity. The protein folding pattern was determined by using Phyre 2 and PFP-FunDSeqE [56] servers where protein coil nature was determined by using PCoils [57] from the Bioinformatics toolkit server.

Generation of protein-protein interaction network

As the proposed investigation seeking a tumor suppressor protein from microorganisms, STRING [58] has been used to summarize the network information of VHL tumor suppressor protein. Because of being a novel microorganism, there is no specific network is available. Here the VHL protein from humans has been used as a supposition model that might give an intellectual knowledge about VHL protein if it may apply to the human.

Results

Identification of sequence homology

The overall workflow of this study has been shown in Fig. 1. The BLASTp result of the FASTA sequence of the selected protein shows the sequence homology with other identical proteins (Tables 1 and 2). Construction of phylogenetic tree using multiple sequence alignment generated from BLASTp result shows the evolutionary relationship of the selected hypothetical protein (WP_130598461.1) (Fig. 2).

Analysis of physicochemical properties

The physicochemical properties of a protein can be characterized by an analysis of the analogous properties of the amino acids (Supplementary Table 2). The hypothetical protein is negatively charged as the theoretical pi: 4.22 and the total number of positively (Arg + Lys) and negatively charged residues (Asp + Glu) were found to be 10 and 27, respectively. The computed instability index was 32.71 classifying the protein as a stable one. The aliphatic index was 77.37 which gives an indication of proteins’ stability over a wide temperature range and
Table 1. Similar proteins obtained from the non-redundant database

<table>
<thead>
<tr>
<th>Accession No.</th>
<th>Description</th>
<th>Scientific name</th>
<th>Total score</th>
<th>Query cover (%)</th>
<th>E-value</th>
<th>Identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP_118961164.1</td>
<td>Hypothetical protein (Colwellia sp. RSH04)</td>
<td>Colwellia sp. RSH04</td>
<td>349</td>
<td>100</td>
<td>5.00E-120</td>
<td>74.18</td>
</tr>
<tr>
<td>WP_033081725.1</td>
<td>Hypothetical protein (Colwellia psychrethraea)</td>
<td>Colwellia psychrethraea</td>
<td>235</td>
<td>100</td>
<td>4.00E-75</td>
<td>51.17</td>
</tr>
<tr>
<td>WP_142932219.1</td>
<td>Hypothetical protein (Aliikangiella sp. M105)</td>
<td>Aliikangiella sp. M105</td>
<td>108</td>
<td>94</td>
<td>2.00E-25</td>
<td>34.78</td>
</tr>
<tr>
<td>WP_155746905.1</td>
<td>Hypothetical protein (Scytonema sp UIC 10036)</td>
<td>Scytonema sp. UIC 10036</td>
<td>61.2</td>
<td>45</td>
<td>3.00E-08</td>
<td>34.02</td>
</tr>
<tr>
<td>BAZ36602.1</td>
<td>Hypothetical protein NIES4101_25210 (Calothrix sp NIES-4101)</td>
<td>Calothrix sp. NIES-4101</td>
<td>57.8</td>
<td>27</td>
<td>5.00E-07</td>
<td>44.83</td>
</tr>
</tbody>
</table>

Table 2. Similar proteins obtained from Swissprote database

<table>
<thead>
<tr>
<th>Entry</th>
<th>Protein names</th>
<th>Identity (%)</th>
<th>Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOA396TZK2</td>
<td>Uncharacterized protein (Colwellia sp. RSH04)</td>
<td>74.2</td>
<td>894</td>
<td>1.3e-120</td>
</tr>
<tr>
<td>AOA545UCJ6</td>
<td>VHL domain-containing protein (Aliikangiella sp. M105)</td>
<td>34.3</td>
<td>81</td>
<td>8.3e-28</td>
</tr>
<tr>
<td>AOA12Z4R2C0</td>
<td>VHL domain-containing protein (Calothrix sp. NIES-4101)</td>
<td>36.6</td>
<td>150</td>
<td>1.5e-9</td>
</tr>
<tr>
<td>AOA116H391</td>
<td>Por secretion system C-terminal sorting domain-containing protein (Robiginitalea myxolifaciens)</td>
<td>37.1</td>
<td>133</td>
<td>7e-6</td>
</tr>
<tr>
<td>AOA257JPT4</td>
<td>VHL domain-containing protein (Limnophibians sp. TS-CS-82)</td>
<td>35.1</td>
<td>124</td>
<td>2e-5</td>
</tr>
</tbody>
</table>

all the other properties have been summarized (Table 3).

**Secondary structure analysis**

The secondary structure of a protein can be able to provide some worthy information about the function. The query hypothetical protein shows the percentages of alpha-helix, beta-turn, extended strand, and the random coil of protein 21.13%, 9.91%, 33.33%, and 36.15%, respectively from SOPMA (Supplementary Figs. 1 and 2, Supplementary Table 3). The results of the secondary structure were also cross-checked by the PRISPRED server which shows a summary of similar results (Supplementary Fig. 3). The representative secondary structure of the hypothetical protein (WP_13059
Table 3. Physicochemical properties of the hypothetical protein (WP_130598461.1)

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular weight</td>
<td>23,229.44</td>
</tr>
<tr>
<td>Theoretical pl</td>
<td>4.22</td>
</tr>
<tr>
<td>Total No. of negatively charged residues (Asp + Glu)</td>
<td>27</td>
</tr>
<tr>
<td>Total No. of positively charged residues (Arg + Lys)</td>
<td>10</td>
</tr>
<tr>
<td>The instability index (II) is computed to be</td>
<td>32.71</td>
</tr>
<tr>
<td>Formula</td>
<td>C_{1024}H_{1552}N_{262}O_{346}S_{5}</td>
</tr>
<tr>
<td>Total No. of atoms</td>
<td>3189</td>
</tr>
<tr>
<td>Aliphatic index</td>
<td>77.37</td>
</tr>
<tr>
<td>Grand average of hydropathicity (GRAVY)</td>
<td>-0.261</td>
</tr>
</tbody>
</table>

8461.1) has been shown (Fig. 3).

Secondary structure predicted from SOPMA server directed (Fig. 3A); having maximum portion of random coil (36.15%), extended strand (33.33%) and alpha-helix (21.13%) and others information displayed in Supplementary Fig. 1 and Table 3. Here, alpha-helix, beta-turn, extended strand and the random coil is indicated as blue, green, red and orange, respectively (Fig. 3A). Simultaneous analyses of secondary structure from the PSIPRED server was presented (Fig. 3B, Supplementary Fig. 3), where the helix, strand and coil sections were indicated by specified color code. Other information is available in Supplementary Figs. 2-6.

Assessment and validation of protein 3-dimensional structure

PROCHECK program was used for the validation of predicted tertiary structure, where the distribution of φ and ψ angle in the model within the limits are shown (Table 4, Fig. 4). The model was presumed to be a good one according to the Ramachandran Plot Statistics, with 91.1% residues in the most favored regions. Finally, the structure validation server Verify3D and ERRAT was implicated in verifying the established model of 3D structure for the target sequence. In the Verify3D graph, 93.75% of the residues have averaged a 3D-1D score ≥ of 0.2 which indicates that the environmental profile of the model is good (Fig. 5) and the overall quality factor predicted by the ERRAT server was 60.7143 indicates a quality model (Supplementary Fig. 7). From ProFunc, the average G-factors of the hypothetical protein is calculated to be –0.20, which indicates a usual protein model.

Active site calculation

The active site of the selected hypothetical protein constituted by 11 amino acids of an area with 52.957 and a volume of 22.609. Chain X of the hypothetical protein shows the amino acids involved in the active site (F, V, Y, T, L, E, V, T, Q, W) (Fig. 6A and 6B).

The selected hypothetical protein has 11 active sites with variable
Fig. 3. Model of secondary structure. (A) Secondary structure information from SOPMA server. (B) Sequential organization and graphical visualization of secondary structure from PSIPRED.

Fig. 4. Graphical representation and assessment of protein 3D structure. Predicted 3-dimensional structure from SAVES server (Pymol view) (A), from SWISS-MODEL (B), and Ramachandran plot analysis of 3D modeled structure validated by PROCHECK program (C).
Table 4. Ramachandran plot statistics of the predicted 3D model for the target protein EMK97_00595 (WP_130598461.1)

<table>
<thead>
<tr>
<th>Plot statistics</th>
<th>No. of amino acid residues (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residues in the most favored regions [A, B, L]</td>
<td>51 (91.1)</td>
</tr>
<tr>
<td>Residues in additional allowed regions [a, b, l, p]</td>
<td>4 (7.1)</td>
</tr>
<tr>
<td>Residues in generously allowed regions [-a, -b, -l, -p]</td>
<td>0</td>
</tr>
<tr>
<td>Residues in disallowed regions</td>
<td>1 (1.8)</td>
</tr>
<tr>
<td>No. of non-glycine and non-proline residues</td>
<td>56 (100)</td>
</tr>
<tr>
<td>No. of end-residues (excl. Gly and Pro)</td>
<td>2</td>
</tr>
<tr>
<td>No. of glycine residues (shown as triangles)</td>
<td>4</td>
</tr>
<tr>
<td>No. of proline residues</td>
<td>2</td>
</tr>
<tr>
<td>Total No. of residues</td>
<td>64</td>
</tr>
</tbody>
</table>

Fig. 5. 3D-structure validation by Verify3D.

size and is constituted by 64 amino acids demonstrated (Fig. 6A and 6B). Different binding pockets of the hypothetical protein were indicated as red, blue, green, purple, orange, and pink region, and where the amino acids contributing to the beta-bridge, beta-strand, bend, turn, and coiled regions were specified by colored bars. The largest active site (red spheres) with the contributing amino acids was directed (Fig. 6C and 6D).

Assessment of protein subcellular localization and topology
The subcellular localization of the hypothetical protein seems to be an extracellular secretory signal peptide. Protein-sol and SOSUI both predict the hypothetical protein as a soluble protein. HMMTOP, TMHMM predicted the protein as a non-transmembrane protein (Table 5). The predicted topology of the protein has shown here from N-terminal to the C-terminal.

Topological orientation of the respective strands depicted (pink arrow) from the amino terminal (N) to the carboxyl terminal (C) end exposed in Fig. 7.

Functional annotation of the hypothetical protein
The initial protein domain was achieved from the CDD of NCBI. The region of the domain, superfamily, and family classifications have been determined by the servers—CDD, Pfam, SMART, Interpro, SCOP, Supfam, MotifFinder, ProFunc, Phyre 2, and CATH-Gene3D. The domain, superfamily, and family were selected based
on the lowest e-value of the following domain. The higher e-value has been filtered out from the selection procedure. The e-value 9.11e-05 of VHL beta domain from ProFunc, 2.71e-09 of VHL superfamly from SCOP, 8.1e-03 of VHL family from Supfam indicate extremely good protein alignment, respectively. The overall alignment range of the VHL beta domain was 133–212, VHL superfamly and family were 144–200, respectively. Protein coil nature was determined by using PCoils from the Bioinformatics tool-kit server. According to Phyre 2, the folding pattern of the following hypothetical protein is pre-albumin-like. On the other hand, PEF-FunSeqE is called the protein immunoglobulin-like. Both are secreted protein as well as soluble protein and hence provide a properly defined similarity indication of VHL protein (Table 6, Supplementary Figs. 4 and 7–9).

Table 5. Assessment of subcellular localization

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Servers</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction of subcellular localization</td>
<td>BUSCA</td>
<td>Extracellular space, signal peptide</td>
</tr>
<tr>
<td></td>
<td>Cello</td>
<td>Extracellular</td>
</tr>
<tr>
<td></td>
<td>PsortB</td>
<td>Unknown, signal peptide</td>
</tr>
<tr>
<td></td>
<td>Cell-Ploc</td>
<td>Extracellular</td>
</tr>
<tr>
<td></td>
<td>PSLpred</td>
<td>Extracellular protein</td>
</tr>
<tr>
<td></td>
<td>SOSUIgramN</td>
<td>Outer membrane</td>
</tr>
<tr>
<td>Signal peptide prediction</td>
<td>Predisi</td>
<td>Secreted protein, signal peptide</td>
</tr>
<tr>
<td></td>
<td>SignalP-5.0 Server</td>
<td>Signal peptide</td>
</tr>
<tr>
<td>Prediction of protein solubility</td>
<td>SOSUI</td>
<td>Soluble protein</td>
</tr>
<tr>
<td></td>
<td>Protein-sol</td>
<td>Soluble protein</td>
</tr>
<tr>
<td>Prediction of transmembrane helices</td>
<td>HMMTOP</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>TMHMM</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Sable</td>
<td>No transmembrane domain</td>
</tr>
</tbody>
</table>

Fig. 6. Active site of the hypothetical protein, binding site of the hypothetical protein indicated by red region (A, C), and amino acids involved in the active site (B, D).
Fig. 7. Topology of hypothetical protein.

Fig. 8. Protein–protein interaction network of the hypothetical VHL protein. VHL, von Hippel–Lindau.

Table 6. Function annotation of hypothetical protein through the analysis of protein domain/superfamily/family

<table>
<thead>
<tr>
<th>Server</th>
<th>Domain/Superfamily/Family</th>
<th>e-value/Confidence</th>
<th>Region/Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional annotation from sequence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conserved domain database (CDD)</td>
<td>Superfamily: pVHL</td>
<td>6.22e-05</td>
<td>146–197</td>
</tr>
<tr>
<td>Pfam</td>
<td>Family: VHL (VHL beta domain)</td>
<td>1.3e-02</td>
<td>144–200</td>
</tr>
<tr>
<td>SMART</td>
<td>VHL</td>
<td>1.2e-02</td>
<td>133–205</td>
</tr>
<tr>
<td>Interpro</td>
<td>VHL superfamily</td>
<td></td>
<td>144–199</td>
</tr>
<tr>
<td></td>
<td>VHL beta domain</td>
<td></td>
<td>131–212</td>
</tr>
<tr>
<td>Superfamily 1.75 (SCOP)</td>
<td>Superfamily: VHL</td>
<td>2.71e-09</td>
<td>144–199</td>
</tr>
<tr>
<td></td>
<td>Family: VHL</td>
<td>8.1e-03</td>
<td></td>
</tr>
<tr>
<td>Supfam</td>
<td>Superfamily: VHL</td>
<td>1.54e-09</td>
<td>144–199</td>
</tr>
<tr>
<td></td>
<td>Family: VHL</td>
<td>8.1e-03</td>
<td></td>
</tr>
<tr>
<td>Motif (from Pfam)</td>
<td>VHL beta domain</td>
<td>8.1e-03</td>
<td>146–200</td>
</tr>
<tr>
<td>Functional annotation from the 3D structure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ProFunc</td>
<td>VHL beta domain</td>
<td>9.11e-05</td>
<td>131–191</td>
</tr>
<tr>
<td>Phyre 2</td>
<td>Superfamily: VHL</td>
<td>99.8% (confidence)</td>
<td>135–212</td>
</tr>
<tr>
<td></td>
<td>Family: VHL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATH-Gene3D (From Interpro)</td>
<td>VHL beta domain</td>
<td></td>
<td>131–212</td>
</tr>
</tbody>
</table>

Analysis of protein network

The STRING interaction of VHL protein from *Homo sapiens* has been shown in Fig. 8 as a model. VHL interacts with various proteins based on their combined score (Table 7). The network has 11 nodes, 40 edges, average node degree 7.27, local clustering coefficient 0.819, expected number of edges 18, and the p-value of protein-protein interaction enrichment 7.07e-06 indicates the network has significantly more interactions than expected.

Because of being a noble microorganism that produces hypothetical VHL protein, the VHL protein from humans has been used as a supposition model that likely to be similar to VHL protein found from microorganisms. The model VHL protein interacts with 10 other proteins such as AKT1, AKT2, CUL2, EGLN1, EPAS1, HIF1A, PPP2CA, RBX1, TCEB, and TCEB2.

Similarity analysis between query (*Litorilittius sediminis, EMK97_00595*) and target (*Homo sapiens, AAB64200.1*) pVHL proteins

The mentioned *L. sediminis* (EMK97_00595) and target (*Homo sapiens, AAB64200.1*) pVHL proteins (Table 8) molecular weight, aliphatic index, and pI value bolster the confidence value between these two pVHL proteins to be more congruous for their almost resemble value [59].

The other properties like helix, coil, and beta sheet contents are also comparable whereas the beta sheet contents were massive in
the query protein rather than target protein which implies that the bacterial query pVHL proteins have higher potentiality to drive role as a tumor suppressor protein comparing with human pVHL proteins. Because the beta domain in the pVHL protein provide the binding site for HIFα degradation. The most intriguing matter from the comparisons, the query protein is highly stable rather than the human protein which implicate to substitute this protein in human is considerable [60].

Even though the helix content is a bit more in the human pVHL protein the consequence of it, in overall amino acid sequences alignment and structure formation are demonstrated following in Fig. 9 and Supplementary Fig. 10.

The human pVHL protein has a greater instability index than the novel bacterial protein, indicating that the bacterial pVHL protein will be very effective as an anti-proliferative drug to substitute in humans, which necessitates additional research (Fig. 10).

Discussion

The sequence information as well as the structural information contributes to understanding the function of a hypothetical protein (Tables 1 and 2, Fig. 2, Supplementary Table 1). This study aims to characterize a hypothetical protein, which showed strong homology with VHL superfamily, involved in tumor suppressor. Therefore, the amino acid sequence of the hypothetical protein EMK97_00595 (Litorilituus sediminis) was retrieved (Supplementary Table 2), and initially, the physicochemical properties were obtained by ExPASy’s ProtParam tool and the prediction results are the deciding factors for the hydrophilicity, stability, and function of the protein [61]. The protein was considered as a stable one even in a wide temperature range as the instability index and the aliphatic index were 32.71 and 77.37, respectively. And the query protein seems to be hydrophilic as the GRAVY was −0.261 (Table 3).

Protein structure is closely associated with its function. The secondary structure, viz. helix, sheet, turn and therefore the coil of any protein has an excellent association with the structure, function, and interaction of the protein (Fig. 3). The query hypothetical protein contains the percentages of alpha-helix, beta-turn, extended strand, and the random coil 21.13%, 9.91%, 33.33%, and 36.15%, respectively (Supplementary Table 3, Supplementary Figs. 1–4).

Table 7. Interacting proteins and their combined score from STRING 11.0 server

<table>
<thead>
<tr>
<th>Interacted protein</th>
<th>Combined score</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKT1 (RAC-alpha serine/threonine-protein kinase)</td>
<td>0.997</td>
</tr>
<tr>
<td>AKT2 (RAC-beta serine/threonine-protein kinase)</td>
<td>0.994</td>
</tr>
<tr>
<td>CUL2 (cullin-2; core component of multiple cullin-RING-based ECS E3 ubiquitin-protein ligase complexes)</td>
<td>0.999</td>
</tr>
<tr>
<td>EGLN1 (Egl nine homolog 1)</td>
<td>0.989</td>
</tr>
<tr>
<td>EPAS1 (endothelial PAS domain-containing protein 1)</td>
<td>0.994</td>
</tr>
<tr>
<td>HIF1A (hypoxia-inducible factor 1-alpha)</td>
<td>0.999</td>
</tr>
<tr>
<td>PPP2CA (serine/threonine-protein phosphatase 2A catalytic subunit alpha isoform)</td>
<td>0.993</td>
</tr>
<tr>
<td>RBX1 (E3 ubiquitin-protein ligase RBX1)</td>
<td>0.982</td>
</tr>
<tr>
<td>TCEB1 (elongin-C)</td>
<td>0.999</td>
</tr>
<tr>
<td>TCEB2 (elongin-B)</td>
<td>0.998</td>
</tr>
</tbody>
</table>

Table 8. Comparison between query and target pVHL protein properties

<table>
<thead>
<tr>
<th>Characteristics of pVHL protein</th>
<th>Litorilituus sediminis</th>
<th>Homo sapiens</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of residues</td>
<td>213</td>
<td>213</td>
</tr>
<tr>
<td>Molecular weight</td>
<td>23,229.44</td>
<td>24,152.78</td>
</tr>
<tr>
<td>Theoretical pI</td>
<td>4.22</td>
<td>4.68</td>
</tr>
<tr>
<td>Aliphatic index</td>
<td>77.37</td>
<td>75.45</td>
</tr>
<tr>
<td>Overall confidence value (%)</td>
<td>75.4</td>
<td>78.2</td>
</tr>
<tr>
<td>Predicted % helix content</td>
<td>11 (24 residues)</td>
<td>28 (60 residues)</td>
</tr>
<tr>
<td>Predicted % beta sheet content</td>
<td>43 (91 residues)</td>
<td>12 (26 residues)</td>
</tr>
<tr>
<td>Predicted % coil content</td>
<td>46 (98 residues)</td>
<td>60 (127 residues)</td>
</tr>
<tr>
<td>Instability index</td>
<td>32.71</td>
<td>68.65</td>
</tr>
</tbody>
</table>

pVHL, von Hippel-Lindau tumor suppressor protein.
Findings from SOPMA revealed that the protein has an abundance of coiled regions that contributes to higher stability and conservation of the protein structure [61]. Moreover, the protein features a reliable helices percentage in its structure, which may facilitate folding by providing more flexibility to the structure; thus, protein interactions could be increased [62].

For the prediction of the protein 3D model, HHpred was employed, where the highest identical template was selected for getting an acceptable model. The query protein WP_012259469.1 showed the highest template identity of 25% with von Hippel-Lindau disease tumor suppressor; E3 ubiquitin ligase, transcription factor, hypoxic signaling, transcription; (Homo sapiens) with lowest E-value: 1.1e-11. Ramachandran plot analysis revealed that 91.1% of residues were located in the most favored regions. Moreover, residues in additional allowed regions and generously allowed regions were 7.1% and 0.0%, respectively, which evaluated the quality of the model to be good and reliable as it is generally accepted that if 90% of residues are in the most favored regions, it is likely to be a reliable model [63], shown in Fig. 4B. The model is compatible with its sequence as Verify 3D analysis implies that 93.75% of the residues had an average 3D–1D score of ≥ 0.2 (Fig. 5). “Overall quality factor” was estimated by ERRAT, which is used to evaluate the amino acid sequence alignment between query and target pVHL protein. The black legends below the two amino acid sequences alignment indicate the consensus amino acid of the protein (from Jalview analysis). pVHL, von Hippel-Lindau tumor suppressor protein.

Fig. 9. The amino acid sequence alignment between query and target pVHL protein. The black legends below the two amino acid sequences alignment indicate the consensus amino acid of the protein (from Jalview analysis). pVHL, von Hippel-Lindau tumor suppressor protein.

Fig. 10. The structural similarity prediction between query and target pVHL protein. (A, B) pVHL proteins contain the beta domain that actually play role as a tumor suppressor protein is superimposed (using PyMOL) to infer how much structural similarity they have, the superimposed result (C) is absolutely congruous each other in the β domain region which dictate the human pVHL proteins can play magnificent role as a tumor suppressor protein even though it contain a domain. pVHL, von Hippel-Lindau tumor suppressor protein.
acid environment for non-bonded atomic interactions. Higher scores indicate higher quality, and the query protein's quality factor was 60.7143, which is greater than the generally accepted range (> 50) for a high-quality model [64]. The average G-factor of the query protein is –0.20 obtained from ProFunc analysis, which indicates a usual protein model.

Protein’s active site was determined by CASTp, containing 11 amino acids (F, V, Y, Y, T, L, E, V, T, Q, W) of an area with 52.957 and a volume of 22.609, shown in Fig. 6A and 6B. The subcellular localization obtained from CELLO, BUSCA, and other similar servers, seems to be an extracellular secretory signal peptide (Supplementary Fig. 6) and non-transmembrane (Table 5). As the functions of secreted proteins are diverse, the query hypothetical protein may work like paracrine, autocrine, endocrine, or neuroendocrine depending on the target [65]. Solubility is the most important factor and an excellent index for protein functionality (Supplementary Fig. 5). Protein-sol and SOSUI both predict the hypothetical protein as a soluble one, so it may possess good dispersibility and lead to the formation of finely dispersed colloidal systems.

The superfamily, family, and domain information have been determined by a combinational sequence and structural informative approach based on the e-value of different sequence and structure analysis servers. These servers suggested the following hypothetical protein EMK97_00595 from the organism L. sediminis to be a VHL beta domain from the VHL superfamily (Table 6, Supplementary Figs. 8 and 9). VHL tumor suppressor protein can play a role in tumor suppression in multiple ways and the most common of them is targeting the HIF that mediated tumor suppression activity through polyubiquitylation and proteasomal degradation [66]. The major contribution of pVHL is to suppress clear-cell renal cell carcinoma in kidney cancer [66,67] and phosphodiesterase 9A gene as novel biomarker in human colorectal cancer [68].

L. sediminis is a novel species and the investigated protein EMK97_00595 is also novel so there is no specific STRING derived protein-protein network is available for this organism. The protein-protein interaction network analysis shown here from H. sapiens is just for a supposition model to evaluate how the protein interacted in humans (Fig. 8). The protein-protein interaction of VHL-HIF1A with a combined score of 0.999 indicated a strong relationship between these two proteins. The interaction between VHL and HIF1A indicating the involvement of the same pathway to suppress tumor activity (Table 7, Supplementary Fig. 11) [13].

Overall, the combinational strategy of computing physicochemical properties, evaluating the secondary structure and tertiary structure information, and domain information analysis denoted the protein as VHL tumor suppressor protein that is associated with VHL disease (Table 8, Supplementary Figs. 10, 11).

Protein is the building block of life that serves both biological processes and molecular functions in living organisms. Hence, this study investigated the functional role of a hypothetical protein from a novel bacterium, L. sediminis that possesses a significant tumor suppression activity. The employment of highly recommended bioinformatics tools to analyze the combinational sequence and structural information revealed the underlying molecular function of the examined hypothetical protein. The current investigation suggested that the hypothetical protein may exhibit a VHL beta domain that is similar to the human VHL beta domain and is also a part of pVHL (Figs. 9 and 10). Therefore, this finding with the aid of bioinformatics tools can soften our viewpoint for further investigation and experimental validation of this hypothetical protein containing VHL beta domain, and the use of this hypothetical protein with the aid of modern biotechnology might be utilized to suppress tumor progression in higher organisms such as human as an alternative to human defective or mutated VHL protein in the near future.

**ORCID**

Md. Abdul Kader: https://orcid.org/0000-0001-8229-2622
Akash Ahammed: https://orcid.org/0000-0002-9757-7855
Md. Sharif Khan: https://orcid.org/0000-0003-1351-3836
Sheikh Abdullah Al Ashiks: https://orcid.org/0000-0002-3923-1266
Md. Shariful Islam: https://orcid.org/0000-0002-7631-882X
Mohammad Uzzal Hossain: https://orcid.org/0000-0002-9957-122X

**Authors' Contribution**

Conceptualization: MAK, SAAA, MUH. Data curation: AA, MSK. Formal analysis: MAK, SAAA, AA, MSK. Methodology: SAAA, MSK, MUH, MSI. Writing - original draft: MAK, AA, SAAA, MSK. Writing - review & editing: MUH, MSI.

**Conflicts of Interest**

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

**Acknowledgments**

We are very grateful to the book of Gobeshonay Bioinformatics-1st Part.
Supplementary Materials

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

References


27. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an
33. Tian W, Chen C, Lei X, Zhao J, Liang J. CASTp 3.0: computed at
43. Mitaku S, Hirokawa T. Physicochemical factors for discriminating between soluble and membrane proteins: hydrophobicity of heli-


Introduction

The process of carcinogenesis is relay on over-expression of cancer genes through gene amplification, inappropriate expression of normal genes, or mutations in genes [1,2]. Accumulation of genetic mutations and epigenetic alterations are needed to induce a normal cell to a transformed cell [3] and eventually lead to cancer formation. Cancer is mainly about the active and "gain-of-function" process of the oncogenes with loss of function of tumor suppressor genes [4]. The acquisition of oncogenes and tumor suppressor genes was shown at different times during the tumor progression in different type of tumors [5].

The complication of carcinogenesis involves several cancer hallmarks. These underlying principles are describing how normal cells transform to malignant or tumor cells. Hallmarks of cancer refers to behaviors of cancer cells that sustained to proliferative sig-
nals, resisted to cell death signals, evading from growth suppressors, enabling replicative immortalization, genome instability and mutation, continues tumor-promoting inflammation, deregulating cellular energetic, avoiding immune destruction, inducing angiogenesis, and activating invasion and metastasis [5]. Multiple signaling pathways regulate each aforementioned cancer hallmark.

Cancer cells are known to have alterations in multiple cellular signaling pathways involving the affected genes [6]. Aberrant cancer signaling pathways always have been characterized with their complexities regarding the alterations and deregulations of the genes involved [7]. These multiple dysfunctional signaling pathways implicated in cancer development have become one of the factors that lead to the complexity of carcinogenesis and difficulties faced in cancer research. With the understanding of cancer complexities in the link between these signaling pathways, the search for curing cancer can possibly done by disrupting these altered signaling pathways [8]. Non-toxic natural products can be used to disrupt these altered signaling pathways; these natural products can activate cell death signals in pre-cancerous or cancerous cells without affecting the normal cells [9]. Besides, these natural products can be used alone or in combination with chemotherapeutic drugs.

More than 10 cancer signaling pathways have been identified [10]. Studies on the aberrant signaling pathways are not necessarily limited to only one or two specific signaling pathways towards cancer progression. In many cases, a few of the pathways would be studied altogether to clearly know the effects of the complexities and the interactions between these signaling networks towards carcinogenesis. In term of inhibiting signal transduction, suppression of multiple signaling pathways is better than suppression on single pathway at a time [11]. In a feedback loop theory, suppression of one signaling pathway would activate another pathway automatically. This is a phenomenon where cancer cells prolong the survival as mutated cell [12]. In all, better understanding and accurate dissection of all these signaling pathways are important; this is because they are vital for elucidating the most appropriate and reliable target molecules for the design of effective cancer therapy [13]. For instance, many potential chemo preventive drugs have been tested on multiple dysfunctional proteins that are involved in multiple dysfunctional cancer signaling pathways.

*Impatiens balsamina*, Linn from the family of Balsaminaceae, is an ornamental plant that has been used to treat various skin diseases locally. 2-Methoxy-1,4-naphthoquinone (MNQ), isolated from pericarps of *Impatiens balsamina* Linn. has been shown to cause cytotoxic towards various cancer cell lines [14], show anti-tumor-promoting activity in HepG2 and Raji cells [15,16], trigger apoptotic pathway and the upper stream modulator of many cancer pathways [17], inhibit protein kinase C expressions in Raji cells [16], and suppress the invasion and migration of MDA-MB-231 [18]. Furthermore, MNQ also has been shown to inhibit WNT signaling in STF/293 cells [19], altered proteins related to cytoskeletal functions and regulations, mRNA processing, protein modifications, and oxidative stress responses [20].

PKC is one of the protein kinases that play a role in carcinogenesis and maintenance of malignant phenotype. Moreover, it is also the core upper stream kinase involved in mitogen-activated protein kinase (MAPK), phosphoinositide 3-kinase (PI3K), and nuclear factor xB (NF-xB) pathways that regulate many down-stream proteins implicated in these cancer signaling pathways. In our previous study, MNQ has been demonstrated to inhibit PKC βI, δ, and ζ [16]; therefore, this study was continued to reveal the genes that are possibly regulated by MNQ in MAPK, PI3K, and NF-xB pathways.

**Methods**

**Chemicals**

Phorbol 12-myristate 13-acetate (PMA) and sodium n-butyrate (SnB) were obtained from Sigma. MNQ was isolated from the pericarps of *Impatiens balsamina*, Linn. The compound was prepared in dimethyl sulfoxide and stored at 4°C. The reagents obtained from Qiagen (Hilden, Germany) were RT² First Strand Kit (containing GE buffer, 5 × BC3 buffer, RE3 reverse transcriptase mix, P2 control, and RNase-free water), RT² STBR Green ROX FAST mastermix (containing HotStart DNA Taq Polymerase, PCR buffer, dNTP mix [dATP, dCTP, dGTP, dTTP], SYBR Green dye ROX mastermix).

**Cell culture**

Raji cells obtained from Riken Cell Bank, Japan were maintained in commercial Roswell Park Memorial Institute Media (RPMI-1640) supplemented with 10% fetal bovine serum (FBS) (Thermo Fisher, Waltham, MA, USA). The cells were incubated in a humidified atmosphere at 37°C with 5% CO₂ incubator (ESCO, Hatboro, PA, USA).

**Structure confirmation of MNQ**

The high-performance liquid chromatography (HPLC) chromatogram of the isolated MNQ was first identified by comparing the HPLC chromatogram of MNQ HPLC chromatogram obtained by Teng [14] in our previous work. After that, mass spectrometry spectrum was compared to NIST mass spectral data. The identified MNQ was then isolated and subjected to HPLC spiking
analysis and gas chromatography and mass spectrometry spiking analysis. Isolated MNQ structure was confirmed with nuclear magnetic resonance (NMR). Proton and carbon NMR chemical shift results were compared to Teng [14].

**Treatment of cells for RT² Profiler PCR arrays**

Raji cells (5 × 10⁵ cells/mL) were incubated in 1 mL of RPMI 1640 medium (supplemented with 10 % FBS) containing 0.05 µM PMA, 3 mM SnB, and MNQ in a 24-well plate and then incubated at 37°C for 6, 12, 24, and 48 h in a CO₂ incubator. PMA and SnB was used as inducer and enhancer respectively to cause Epstein-Barr virus activation in Raji cells and directly transform Raji cells into promotion stage.

**RNA extraction and cDNA preparation**

RNA of the treated Raji cells was extracted according to the manufacturer’s instruction using RNeasy Mini Kit (Qiagen). RT² RNA QC PCR Array was used to confirm the purity of RNA and excluded substandard samples prior to RT² Profiler PCR Arrays analysis. The RNA samples extracted were at high-quality RNA, in which the A260:A230 ratio obtained was greater than 2.0 and the A260:A280 ratio attained was in between 1.8 to 2.0. Next, the cDNA was synthesized using the RT² First Strand Kit (Qiagen) following the manufacturer’s instructions. The genomic DNA elimination mix was prepared accordingly to the manual given by Qiagen.

**RT² Profiler PCR Array**

The cDNA was added with RT² SYBR Green ROX FAST Mastermix according to manufacturer’s protocol. RT² Profiler PCR Array for the detection of key cancer genes regulation in MAPK, PI3K, and NF-κB pathways were tested in three independent experiments. The genes expression results were calculated using results generated by the RT cycler and the ΔΔCₚ method. The Cₚ values for all wells were exported to PCR Array Data Analysis Template Excel sheet (downloaded at http://www.sabiosciences.com/data-analysis.php) and uploaded to web-based software at www.sabiosciences.com/pcrarraydataanaalysis.php. This software helps in calculating Cₚ values collected and shows p-values and fold change for each gene. The connection between genes was analyzed using web-based resources Gene Network Central Pro at http://gnpcpro.sabiosciences.com/gncpro/gncpro.php.

**Pathway and network analysis**

The identified key cancer genes were first uploaded onto Qiagen’s Ingenuity Pathway Analysis (IPA) system for core analysis and then overlaid with the global molecular network in the ingenuity pathway knowledge base. IPA was performed to identify canonical pathways, diseases and functions, and gene networks that are most significant to key cancer genes and to categorize differentially expressed genes in specific molecular and cellular functions. Thresholds of two-fold or greater in changes in expression and a p-value of 0.05 or less for significance were used to filter the findings from the analysis with IPA software.

**Statistical analysis**

The results were obtained from three separate experiments. In terms of the sensitivity and accuracy, the cutoff fold change (ΔΔCₚ was set at greater than 5.0 or less than −4.0.

**Results**

**Effect of MNQ on MAPK, PI3K, and NF-κB pathway-focused gene expression profiling**

A total of 43 statistically significant expressed genes were identified, which was 19.72 % of the total of 218 cancer key genes studied. All the 43 genes were statistically significant expressed in Raji cells after treated with 0.05 µM PMA, 3 mM SnB, and 40 µM MNQ (p < 0.05, fold change > 4.0). The fold changes of all statistically significant genes involved in these pathways are shown in Fig. 1 (MAPK pathway), Fig. 2 (PI3K pathway), and Fig. 3 (NF-κB pathway). While Fig. 4 shows bar chart for fold change of all statistically significant genes involved in more than one pathway. Negative and positive values denoted down-regulation and up-regulation of gene expression, respectively.

From the 43 genes identified, MNQ up-regulated 21 genes and down-regulated 22 genes. The identified 21 up-regulated genes were apoptotic genes (BAD, EGR1, FOXO1, FOXO3, ITGB1, LTA, TNF, TNFRSF1A, TNFRSF1B, TRADD, YWHAH), tumor suppressor genes (APC, RB1, and TP53) and cell cycle regulation genes (CDKN1A, CDKN1B, CDKN2A, CDKN2D, MTOR, LAMTOR3, and PTEN). The confirmed 22 down-regulated genes were anti-apoptotic genes (HSBP1, NFKB1, NFKB2, NFKB1A, NFKB1E, RPS6KA1), angiogenesis genes (IL6, TIMP1), cell cycle regulation genes (ARAF, AKT2, AKT3, BRAF, EGFR, HRAS, IGF1, IGF1R, KRAS, NRAS, PDGFRA), transcription factors (FOS and JUN), proto-oncogenes (BCL3).

Among all the 43 statistically significant expressed genes, 12 genes (ARAF, BRAF, CDKN1A, CDKN1B, CDKN2A, CDKN2D, EGFR, KRAS, LAMTOR3, NRAS, RB1, and TP53) are involved in MAPK pathway (Table 1), 14 genes (AKT2, AKT3, APC, BAD, FOXO1, FOXO3, IGF1, IGF1R, ITGB1, MTOR, PDGFRA, PTEN,
RPS6KA1, and YWHAH) are involved in PI3K pathway (Table 2), 10 genes (BCL3, IL6, LTA, NFKB2, NFKB1E, TIMP1, TNF, TNFRSF1A, TNFRSF1OB, and TRADD) are involved in NF-κB pathway (Table 3), and seven genes (EGR1, FOS, HRAS, HSPB1, JUN, NFKB1, and NFKB1A) are involved in all three pathways (Table 4).
Table 1. Up-regulated and down-regulated genes in MAPK pathway

<table>
<thead>
<tr>
<th>Gene</th>
<th>Name</th>
<th>Time interval (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48</td>
</tr>
<tr>
<td>BRAF</td>
<td>V-raf murine sarcoma viral oncogene homolog B1</td>
<td>–12.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–10.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–11.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–12.33</td>
</tr>
<tr>
<td>KRAS</td>
<td>V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog</td>
<td>–10.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–11.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–11.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–10.47</td>
</tr>
<tr>
<td>ARAF</td>
<td>V-raf murine sarcoma 3611 viral oncogene homolog</td>
<td>–5.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–6.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–7.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–4.27</td>
</tr>
<tr>
<td>EGFR</td>
<td>Epidermal growth factor receptor</td>
<td>–5.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–6.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–12.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–10.69</td>
</tr>
<tr>
<td>N Ras</td>
<td>Neuroblastoma RAS viral (v-ras) oncogene homolog</td>
<td>–4.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–6.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–5.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–5.53</td>
</tr>
<tr>
<td>LAMTOR3</td>
<td>Late endosomal/lysosomal adaptor, MAPK and MTOR activator 3</td>
<td>4.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11.35</td>
</tr>
<tr>
<td>TP53</td>
<td>Tumor protein p53</td>
<td>5.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.35</td>
</tr>
<tr>
<td>RB1</td>
<td>Retinoblastoma 1</td>
<td>5.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.37</td>
</tr>
<tr>
<td>CDKN1B</td>
<td>Cyclin-dependent kinase inhibitor 1B (p27, Kip1)</td>
<td>5.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.31</td>
</tr>
<tr>
<td>CDKN2D</td>
<td>Cyclin-dependent kinase inhibitor 2D (p19, inhibits CDK4)</td>
<td>5.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.35</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>Cyclin-dependent kinase inhibitor 1A (p21, Cip1)</td>
<td>7.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.37</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>Cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)</td>
<td>10.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.38</td>
</tr>
</tbody>
</table>

MAPK, mitogen-activated protein kinase.

Table 2. Up-regulated and down-regulated genes in PI3K pathway

<table>
<thead>
<tr>
<th>Gene</th>
<th>Name</th>
<th>Time interval (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48</td>
</tr>
<tr>
<td>PDGFRα</td>
<td>Platelet-derived growth factor receptor, alpha polypeptide</td>
<td>–10.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–8.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–9.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–10.27</td>
</tr>
<tr>
<td>AKT2</td>
<td>V-akt murine thymoma viral oncogene homolog 2</td>
<td>–7.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–9.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–20.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–12.37</td>
</tr>
<tr>
<td>AKT3</td>
<td>V-akt murine thymoma viral oncogene homolog 3 (protein kinase B, gamma)</td>
<td>–6.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–9.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–12.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–5.32</td>
</tr>
<tr>
<td>RPS6KA1</td>
<td>Ribosomal protein S6 kinase, 90 kDa, polypeptide 1</td>
<td>–5.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–6.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–4.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–6.30</td>
</tr>
<tr>
<td>IGFl</td>
<td>Insulin-like growth factor 1 (somatomedin C)</td>
<td>–5.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–6.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–7.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–6.34</td>
</tr>
<tr>
<td>IGFlR</td>
<td>Insulin-like growth factor 1 receptor</td>
<td>–5.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–6.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–7.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–6.93</td>
</tr>
<tr>
<td>Pten</td>
<td>Phosphatase and tensin homolog</td>
<td>4.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.36</td>
</tr>
<tr>
<td>FOXO1</td>
<td>Forkhead box 01</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.28</td>
</tr>
<tr>
<td>FOXO3</td>
<td>Forkhead box 03</td>
<td>4.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.90</td>
</tr>
<tr>
<td>YWHAH</td>
<td>Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide</td>
<td>5.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.44</td>
</tr>
<tr>
<td>Mtor</td>
<td>Mechanistic target of rapamycin (serine/threonine kinase)</td>
<td>6.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.24</td>
</tr>
<tr>
<td>ItgB1</td>
<td>Integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)</td>
<td>6.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.32</td>
</tr>
<tr>
<td>Bad</td>
<td>BCL2-associated agonist of cell death</td>
<td>7.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.36</td>
</tr>
<tr>
<td>ApC</td>
<td>Adenomatous polyposis coli</td>
<td>12.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11.45</td>
</tr>
</tbody>
</table>

PI3K, phosphoinositide 3-kinase.

Table 3. Up-regulated and down-regulated genes in NF-κB pathway

<table>
<thead>
<tr>
<th>Gene</th>
<th>Name</th>
<th>Time interval (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48</td>
</tr>
<tr>
<td>NFκB2</td>
<td>Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (p49/p100)</td>
<td>–5.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–17.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–12.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–10.43</td>
</tr>
<tr>
<td>NFκB1E</td>
<td>Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon</td>
<td>–4.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–5.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–6.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–9.62</td>
</tr>
<tr>
<td>Il6</td>
<td>Interleukin 6 (interferon, beta 2)</td>
<td>–4.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–4.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–4.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–5.04</td>
</tr>
<tr>
<td>Timp1</td>
<td>TIMP metallopeptidase inhibitor 1</td>
<td>–4.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–4.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–3.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–4.05</td>
</tr>
<tr>
<td>Bcl3</td>
<td>B-cell CLL/lymphoma 3</td>
<td>–4.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–4.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–5.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–5.39</td>
</tr>
<tr>
<td>Tradd</td>
<td>TNFRSF1A-associated via death domain</td>
<td>4.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.09</td>
</tr>
<tr>
<td>Lta</td>
<td>Lymphotoxin alpha (TNF superfamily, member 1)</td>
<td>6.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.39</td>
</tr>
<tr>
<td>Tnfrsf1A</td>
<td>Tumor necrosis factor receptor superfamily, member 1A</td>
<td>6.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.48</td>
</tr>
<tr>
<td>Tnfrsf10B</td>
<td>Tumor necrosis factor receptor superfamily, member 10b</td>
<td>7.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.09</td>
</tr>
<tr>
<td>Tnf</td>
<td>Tumor necrosis factor</td>
<td>8.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12.59</td>
</tr>
</tbody>
</table>

NF-κB, nuclear factor κB.

https://doi.org/10.5808/gi.21041
Table 4. Down-regulated genes in MAPK, PI3K, or NF-κB pathways

<table>
<thead>
<tr>
<th>Gene</th>
<th>Name</th>
<th>Pathway involved</th>
<th>Time interval (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSPB1</td>
<td>Heat shock 27 kDa protein 1</td>
<td>MAPK, PI3K</td>
<td>-6.96 -12 -24 -48</td>
</tr>
<tr>
<td>JUN</td>
<td>JUN proto-oncogene</td>
<td>MAPK, PI3K, NF-κB</td>
<td>-5.02 -5.30 -5.37 -6.02</td>
</tr>
<tr>
<td>NFkB1</td>
<td>Nuclear factor of kappa light polypeptide gene enhancer in B-cells</td>
<td>PI3K, NF-κB</td>
<td>-4.89 -5.93 -4.95 -5.94</td>
</tr>
<tr>
<td>FOs</td>
<td>FBJ murine osteosarcoma viral oncogene homolog</td>
<td>MAPK, PI3K, NF-κB</td>
<td>-4.28 -4.39 -5.29 -6.47</td>
</tr>
<tr>
<td>Hras</td>
<td>V-H-ras Harvey rat sarcoma viral oncogene homolog</td>
<td>MAPK, PI3K, NF-κB</td>
<td>-4.28 -4.30 -5.38 -6.34</td>
</tr>
<tr>
<td>NFkB1A</td>
<td>Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha</td>
<td>PI3K, NF-κB</td>
<td>-4.26 -3.95 -4.39 -4.88</td>
</tr>
<tr>
<td>Egr1</td>
<td>Early growth response 1</td>
<td>MAPK, NF-κB</td>
<td>5.38 4.48 8.32 5.32</td>
</tr>
</tbody>
</table>

MAPK, mitogen-activated protein kinase; PI3K, phosphoinositide 3-kinase; NF-κB, nuclear factor κB.

Table 5. Functions of 21 up-regulated genes by MNQ

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Pathway</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP53</td>
<td>Tumor protein p53</td>
<td>MAPK</td>
<td>Tumor suppressor</td>
</tr>
<tr>
<td>RB1</td>
<td>Retinoblastoma 1</td>
<td>MAPK</td>
<td>Tumor suppressor</td>
</tr>
<tr>
<td>CDKN1B</td>
<td>Cyclin-dependent kinase inhibitor 1B (p27, Kip1)</td>
<td>MAPK</td>
<td>Suppress tumor growth and prevent cell cycle progression</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>Cyclin-dependent kinase inhibitor 1A (p21, Cip1)</td>
<td>MAPK</td>
<td>Suppress tumor growth and prevent cell cycle progression</td>
</tr>
<tr>
<td>CDKN2D</td>
<td>Cyclin-dependent kinase inhibitor 2D (p19, inhibits CDK4)</td>
<td>MAPK</td>
<td>Important tumor suppressor gene in cell cycle regulation</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>Cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)</td>
<td>MAPK</td>
<td>Important tumor suppressor gene in cell cycle regulation</td>
</tr>
<tr>
<td>LAMTOR3</td>
<td>Late endosomal/lysosomal adaptor, MAPK and Mtor activator 3</td>
<td>MAPK</td>
<td>Activation of MAP2 and Mtor induce cell proliferation</td>
</tr>
<tr>
<td>TRADD</td>
<td>TNFRSF1A-associated via death domain</td>
<td>NF-κB</td>
<td>Induction of apoptosis, suppresses TRAF2 mediated apoptosis</td>
</tr>
<tr>
<td>LTA</td>
<td>Lymphotoxin alpha (TNF superfamily, member 1)</td>
<td>NF-κB</td>
<td>Induce apoptosis in gastric cancer</td>
</tr>
<tr>
<td>TNF</td>
<td>Tumor necrosis factor</td>
<td>NF-κB</td>
<td>Death domain receptor, transducer for apoptosis signal</td>
</tr>
<tr>
<td>TNFRSF1A</td>
<td>Tumor necrosis factor receptor superfamily, member 1A</td>
<td>NF-κB</td>
<td>Death domain receptor, transducer for apoptosis signal</td>
</tr>
<tr>
<td>TNFRSF1OB</td>
<td>Tumor necrosis factor receptor superfamily, member 10b</td>
<td>NF-κB</td>
<td>Death domain receptors, induce apoptosis, activates MAPK and NF-κB pathway</td>
</tr>
<tr>
<td>APC</td>
<td>Adenomatous polyposis coli</td>
<td>PI3K</td>
<td>Tumor suppressor mutated in colon cancer</td>
</tr>
<tr>
<td>BAD</td>
<td>BCL2-associated agonist of cell death</td>
<td>PI3K</td>
<td>Induce apoptosis in breast cancer</td>
</tr>
<tr>
<td>FOXO1</td>
<td>Forkhead box O1</td>
<td>PI3K</td>
<td>Tumor suppressors in a variety of cancers, transcriptional activator that regulates apoptosis and cell cycle progression, activating BCL2</td>
</tr>
<tr>
<td>FOXO3</td>
<td>Forkhead box O3</td>
<td>PI3K</td>
<td>Tumor suppressors in a variety of cancers, transcriptional activator that triggers apoptosis in the absence of survival factors</td>
</tr>
<tr>
<td>ITGB1</td>
<td>Integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)</td>
<td>PI3K</td>
<td>Involve in tumor progression and metastasis</td>
</tr>
<tr>
<td>Mtor</td>
<td>Mechanistic target of rapamycin</td>
<td>PI3K</td>
<td>Cell cycle arrest and immune-suppressive effects</td>
</tr>
<tr>
<td>YWHAH</td>
<td>Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide</td>
<td>PI3K</td>
<td>Relates to tumor cell proliferation and malignant outcome of gastric carcinoma</td>
</tr>
<tr>
<td>Pten</td>
<td>Phosphatase and tensin homolog</td>
<td>PI3K</td>
<td>Tumor suppressor gene, regulating cell cycle</td>
</tr>
<tr>
<td>Egr1</td>
<td>Early growth response 1</td>
<td>MAPK, NF-κB</td>
<td>Transcription factor, suppress cell growth, transformation, and induce apoptosis</td>
</tr>
</tbody>
</table>

MNQ, 2-methoxy-1,4-naphthoquinone; MAPK, mitogen-activated protein kinase; NF-κB, nuclear factor κB; PI3K, phosphoinositide 3-kinase.

Overall, the functions of 21 up-regulated and 22 down-regulated genes regulated by MNQ are listed in Tables 5 and 6, respectively.

The IPA software system enables systemic analysis of 43 statistically significant genes in a biologic context. Network of up/down-regulated genes were then algorithmically generated based on their inter-relationships. Five major networks were identified and included with functions related to cell death and survival, cellular development, cell cycle, proliferation and cell to cell signaling (Table 7). Canonical pathways were then identified and analyzed from the IPA libraries that were most significant to our common
Table 6. Functions of 22 down-regulated genes by MNQ

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Pathway</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRAF</td>
<td>V-raf murine sarcoma viral oncogene homolog B1</td>
<td>MAPK</td>
<td>Proto-oncogene involved in cell growth and differentiation of melanoma</td>
</tr>
<tr>
<td>NRAS</td>
<td>Neuroblastoma RAS viral (v-ras) oncogene homolog</td>
<td>MAPK</td>
<td>Proto-oncogene regulating cell division in response to growth factor stimulation</td>
</tr>
<tr>
<td>ARAF</td>
<td>V-raf murine sarcoma 3611 viral oncogene homolog</td>
<td>MAPK</td>
<td>Proto-oncogene involved in cell growth and development</td>
</tr>
<tr>
<td>KRAS</td>
<td>V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog</td>
<td>MAPK</td>
<td>Proto-oncogene regulating cell division in response to growth factor stimulation</td>
</tr>
<tr>
<td>EGFR</td>
<td>Epidermal growth factor receptor</td>
<td>MAPK</td>
<td>Cell cycle regulation</td>
</tr>
<tr>
<td>NFKB2</td>
<td>Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (p49/p100)</td>
<td>NF-κB</td>
<td>Anti-apoptosis</td>
</tr>
<tr>
<td>NFKB1E</td>
<td>Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon</td>
<td>NF-κB</td>
<td>Inhibitor in colon cancer</td>
</tr>
<tr>
<td>TIMP1</td>
<td>TIMP metalloproteinase inhibitor 1</td>
<td>NF-κB</td>
<td>Promote cell proliferation, angiogenesis, and anti-apoptotic</td>
</tr>
<tr>
<td>IL6</td>
<td>Interleukin 6 (interferon, beta 2)</td>
<td>NF-κB</td>
<td>Inflammatory cytokine, inhibit apoptosis, induce angiogenesis</td>
</tr>
<tr>
<td>BCL3</td>
<td>B-cell CLL/lymphoma 3</td>
<td>NF-κB</td>
<td>Proto-oncogenes, activates by NF-κB</td>
</tr>
<tr>
<td>AKT2</td>
<td>V-akt murine thymoma viral oncogene homolog 2</td>
<td>PI3K</td>
<td>Cell proliferation, anti-apoptosis</td>
</tr>
<tr>
<td>AKT3</td>
<td>V-akt murine thymoma viral oncogene homolog 3 (protein kinase B, gamma)</td>
<td>PI3K</td>
<td>Cell proliferation, anti-apoptosis</td>
</tr>
<tr>
<td>RPS6KA1</td>
<td>Ribosomal protein S6 kinase, 90 kDa, polypeptide 1</td>
<td>PI3K</td>
<td>Interact with NFKB1A and MAPK1, important mediator of survival signals that protect cells from undergoing apoptosis</td>
</tr>
<tr>
<td>IGF1</td>
<td>Insulin-like growth factor 1 (somatomedin C)</td>
<td>PI3K</td>
<td>Cell cycle regulation</td>
</tr>
<tr>
<td>IGF1R</td>
<td>Insulin-like growth factor 1 receptor</td>
<td>PI3K</td>
<td>Cell cycle regulation</td>
</tr>
<tr>
<td>PDGFRα</td>
<td>Platelet-derived growth factor receptor, alpha polypeptide</td>
<td>PI3K</td>
<td>Cell cycle regulation</td>
</tr>
<tr>
<td>FOS</td>
<td>FBJ murine osteosarcoma viral oncogene homolog</td>
<td>MAPK, PI3K, NF-κB</td>
<td>Transcription factors</td>
</tr>
<tr>
<td>HRAS</td>
<td>V-Ha-ras Harvey rat sarcoma viral oncogene homolog</td>
<td>MAPK, PI3K, NF-κB</td>
<td>Proto-oncogene regulating cell division in response to growth factor stimulation</td>
</tr>
<tr>
<td>JUN</td>
<td>JUN proto-oncogene</td>
<td>MAPK, PI3K, NF-κB</td>
<td>Transcription factors</td>
</tr>
<tr>
<td>NFKB1</td>
<td>Nuclear factor of kappa light polypeptide gene enhancer in B-cells 1</td>
<td>PI3K, NF-κB</td>
<td>Anti-apoptosis</td>
</tr>
<tr>
<td>NFKB1A</td>
<td>Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha</td>
<td>PI3K, NF-κB</td>
<td>Rel protein inhibitor in colon cancer</td>
</tr>
<tr>
<td>HSPB1</td>
<td>Heat shock 27 kDa protein 1</td>
<td>MAPK, PI3K</td>
<td>Inhibition of apoptosis, regulation of cell development, and cell differentiation, cytoprotection, and support of cell survival under stress conditions, enhances the activation of the NF-κB pathway</td>
</tr>
</tbody>
</table>

MNQ, 2-methoxy-1,4-naphthoquinone; MAPK, mitogen-activated protein kinase; NF-κB, nuclear factor κB; PI3K, phosphoinositide 3-kinase.

Table 7. Molecular and cellular functions

<table>
<thead>
<tr>
<th>Name</th>
<th>p-value</th>
<th>No. of molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell death and survival</td>
<td>5.58E-14–2.02E-38</td>
<td>40</td>
</tr>
<tr>
<td>Cellular development</td>
<td>3.95E-14–2.60E-37</td>
<td>40</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>5.33E-14–1.66E-33</td>
<td>35</td>
</tr>
<tr>
<td>Cellular growth and proliferation</td>
<td>3.95E-14–4.31E-33</td>
<td>39</td>
</tr>
<tr>
<td>Cell-to-cell signaling and interaction</td>
<td>1.12E-14–1.81E-29</td>
<td>32</td>
</tr>
</tbody>
</table>

gene data set. Top statistically significant canonical pathways included PTEN, glioblastoma multiforme, PI3K/AKT, glioma signaling 1 and molecular mechanisms of cancer signaling pathways are shown in Table 8 (p < 0.05) and network contains differentially expressed 43 genes are shown in Fig. 5.

Discussion

The common cancer pathways are MAPK, PI3K, NF-κB, STAT, NOTCH, TP53, RB1, WNT, and Hedgehog pathways. These pathways play crucial roles in survival and maintain tumourigenesis properties of cancer cells [21]. All these pathways are complement to each other forming a very complicated network, which has not been fully studied yet. In term of inhibiting signal transduction, suppression of multiple signaling pathways is better than suppression on single pathway at a time [11].
Table 8. List of the genes in most significantly up/down-regulated top five canonical pathways

<table>
<thead>
<tr>
<th>Pathway</th>
<th>–Log (p-value)</th>
<th>Ratio</th>
<th>Overlap, n (%)</th>
<th>Molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTEN signaling</td>
<td>2.77E01</td>
<td>1.44E-01</td>
<td>17/118 (14.4)</td>
<td>ITGB1, AKT2, NRAS, YWHAH, BAD, HRAS, KRAS, NFKB2, NFKB1, PTEN, FOXO1, CDKN1A, FOXO3, PDGFR, IGF1R, CDKN1B, EGFR</td>
</tr>
<tr>
<td>Glioblastoma multi-forme signaling</td>
<td>2.6E01</td>
<td>1.16E-01</td>
<td>17/146 (11.6)</td>
<td>TP53, CDKN2A, AKT2, NRAS, HRAS, KRAS, APC, PTEN, RB1, MTOR, IGF1, FOXO1, CDKN1A, PDGFR, IGF1R, CDKN1B, EGFR</td>
</tr>
<tr>
<td>PI3K/AKT signaling</td>
<td>2.52E01</td>
<td>1.3E-01</td>
<td>16/123 (13.0)</td>
<td>TP53, ITGB1, AKT2, NRAS, YWHAH, BAD, HRAS, KRAS, NFKB2, NFKB1, PTEN, FOXO1, CDKN1A, PDGFR, IGF1R, CDKN1B, EGFR</td>
</tr>
<tr>
<td>Glioma signaling molecular</td>
<td>2.5E01</td>
<td>1.58E-01</td>
<td>15/95 (15.8)</td>
<td>TP53, CDKN2A, AKT2, NRAS, HRAS, KRAS, PTEN, RB1, MTOR, CDKN2D, IGF1, CDKN1A, PDGFR, IGF1R, CDKN1B, EGFR</td>
</tr>
<tr>
<td>Mechanisms of cancer</td>
<td>2.42E01</td>
<td>5.48E-02</td>
<td>20/365 (5.5)</td>
<td>TP53, ITGB1, CDKN2A, AKT2, NRAS, BAD, HRAS, KRAS, NFKB2, NFKB1, APC, BRAF, RB1, FOS, MTOR3, CDKN2D, JUN, FOXO1, CDKN1A, CDKN1B</td>
</tr>
</tbody>
</table>

PI3K, phosphoinositide 3-kinase.

Fig. 5. Network contains differentially expressed 43 genes. Top functions of the genes were related to molecular and cellular functions and physiological system development and function. Ingenuity Pathway Analysis network legend is on the left side.
In a feedback loop theory, suppression of one signaling pathway would activate another pathway automatically. This is a phenomenon where cancer cells can prolong the survival as mutated cell [12]. Thus, this study focuses to determine the modulatory effects of MNQ on the key genes involved in MAPK, PI3K, and NF-kB pathways.

Overall, MNQ significantly regulated 43 genes (19.72%) from the total 218 cancer key genes studied in these three signaling pathways (Figs. 1–3). These 43 genes consist of 21 up-regulated genes (Table 5) and 22 down-regulated genes (Table 6) have different functions in regulating tumorigenesis in cancer cells (Table 7). The regulatory effects of these 43 genes in various cellular processes are discussed in following sections.

In this study, MNQ activated the apoptosis, tumor suppressor and cell cycle regulatory activities in MAPK pathway, where the genes of LAMTOR3, TP53, RB1, CDKN1A, CDKN1B, CDKN2A and CDKN2D were up-regulated across all time points (Table 1, Fig. 1). MNQ induced the expression of P53 and RB1 transcription factors to exert its suppressive effect in preventing the progression of cell cycle in Raji cells. Upon the treatment of MNQ, MNQ could possibly cause DNA damage [22] and thus induce the expression of P53. In respond to DNA damage, P53 induces cyclin-dependent kinase inhibitors such as CDKN1A, CDKN1B, CDKN2A, CDKN2D that responsible for cell cycle arrest. The expression of late endosomal/lysosomal adaptor, MAPK and MTOR activator 3 (LAMTOR3) in MNQ-treated Raji could possibly explain that Raji cells tried to sustain their survival in the presence of the abovementioned genes in inducing cell cycle arrest and apoptosis. This phenomenon can be observed from Fig. 1 where the LAMTOR3 expression were gradually increased in amount from 6 to 48 h treatment. In addition, ERGF, epidermal growth factor receptor involved in cell cycle was down-regulated by MNQ. Furthermore, proto-oncogene genes that are involved in cell growth and differentiation such as ARAF, BRAF, KRAS, HRAS and NRAS were also down-regulated across all time points. Overall, MNQ mainly caused cell cycle arrest, suppressed cell proliferation and induced apoptosis in Raji cells via MAPK pathway.

In this study, MNQ activated tumor suppressor, apoptosis and cell cycle regulation through upregulating PTEN, FOXO1, FOXO3, YWHAH, MTOR, ITGB1, BAD, and APC genes in PI3K pathway (Table 2, Fig. 2), respectively. Among these up-regulated genes, adenomatous polyposis coli (APC) was found to be the most highest expressed gene across all time points. APC is a tumor suppressor that indirectly regulate a number of key genes involved in cell proliferation. Besides, it also plays a role in cell migration [23]. PTEN is another tumour suppressor gene up-regulated by MNQ that generally involved in the regulation of cell cycle, and it is known to be the target of many anticancer drugs. FOXO1 and FOXO3 have been reported to trigger apoptosis via regulating genes that responsible for cell death such as BIM and PUMA [24]. BAD is a known pro-apoptotic gene that initiate apoptosis. YWHAH encodes for 14-3-3 protein eta has been reported to regulate a wide variety of signaling pathways, and mainly mediated these signaling pathways by binding to phosphoserine-containing proteins. MTOR that encode for mTOR complex 1 and 2 proteins has been reported to regulate cell growth, cell proliferation, cell motility and cell survival [25]. ITGB1 gene encodes Integrin beta-1 surface receptor that are involved in cell proliferation, cell adhesion and recognition as well as metastatic diffusion of tumor cells [26]. In our previous study [27], MNQ was demonstrated to induce apoptosis in A549 lung adenocarcinoma cell through oxidation triggered JNK and p38 MAPK signaling pathways. In this study, the abovementioned listed gene were up-regulated throughout all time points of treatment, this could explain Raji cells strived to survive upon the treatment of MNQ that creating the oxidative stress environment. Raji cells strived to survive via the abovementioned genes that are mainly involved in cell growth, proliferation, and survival. On the contrary, MNQ down-regulated PDGFRA, AKT2, AKT3, RPS6KA1, IGF1, and IGF1R, genes. PDGFRA gene has been reported to instruct the making of platelet-derived growth factor receptor alpha (PDGFRA). PDGFRA proteins activate signaling pathways to control many important cellular processes such as cell growth, division and survival. Mutation on PDGFRA has been reported in gastrointestinal stromal tumors and in chronic myeloid leukemia. Besides, highly phosphorylated PDGFRA has been observed in non-small cell lung cancer and rhabdomyosarcoma [22]. AKT2 and AKT3 are the serine/threonine kinases that regulate cellular metabolism, cell proliferation, cell survival, cell growth and angiogenesis.

These genes are highly activated in many cancer cells. The AKT protein kinase has been indicated to transduce growth factors to oncogenes and target the proteins to induce tumor development. Thus, inhibition of the AKT kinase is workable in cancer prevention as down-regulation of AKT signaling is a strategy to prevent cancer formation [28]. Akt kinases has been known to inhibit FOXOs transcriptional functions and contribute to cell survival, cell growth and cell proliferation. RPS6KA1 encodes ribosomal protein S6 kinase alpha-1 kinase that phosphorylate various substrates of the MAPK singling pathway, which is implicated in controlling cell growth and differentiation. RPS6KA1 is also known as RSK1 that present in cytoplasm. RSK1 is an ERK effector that is involved mainly in nuclear signaling, cell growth, survival, and cell
proliferation. It has been reported to regulate SRF, c-Fos, and Nur77 transcription factors. RSK1 stimulated cell growth by regulating c-Fos and promotes cyclin D1 expression. Besides, RSK1 has been demonstrated to inhibit neuronal NO synthase in response to mitogenic signaling. RSK1 is highly expressed in prostate and breast cancers. RSK1 triggered survival signals via the Ras/ERK signaling pathway and protected cancer cells from apoptosis. RSK1 is also involved in cell cycle regulation by phosphorylating CDKN1B and preventing CDKN1B translocation into nucleus and thus inhibited G1 progression. RSK1 also promoted cell survival by suppressing the functions of pro-apoptotic proteins BAD and DAPK1. MNQ down-regulated insulin-like growth factor 1 (IGF1) and insulin-like growth factor 1 receptor (IGF1R) significantly. IGF family is a humoral mediator of growth hormone (GH) [29]. The IGF signaling pathway that control endocrine system and regulate cell growth and development has been reported to have a pathogenic role in cancer [30]. IGF has been reported to increase cancer cells growth and cells resistance to chemotherapy and radiation by decreasing functions of targeted agents in GH independent manner [31]. IGF1R is highly expressed and biologically active in small cell lung cancer, pediatric high-grade gliomas, and Ewing’s Sarcoma [32].

For NF-κB pathway, MNQ was found to up-regulate TRADD, LTA, TNFRSF1A, TNFRSF10B, and TNF genes, and down-regulate NFκB2, NFκB1E, IL6, TIMP1, and BCL3 genes. LTA is a cytokine produced by lymphocytes that belong to tumor necrosis factor family. LTA bound to TNFRSF1A/TNFR1, TNFRSF1B/TNFBR, TNFRSF14 and induced apoptosis. TNFRSF1A is expressed in transformed cells to trigger apoptosis or inflammation [33]. Mutations in TNFRSF1A has been studied to display resistance towards apoptosis [34]. TNFRSF1A signals apoptosis through caspase-8 [35]. Tumor necrosis factor-a (TNFa) has been indicated to promote apoptosis in human endothelial cells through TNFRSF1A and trigger caspase-2 and p53 activation. It took part in variety cellular responses, such as survival, differentiation, proliferation, apoptosis, and migration [33]. TNFa is a pro-inflammatory cytokine regulated by TNFRSF1A and TNFRSF1B [36]. TNFRSF1A has been shown to promote the recruitment of TNFR-associated factors (TRAFs), FAS-associated via death domains (FADDs), and TNFR-associated via death domains (TRADDs). TNFRs interact with TRAFs, FADDs and TRADDs to control apoptosis. TNFa has been reported to bind with TNFRSF1A to induce apoptosis by activating caspase-8 [37]. Tumor necrosis factor receptor superfamily, member 10b (TNFRSF10B) is a member of the TNF-receptor superfamily, it has the similar function with TNFRSF1A. TNFRSF10B has been shown to interact with FADD [38], caspase 10, and caspase 8 and induced apoptosis in cancer cells. TRADD contains a death domain that interact with TNFRSF1A or TNFR1 to activate apoptosis. It has been reported to bind with TRAF2 to reduce the inhibitor-of-apoptosis proteins expression, and thus suppressed TRAF2-mediated apoptosis. TRADD activated apoptosis through FADD and caspase-8 activation [39]. STAT1, FADD, TNFRSF1A, and TNFRSF25 are the proteins interact with TRADD. NFκB1, NFκB2, NFκB1A, and NFκB1E are the genes encode for protein members of NF-κB family. Two angiogenesis genes of IL6 and TIMP1 were down-regulated by MNQ. IL6 is an inflammatory cytokine that modulate growth and differentiation in tumor cells through STAT3 signaling pathway. Overexpression of IL6 has been reported to be associated with tumor progression through inhibition of cancer cell apoptosis, stimulation of angiogenesis, and drug resistance. Clinical studies have been reported IL6 protein as a regulator that associate with many types of cancers such as multiple myeloma, non-small cell lung adenocarcinoma, prostate cancer, colorectal cancer, renal cell carcinoma, breast cancer, and ovarian cancer [40,41]. Tissue inhibitor of metallopeptinases 1 (TIMP1) has become a marker of prognosis and indicator for checking the clinical response in cancer treatment because TIMP1 is elevated in breast, colon, and prostate cancer patient plasma [42,43]. BCL3 was the only down-regulated proto-oncogene by MNQ. BCL3 is a proto-oncogene deregulated in solid tumors. BCL3 has been revealed to induce proliferation and inhibit apoptosis. BCL3 activated STAT3, an aggressive oncogene in human cancer and promoted metastasis [44,45]. In leukemia cells, BCL3 has been demonstrated to up-regulate myc genes and lead to the formation of aggressive B-cell leukemia. Other report mentioned that NF-κB activated BCL3 to act as transcriptional co-activator. BCL3 contains two transactivating domain and can form homodimers with NF-κB1 (p50) or NF-κB2 (p52). It has been reported that BCL3 was up-regulated by cytokines such as TNFa, interleukin (IL)-1, IL-4, IL-6, IL-10, IL-12, and adiponectin [46-54]. These cytokines were activated by AP1 [46] and STAT3 [45,55]. In addition, BCL3 was down-regulated by p53 [56].

In a nutshell, MNQ up-regulated apoptotic genes (BAD, EGR1, FOXO1, FOXO3, ITGB1, LTA, TNF, TNFRSF1A, TNFRSF10B, TRADD, and YWHAH), tumor suppressor genes (APC, RB1, and TP53), and cell cycle genes (CDKN1A, CDKN1B, CDKN2A, CDKN2D, MTOR, LAMTOR3, and PTEN). On the contrary, MNQ down-regulated anti-apoptotic genes (HSPB1, NFKB1, NFKB2, NFKB1A, NFKB1E, and RPS6KA1), angiogenesis genes (IL6 and TIMP1), cell cycle genes (ARAF, AKT2, AKT3, BRAF, EGFR, HRAS, IGF1, IGF1R, KRAS, NRAS, and PDGFA), transcription
factors (FOS and JUN), and proto-oncogenes (BCL3).

The findings of this study have revealed the capability of MNQ in regulating the key cancer genes in MAPK, PI3K, and NF-κB signaling pathways. The regulatory effects exerted by MNQ are targeting at cell proliferation, cell differentiation, cell transformation, induce apoptosis, reduce inflammatory responses, inhibits angiogenesis and metastasis. These results underline the need to further investigate the mechanism of actions of MNQ at each molecular level of these genes.

**ORCID**

Subramaniam Menaga: https://orcid.org/0000-0002-3388-9534
Chi-Ying F. Huang: https://orcid.org/0000-0003-4466-8486
Siong Hock Anthony Ho: https://orcid.org/0000-0002-8455-8604
Seng Chiew Gan: https://orcid.org/0000-0002-8271-3420
Yang Mooi Lim: https://orcid.org/0000-0003-1377-7655

**Authors' Contribution**

Conceptualization: YML. Data curation: TYW, SM. Formal analysis: TYW, YML, SHAH, SCG, CYFH. Writing - original draft: TYW. Writing - review & editing: TYW, SM, CYFH, SHAH, SCG, YML.

**Conflicts of Interest**

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

**Acknowledgments**

This work was supported by Fundamental Research Grant Scheme (FRGS) Grant (FRGS/1/11/SKK/UTAR/03/5) from Ministry of Higher Education (MOHE), Malaysia.

**References**


https://doi.org/10.5808/gi.21041


Introduction

In recent years, genome-wide association studies (GWAS) for longitudinal traits (e.g., body weight or cholesterol levels) have been carried out in cohorts, where multiple measurements have been collected from each individual [1-7]. Although GWAS have successfully discovered a large number of novel genetic variants associated with these traits, the identified variants typically account for only a small proportion of overall heritability [8-10]. A presumed explanation for the “missing heritability” is that existing methods have low power to identify gene-gene and gene-time/environment interactions [11]. Since traditional methodologies are limited to the identification of variants with marginal effects using a single measurement per individual, a large amount of useful information in longitudinal data is lost and variants that interact with other variants or have time-varying effects may not be detected [12]. It is more appropriate to analyze multiple variants simultaneously, using all available measurements, for longitudinal genetic studies.

There are methodological challenges associated with the genetic analysis of longitudinal traits for multiple variants. Most complex traits are typically controlled by multiple...
variants that interact with each other or environmental factors. It may be exceedingly difficult to model all candidate variants with epistatic effect and gene-time/environment interactions for longitudinal traits because genetic data are generally high-dimensional relative to the number of samples. Bayesian multiple quantitative trait loci (QTL) mapping methods [13-16] have been proposed for modeling epistatic effects. Multiple QTL can be simultaneously detected by treating the number of QTL as a random variable using the reversible jump Markov-chain Monte Carlo (MCMC) method [13,14]. Alternatively, multiple QTL can be viewed as a variable selection problem [15,16]. Bayesian model selection approaches are used for identifying QTL with main and epistatic effects [17], as well as QTL that interact with other covariates [18] based on the composite model space framework. These approaches use a fixed-dimensional parameter space by setting an upper bound on the number of detectable QTL and introduce latent binary variables for deciding which variables will be included in the model. This technique reasonably reduces the model space using efficient MCMC algorithms. For multiple QTL mapping with multivariate traits, Banerjee et al. [19] extended the Bayesian variable selection method of Yi [16] via a model that allows different genetic models for different traits. This method provides a multiple QTL mapping strategy for correlated traits, but it does not account for the dependence structure among repeated measurements from each subject.

Several statistical methods have been proposed for dealing with within-subject variation. For data collected at the same time points across all individuals, the measured values at each time point can be treated as one variable. The data can then be treated as multivariate outcomes and jointly analyzed [20-23]. For data collected at different time points across some or all individuals, the measured values cannot be effectively grouped; thus, standard multivariate analysis is no longer applicable. Alternatively, mixed models are used for longitudinal data to map QTL [24]. Mixed models are flexible in modeling such unbalanced data because they allow non-constant correlations among observations. Chung and Zou [25] developed a Bayesian multiple association-mapping algorithm based on a mixed model with a built-in variable selection feature. It models multiple genes simultaneously and allows gene-gene and gene-time/environment interactions for repeatedly measured phenotypes. However, in that model, we made the strong assumption that the covariance matrix is known up to a constant. We plan to relax that assumption here.

In this paper, we develop a Bayesian variable selection method for longitudinal data where phenotypes are not measured at a fixed set of time points for all samples. It jointly models the main and pairwise interactions of all candidate genetic variants. We propose a novel grid-based approach to parsimoniously model each subject’s covariance matrix as a function of a covariance matrix defined on a set of pre-selected time points where each observed time point is mapped to its two adjacent grid time points via linear interpolation. This approach thus deals only with a covariance matrix of a fixed dimension. The covariance matrix is then modeled nonparametrically using the modified Cholesky decomposition of Chen and Dunson [26], which facilitates the use of normal conjugate priors. The deviance information criterion (DIC) and the Bayesian predictive information criterion (BPIC) are proposed for the selection of an optimal number of grid points. The paper is organized as follows. In the Methods section, we introduce a novel grid-based Bayesian method for longitudinal genetic data and provide its theoretical basis. In the Results section, we show numerous simulation results using whole-genome sequencing data from the 1000 Genome Project to evaluate the performance of the proposed methods and assess the effects of sample size, number of variants, causal variants, and heritability. We conclude the paper with some discussions on the proposed methods and future research.

Methods

Genotype data
For our simulation studies, we utilized the whole-genome sequencing data from the 1000 Genome Project, which created a catalogue of common human variations using samples from people who provided open consent who declared themselves healthy. It ran between 2008 and 2015, generating a large public catalogue of human variations and genotype data. We randomly selected 400 out of 504 individuals of East Asian (EAS) ancestry from the 1000 Genome Project data (phase 3 version S) and then removed single-nucleotide polymorphisms (SNPs) with a minor allele frequency < 5% and p(Hardy-Weinberg equilibrium) < 10^-6, which resulted in 6,247,288 SNPs.

Bayesian mixed models
For a given trait, suppose we have n individuals where individual i has phenotypes measured at n_i time points (i = 1, ..., n) and p SNPs. Let N = Σ_i=1^n n_i. We set the number of main effect terms equal to p, the number of SNP-SNP interaction terms s to °p(p-1)/2°, and the number of SNP-covariate interaction terms to p_q, where q is the number of covariates in the model, including time. We define λ = (λ_1, ..., λ_s)^T as the SNP positions associated with the above genetic
effects, where \( d = p + \frac{p(p-1)}{2} + pq \). Each SNP can be associated with the trait through its main effect or interactions with other SNPs (epistatic effects) or covariates. We introduce latent binary variables \( \gamma = (\gamma_1, ..., \gamma_p)^T \) for the selection of genetic effects to be included in \( (\gamma_\gamma = 1) \) or excluded from \( (\gamma_\gamma = 0) \) the model. The vector \( (\gamma, \lambda) \) determines the number and positions of SNPs. For the \( i \)th individual, \( x_i \) denotes the \( n \times q \) design matrix of time/environmental covariates, \( x_{i0} \) denotes the \( n \times p \) design matrix of the \( p \) SNPs, \( x_{i0} \) denotes the \( n \times p \) design matrix of the epistatic effects, and \( x_{i0} \) denotes the \( n \times pq \) design matrix of the SNP-time/environment interactions. We define the final design matrix as \( x_i = (x_{i0}, x_{i1}, x_{i2}, x_{i3}) \).

Given \( \gamma, \lambda, \) and \( x_i \) we consider the following mixed model:

\[
y_i = \mu + x_i \beta + p + d + e_i(i = 1, ..., n), \tag{1}
\]

where \( y_i = (y_{i1}, ..., y_{in})^T \) is an \( n \times 1 \) phenotype vector of individual \( i \); \( \mu \) is an \( n \times 1 \) overall mean vector; \( T \) is a diagonal matrix with upper diagonal elements \( 1 \) (i.e., the model always contains all non-genetic covariates) and lower diagonal elements \( \gamma_j \beta = (\beta_1, \beta_2, \beta_{pq})^T \) is a vector of genetic effects, time/environmental effects, epistatic effects, and SNP-time/environment interactions; and \( e_i \) is an \( n \times 1 \) vector of random errors with \( e_i \sim N(0, \sigma^2 I_n) \). To model the correlation among repeated measurements of the same individual, we partition the observed time interval by \( k \) pre-specified grid points, \( t = (t_1, ..., t_k) \), and define \( x_i \) as a \( k \times 1 \) vector of random effects at the grid time points with \( x_i \sim N(0, D) \) where \( D = D_{kk} \) a \( k \times k \) covariance matrix. Let \( p = (p_{i1}, ..., p_{in})^T \) and \( p = \text{diag}(p_{i1}, ..., p_{in}) \) where \( p \) is defined as follows. If all subjects have \( k \) observations measured exactly on the \( k \) grid time points, then \( p \) becomes an identity matrix. We apply an interpolation procedure (e.g., linear, polynomial, or spline) to any observation that does not fall on any of the \( k \) grid time points. For simplicity, we choose a linear interpolation here. When the \( 0 \)th measurement of individual \( i \) falls at time \( t \), which is between the grid points \( t_0, t_1, ..., t_{k-1}, \), we set \( p_{i0} = \left( 0_{0-1}^T, t, -t, 0_{-1}^T \right)^T \). When \( t = t_0 \), we get \( p_{i0} = (0_{-1}^T, 1, 0_{-1}^T) \). We can re-express \( p_{i0} \) as \( p_{i0} = a_{i0} e_i \) where \( a_{i0} \) is the \( r \)th element of \( p_{i0} \) and \( e_i \) (1 \( \leq r \leq k \)) is a \( 1 \times k \) vector whose elements are all zero except the \( r \)th component, which equals 1. Note that \( e_i = \sum_{r=1}^{k} e_{ir} = 1 \), \( 0 \leq a_{i0} \), ..., \( a_{ik} \leq 1 \) and at most two adjacent \( a_{ir} \) values can be non-zero due to the linear interpolation we employ here.

Re-parameterized model

For Bayesian estimation of the mixed model (1), we factor \( D \), the covariance matrix of the random effects, by employing the modified Cholesky decomposition of Chen and Dunson [26]. Let \( L \) denote a \( k \times k \) lower triangular Cholesky decomposition matrix that has nonnegative diagonal elements, such that \( D = LL^T \). Let \( L = L \Delta \), where \( \Delta = \text{diag}(\delta_1, ..., \delta_k) \) and \( \Delta \) is a \( k \times k \) matrix with the \((i, m)\)th element denoted by \( \psi_{im} \). To make \( \Delta \) and \( \Psi \) identifiable, we make the following assumptions: \( \delta_i \geq 0 \), \( \psi_{im} = 1 \) and \( \psi_{im} = 0 \), for \( i = 1, ..., k \), \( m = i+1, ..., k \). These conditions make \( \Delta \) a nonnegative \( k \times k \) diagonal matrix and \( \Psi \) a lower triangular matrix with \( 1 \)’s in the diagonal elements. This leads to the decomposition \( D = \Delta \Psi \Psi^T \Delta \), and thus we reparameterize model (1) as

\[
y_i = \mu + x_i \beta + p \Delta \Psi b_i + e_i(i = 1, ..., n), \tag{2}
\]

where \( b_i = (b_{i1}, ..., b_{ik})^T \) such that \( b_i - N(0, 1) \) and \( b_{i1} \perp \perp b_{ij} \) (\( j \neq i \)), \( i = 1, ..., k \). For later use, we define \( v_i = p \Delta \Psi = (v_{i1}, ..., v_{in})^T \) and \( v = \text{diag}(v_{i1}, ..., v_{in}) \).

Model identifiability

Model identifiability is a property that a model must satisfy for accurate inference to be possible. A model is identifiable if it is theoretically possible to estimate the true values of the underlying parameters of the model, while a model is non-identifiable or unidentifiable if two or more parametrizations are observationally equivalent [27]. The proposed Bayesian model has an identifiability issue associated with the covariance matrix of \( y = (y_{i1}, ..., y_{in})^T \), which equals \( PDP^T + \sigma^2 I_n \) where \( D = I_{n \times n} D \). The condition is that \( PDP^T + \sigma^2 I_n = PDP^T + \sigma^2 I_n \) for only if \( \vec{D} = \hat{D} \) and \( \vec{D} = \vec{D} \). This is equivalent to the system of equations \( PDP^T + \sigma^2 I_n = 0 \) having no non-zero solutions for \( \vec{D} \) and \( \vec{D} \) when \( \vec{D} = \vec{D} \) and \( \vec{D} = \vec{D} \). Let the \((r, s)\)th element of \( \vec{D} = \vec{D} \). The system of equations \( PDP^T + \sigma^2 I_n = 0 \) is equivalent to the system of equations \( AX = 0 \), where \( A = (A_{i1}, ..., A_{ik})^T \) is a \( 2^r \times \left( \sum_i n_i + 1 \right) \times 2^r \) matrix whose elements are functions of the \( \vec{a}_{i1}, ..., \vec{a}_{ik} \). Therefore, the proposed Bayesian model (2) is identifiable if and only if \( \vec{A} = \left( 2^{-r} \right)^k \) (see proof of Theorem 1 in Supplementary Data 1).

Lemma 1 and Theorem 1 enable us to check whether a given model is identifiable. A toy example is provided below. Suppose there are 3 grid points that produce 2 time intervals. According to the theorem, the rank of \( A \) must be \( \lfloor \frac{1}{2} (3+1) \rfloor = 7 \) for the model to be identifiable. Suppose the phenotypes of all individuals are
observed exactly on the 3 grid points.\text{Then} 
\[ p_r = I_1, A_1 = \cdots = A_n = \begin{pmatrix} 10000001 \\ 01000000 \\ 00100000 \\ 00010000 \\ 00001000 \\ 00000100 \\ 00000011 \end{pmatrix} \text{and } X = \begin{pmatrix} \tilde{a}_{1,1} \\ \tilde{a}_{1,2} \\ \tilde{a}_{1,3} \\ \tilde{d}_{2,1} \\ \tilde{d}_{2,2} \\ \tilde{d}_{2,3} \\ \tilde{d}_{3,3} \end{pmatrix} \sigma^2.
\]

The rank of $A$ is $\frac{1}{2} \times 3(3+1) = 6$. Therefore, $PDP^T + \sigma^2 I_n$ is non-identifiable. If we have one additional individual who has one phenotype measured not on any of the grid points, the model becomes identifiable since the rank of $A$ now increases to 7. If we do not have any additional individuals, we can avoid the identifiability issue simply by setting $\sigma^2 = 0$ and modeling $D$ directly.

**Prior specifications**

For the random effects of the proposed Bayesian model, we employ the priors presented by Chen and Dunson [26]. Specifically, independent half normal priors are imposed on the diagonal elements of $\Delta$ and normal priors on the lower triangular elements of $\Psi$. For the fixed effects, we straightforwardly extend the priors presented in Yi et al. [17, 18].

**Priors on $\gamma$ and $\lambda$**

Let $\omega_r = P(\gamma_r = 1)$ be the inclusion probability of the $r$th genetic effect. We assume that all these inclusion probabilities are independent of each other and thus the prior of $\gamma$ is $\Pi_{r=1}^n \omega_r^\gamma (1-\omega_r)^{1-\gamma}$. The inclusion probability $\omega_r$ is pre-determined and can vary according to whether it corresponds to a main genetic effect, SNP-SNP interaction, or SNP-covariate interaction [17]. To specify a prior on $\lambda$, we assume that the locations are again independent and uniformly distributed over all SNPs. For the number of SNPs (i.e., $p$), the prior distribution of genetic variant location $\lambda$ is therefore given by $P(\lambda) = \Pi_{r=1}^p P(\lambda_r)$.

**Priors on $b$, $\Delta$, and $\Psi$**

In model (2), we let the distribution of each $b_i$ independently follow a standard normal distribution. Thus, the joint prior distribution of $b = (b_1, \cdots, b_r)^T$ is $P(b) \overset{d}{=} N(0, \Sigma_b)$. As priors for $\Delta$ and $\Psi$, we define two vectors $\delta = (\delta_l : l = 1, \cdots, k)^T$ and $\psi = (\psi_m : m = 2, \cdots, k; l = 1, \cdots, m-1)^T$. The prior distribution for $\delta$ is $P(\delta) = \Pi_{l=1}^k N(\delta_l | m_{\delta_l}, s_{\delta_l}^2)$, where $N(\delta_l | m_{\delta_l}, s_{\delta_l}^2)$ is the density of a half normal distribution that is a $N(\delta_l | m_{\delta_l}, s_{\delta_l}^2)$ density truncated below by zero. The prior distribution for $\psi$ is $P(\Psi) \overset{d}{=} N(\Psi_0, \Sigma_0)$, where $\psi_0$ and $\Sigma_0$ are pre-specified hyperparameters.

**Priors on $\beta$, $\mu$, and $\sigma^2$**

The prior for the $r$th genetic effect is a normal distribution, $P(\beta_r | y_r, \sigma^2_r) \overset{d}{=} N(0, \gamma_r \sigma^2_r)$ and the prior for the variance $\sigma^2_r$ is a scaled inverse $\chi^2$ distribution, $P(\sigma^2_r) \overset{d}{=} \text{inv-}\chi^2(\nu_r, s^2_r)$ whose expectation is $E(\sigma^2_r) = \frac{\nu_r s^2_r}{\nu_r - 2}$. The degree of freedom $\nu_r$ controls the skewness of the prior for $\sigma^2_r$ (we set $\nu_r = 6$) and the scale parameter $s^2_r$ controls the prior or confidence region for the heritability of the associated genetic factor. Let $V$ be the total phenotypic variance and $V_q$ be the sample variance of the column of $x$ associated with $\beta_r$. The heritability of the $r$th genetic factor, $h_r$ is therefore $V_r \beta_r^2 / V$. Setting $E(\sigma^2_r) = E(\beta_r^2)$, we have $s^2_r = (\nu_r - 2)E(\beta_r^2)/V_r = (\nu_r - 2)E(h_r) V_r (\nu_r V_r)$, with $E(h_r) = 0.1$. The prior for the overall mean $\mu$ is given by $P(\mu) \overset{d}{=} N(\eta_\mu, \tau^2_\mu)$.

We empirically set $\eta_\mu = \tilde{y} = \left(\frac{1}{n}\right)\sum_{r=1}^n y_r$ and $\tau^2_\mu = s^2_\mu = \left(\frac{1}{n-1}\right)\sum_{r=1}^n \sum_{l=1}^m (y_r - \bar{y})^2$. The prior for the residual variance $\sigma^2$ is chosen as an scaled inverse $\chi^2$ distribution, $P(\sigma^2) \overset{d}{=} \text{inv-}\chi^2(\nu_\sigma, s^2_\sigma)$.

**Posterior calculation and MCMC algorithm**

The joint posterior distribution is proportional to the product of the likelihood and the prior distributions of all unknown parameters, which can be expressed as

\[
P(\gamma, \theta | y) \propto P(y | \gamma, \theta) P(\gamma) P(\beta) P(\delta) P(\psi) P(\mu) P(\sigma^2),
\]

where $\theta = (\lambda, \beta, \delta, \psi, \mu, \sigma^2)^T$. To obtain MCMC samples of all parameters, we utilize the Metropolis-Hastings and Gibbs sampling algorithms, and alternately update each unknown parameter or set of unknown parameters conditional on all the other parameters and the observed data.

For $\gamma$ and $\lambda$, we use the Metropolis-Hastings algorithm within Gibbs sampler since their conditional distributions have no known distributional forms. To update those parameters, we straightforwardly extend the Metropolis-Hastings algorithm proposed by Yi et al. [18] for our Bayesian model. These algorithms are described in the Supplementary Data 1. For the other parameters, we applied the Gibbs sampling algorithm. Specifically, since $b$, $\delta$, and $\psi$ have multivariate normal or half normal priors, the full conditional distributions are easy to derive by their conjugacy properties. The full conditional posterior distributions of $b$, $\delta$ and $\psi$ are $P(b | y, \gamma, \theta, \delta, \psi) \overset{d}{=} N(\tilde{b}_b, \Sigma_b), P(\delta | y, \gamma, \theta, \delta, \psi) \overset{d}{=} N(\tilde{\delta}_b, \Sigma_b)$, and $P(\psi | y, \gamma, \theta, \delta, \psi) \overset{d}{=} N(\tilde{\psi}_b, \Sigma_b)$, respectively, where $\theta$ represents all the elements of $\theta$ except $f$. The expressions for $\tilde{b}_b, \Sigma_b, \tilde{\delta}_b, \Sigma_b, \tilde{\psi}_b, \psi$, and $\Sigma_b$ are again given in Supplementary Data 1. The full conditional distributions of $\beta$, $\sigma^2_r$, $\mu$
and $\sigma^2$ are $P(\beta_1|\gamma_1=1, \gamma_{-1}, \theta, \beta_{-1}, y) \overset{d}{=} N(\tilde{\beta}_0, \tilde{\sigma}^2)$, $P(\sigma|\tilde{\beta}_0) \overset{d}{=} \text{Inv}-\chi^2(\nu+1, (\tilde{\beta}_0^T\tilde{\beta}_0)/\nu)$, $P(\mu|\gamma, \theta, y) \overset{d}{=} (\mu^*, \sigma^*)$, and $P(\sigma|\gamma, \theta, \nu, \nu_t, \nu_s) \overset{d}{=} \text{Inv}-\chi^2(\nu+\nu_s+t$, $\nu_s+\nu_t+N\bar{\sigma}^2_{\nu_t, \nu_s})$, respectively, where $\tilde{\mu}_0, \tilde{\sigma}^2, \mu^*, \sigma^*$, and $\delta^2$ are given in Supplementary Data 1 as well.

Posterior analysis

The posterior samples can be used to approximate the posterior distribution of the parameters. MCMC samples from the initial iterations are discarded as “burn-in” and the subsequent samples are thinned by keeping every cth MCMC sample, where c is an integer, and discarding the rest. The posterior inclusion probability of each SNP can be calculated using its inclusion proportion in the MCMC samples as $P(\kappa_i|y) = \frac{1}{T}\sum_{t=1}^{T} I(\psi_0^\kappa_i = \kappa_i, \psi_0^\kappa_{-i} = 1)$ where $\kappa_i$ is the SNP position ($i=1, ..., h$) and $T$ is the total number of MCMC samples. With the prior $P(\kappa_i) = \frac{P}{h}$, the Bayes factor can be calculated to quantify the evidence for inclusion of the ith SNP ($\kappa_i$) against exclusion of the ith SNP as

$$BF(\kappa_i) = \frac{P(\kappa_i|y)/P(\kappa_i)}{(1-P(\kappa_i|y))/(1-P(\kappa_i))} = \frac{(P(\kappa_i|y))}{(1-P(\kappa_i|y))} \cdot \frac{(1-P(\kappa_i))}{P(\kappa_i)}.$$  

The Bayes factor $BF(\kappa_i)$ reflects how our belief in the importance of the ith SNP changes as we move from prior knowledge to posterior knowledge. Jeffreys [28] and Yandell et al. [29] suggest the following criteria for judging the significance of each SNP: weak support if $BF(\kappa_i)$ falls between 3 and 10; moderate support if $BF(\kappa_i)$ falls between 10 and 30; and strong support if $BF(\kappa_i)$ is larger than 30.

Choice of the number of grid points

A critical issue with the proposed Bayesian model is how to choose an optimal number of grid points, k. We achieve this goal by evaluating the goodness of the predictive distributions of our Bayesian models. Spiegelhalter et al. [30] proposed the DIC as $\text{DIC} = -2E_{\theta|y}[\log P(y|\gamma, \theta)] + P_p$. The second term of the DIC, $P_p$, is the effective number of parameters, which is defined as $P_p = -2E_{\theta|y}[\log P(y|\gamma, \theta)] + 2\log P(y|\tilde{\gamma}, \tilde{\theta})$, where $\tilde{\gamma}$ and $\tilde{\theta}$ are the posterior means of $\gamma$ and $\theta$. Since $P(y|\gamma, \theta) \overset{d}{=} N(\mu + x_i \bar{\beta}_i, \sigma^2 I_n)$ in model (1), the DIC is easy to compute with the MCMC samples. However, as stated by Robert and Titterington [31], the observed data are used twice to calculate $P_p$, and thus the predictive distribution from the DIC tends to overfit the data. To overcome the overfitting problem, Ando [32] developed the BPIC, which is defined as $\text{BPIC} = -2E_{\theta|y}[\log P(y|\gamma, \theta)] + 2\bar{n}b$ where $\bar{b}$ is the asymptotic bias in the posterior mean of the expected log-likelihood. Under a certain mild regularity condition, the bias term can be approximated by $\bar{b} = P_p$, resulting in the simplified $\text{BPIC} = 2E_{\theta|y}[\log P(y|\gamma, \theta)] + 2P_p$. It should be noted that the penalty term of the simplified BPIC is twice that of the original DIC. We select the optimal number of grid points for our model by minimizing DIC or simplified BPIC scores.

Implementation in gridbayes

The proposed grid-based Bayesian mixed models have been implemented in an R package named gridbayes [33], which is built on top of the R packages, qtlbim [34] and qtlbim [29]. The MCMC algorithm in C and the data manipulation procedure in R were modified for longitudinal analysis. The gridbayes package employs both DIC and simplified BPIC scores to select the optimal number of grid points. The software package and the source code are available for download at https://github.com/wonilchung/GridBayes.

Results

Simulation I

To evaluate the performance of the proposed method, we conducted the following simulations. We first used 400 individuals and 1,000 SNPs from the 1000 Genome Project data. The number of measurements for each individual ranged from 3 to 7 and the total number of observations was set to 2,000. Six different setups (Setups 1–6) were considered. We simulated the datasets containing 10 causal SNPs, which are randomly selected (i.e., the proportion of causal SNPs = 1%) with only main effects (Setup 1). For individual i, the phenotype values were generated from the model: $y_i = \epsilon_i \cdot \left(\sum_{a=1}^{10} x_i + t_i \right) + p_i \gamma + \epsilon_i$, where $x_i$ ($a=1, ..., 10$) were genotype values of the causal SNPs $c_i$ is used to set trait-heritability to 40%, $t_i = \left(t_{i1}, ..., t_{in}\right)^T$ were the time covariates generated from the uniform distribution $U[0, 1]$ and then standardized to have mean 0 and variance 1, and $\epsilon_i \sim N(0, \sigma^2 I_n)$. We set $\sigma^2 = 1$. The true number of grid points was set to 3 (i.e., true $k = 3$), and $p_i$ was calculated from $t_i$ by the linear interpolation as we described in the Methods section. We set $\delta = (\delta_{11}, \delta_{22}, \delta_{33}) = (1, 1, 2, 0.8)$ and $\psi = (\psi_{11}, \psi_{12}, \psi_{13}) = (0.6, 0.4, 0.6)$. That is, $\psi \sim N(0, D)$ with $\text{diag}(D) = (1, 1.96, 0.97)$ and the lower triangle elements $(d_{12}, d_{13}, d_{23}) = (0.72, 0.32, 0.81)$. The prior distributions for the elements in $\delta$ were independent $N(0, 30)$ and the prior distributions for the elements in $\psi$ were independent $N(0, 0.5)$. For each simulated dataset, the MCMC algorithm ran for $4 \times 10^5$ iterations after discarding the first 1,000 burn-in iterations. The remaining samples were further thinned for every 40 iterations, yielding $10^4$ MCMC samples for

https://doi.org/10.5808/gi.21080
the posterior analysis.

To further investigate the Bayesian mixed model, we analyzed additional datasets containing two SNP-SNP interactions (Setup 2), five SNP-SNP interactions (Setup 3), two SNP-time interactions (Setup 4), five SNP-time interactions (Setup 5), or ten SNP-time interactions (Setup 6). Specifically, we simulated data according to the following models:

\[ y_i = c_2 \left( \sum_{a=1}^{5} x_{ia} x_{ia} + x_{ia} x_{ia} + x_{ia} x_{ia} + t_i \right) + p \psi_1 + e_i \] for Setup 2,

\[ y_i = c_3 \left( \sum_{a=1}^{5} x_{ia} x_{ia} + x_{ia} x_{ia} + x_{ia} x_{ia} + x_{ia} x_{ia} + t_i \right) + p \psi_1 + e_i \] for Setup 3,

\[ y_i = c_4 \left( \sum_{a=1}^{5} x_{ia} x_{ia} + \sum_{a=1}^{5} x_{ia} x_{ia} \right) + p \psi_1 + e_i \] for Setup 4,

\[ y_i = c_5 \left( \sum_{a=1}^{5} x_{ia} x_{ia} + \sum_{a=1}^{5} x_{ia} x_{ia} \right) + p \psi_1 + e_i \] for Setup 5 and

\[ y_i = c_6 \left( \sum_{a=1}^{5} x_{ia} x_{ia} + t_i \right) + p \psi_1 + e_i \] for Setup 6. In our simulations, \( c_g \) were varied to ensure that trait-heritability to 40%. To display time-dependent SNP effects for Setups 4 and 5, we compared the time-dependent curves of averaged phenotype values for three different genotypes (0, 1, 2) at the first causal SNP (with no SNP-time interaction) and 10th one (with SNP-time interaction).

Supplementary Fig. 1 clearly showed that the first causal SNP had only a main effect, but the 10th causal SNP interacted with time. We first conducted gridbayes [33] using all the data. For model comparisons, we then conducted qtlbim [29] in two ways: once on a subset of each simulated data, where only one measurement from each subject was randomly selected, and once with all the data by (incorrectly) assuming that all the measurements were independent. We named the two qtlbim analyses “qtlbim-sub” and “qtlbim-all,” respectively.

The one-dimensional genome-wide profiles of 2log(BF) for the combined main, epistatic effects, and SNP-time interactions of each SNP under the six setups were presented in Figs. 1 and 2. The dashed vertical lines indicate the locations of the 10 causal SNPs. The gridbayes analysis of all the data and qtlbim-sub detected the causal SNPs reasonably well, but gridbayes clearly outperformed qtlbim-sub in general. The qtlbim-all method occasionally identified the true causal SNPs, but it produced far more false-positive findings than gridbayes and qtlbim-sub.

To evaluate the performance of our Bayesian model, we further calculated the receiver operating characteristic (ROC) curves. For each setup, we conducted 100 simulations. The ROC curves with a false-positive rate less than 0.2 are presented in Fig. 3. The solid lines represent the results of gridbayes, the dot-dashed lines correspond to qtlbim-sub and the results from qtlbim-all are summarized by the long-dashed lines. The ROC curves demonstrated that gridbayes with all measurements appeared to outperform the qtlbim analyses in terms of improved true positive rates.

To diagnose the convergence of the MCMC samples, we conducted 10 parallel chains with different, over-dispersed initial values with respect to the true posterior distribution. Using 10^4 iterations, Geweke’s Z-scores [35] for each chain based on the first 10% and last 50% of the samples indicated good convergence of all parameters. Based on 10 chains, Gelman and Rubin’s potential scale reduction factors [36] were calculated, and the upper limits were less than 1.01 for all parameters. Supplementary Fig. 2 presents the trace plots of \( \psi_i \), \( \delta_1 \), \( \delta_2 \), \( \delta_3 \), \( \psi_{11} \), \( \psi_{21} \), \( \psi_{31} \), and \( \psi_{32} \) for each setup, showing that all chains moved around the true values for all parameters, indicating good convergence. We plotted the marginal posterior and prior densities of all parameters based on 10, 000 random draws (Supplementary Fig. 3). It appeared that the random draws were approximately normal, with means close to the simulated values. Supplementary Fig. 4 displays the 95% highest posterior density (HPD) intervals for \( \psi_i \), \( \delta_1 \), \( \delta_2 \), \( \delta_3 \), \( \psi_{11} \), \( \psi_{21} \), \( \psi_{31} \), and \( \psi_{32} \) for each setup. Most of the 95% HPD intervals contained the corresponding true values. Table 1 summarizes the posterior estimates of all parameters. The posterior means and medians were close to the true values and all the 95% HPD intervals contained the true values, demonstrating the good performance of our algorithm.

**Simulation II**

We conducted another simulation to estimate the number of true grid points using the DIC [30] and simplified BPIC [32,37]. The settings were almost the same as those in the previous simulations, except that the true number of grid points now varied from 2 to 4 (i.e., true \( k = 2, 3, 4 \)). We simulated 100 datasets with 400 individuals and 1,000 SNPs containing 10 causal SNPs (i.e., the proportion of causal SNPs = 1%) with only main effects. The causal SNPs were randomly assigned. The trait-heritability was set to 40%. The phenotype values were generated from the model:

\[ y_i = c_g \left( \sum_{a=1}^{5} x_{ia} x_{ia} \right) + p \psi_1 + e_i \], where \( x_{ia} \) are genotypes of the causal SNPs and \( t_i = (t_{i1}, ..., t_{i6})^T \) are the time points of the \( i \)th individual. We set \( (\delta_1, \delta_2, \delta_3, \delta_4) = (1, 1.2, 0.8, 0.7) \) and \( (\psi_{11}, \psi_{12}, \psi_{21}, \psi_{22}, \psi_{31}, \psi_{32}) = (0.6, 0.4, 0.6, 0.2, 0.4, 0.6) \). Table 2 shows the average DIC, simplified BPIC scores over 100 simulations, and the proportion of times that the number of true grid points was correctly selected. All average DIC and average BPIC scores achieved the minimums at the true grid point number, and the percentages correctly selecting the true number of true grid points were 79%, 91%, and 100% for setups with 2, 3, and 4 true grid points using the DIC, and 94%, 98%, and 93% using the simplified BPIC. This illustrated the usefulness of the DIC and simplified BPIC in selecting the true number of grid points.
For a more detailed evaluation of our Bayesian method, we conducted the following simulations with 100 replications for each scenario. We first considered 400 individuals with three to seven time points, resulting in 2,000 observations, and decreased the sample size from 400 to 100 to assess the effect of sample size in ROC curves (Fig. 4A). The simulation data contained 1,000 SNPs with 1% causal SNPs (i.e., 10 causal SNPs) with only main effects. The trait values were generated from the model: 
\[ y_i = c_1 \sum_{a} x_i a \cdot t_i + p v_i + e_i, \]
where \( x_{i a} \) are genotypes of the causal SNPs and \( t_i = (-t_{i1}, ..., t_{in})^T \) are the time points of the \( i \)th individual. As in the previous simulations, we set the number of grid points to \( k = 3 \) and \( \sigma^2 = 1 \), \( \delta = (\delta_1, \delta_2, \delta_3) = (1, 1.2, 0.8) \), \( \psi = (\psi_{21}, \psi_{31}, \psi_{32}) = (0.6, 0.4, 0.6) \). The trait-heritability was set to 40%. As the sample size decreased from 400 to 100, the true positive rates decreased in ROC curves, indicating that including more samples increased the true positive rates with fixed false positive rates. Next, we evaluated the effect of the number of SNPs (Fig. 4B). The simulation data were

**Simulation III**

For a more detailed evaluation of our Bayesian method, we conducted the following simulations with 100 replications for each scenario. We first considered 400 individuals with three to seven time points, resulting in 2,000 observations, and decreased the sample size from 400 to 100 to assess the effect of sample size in ROC curves (Fig. 4A). The simulation data contained 1,000 SNPs with 1% causal SNPs (i.e., 10 causal SNPs) with only main effects. The trait values were generated from the model: 
\[ y_i = c_1 \sum_{a} x_i a \cdot t_i + p v_i + e_i, \]
where \( x_{i a} \) are genotypes of the causal SNPs and \( t_i = (-t_{i1}, ..., t_{in})^T \) are the time points of the \( i \)th individual. As in the previous simulations, we set the number of grid points to \( k = 3 \) and \( \sigma^2 = 1 \), \( \delta = (\delta_1, \delta_2, \delta_3) = (1, 1.2, 0.8) \), \( \psi = (\psi_{21}, \psi_{31}, \psi_{32}) = (0.6, 0.4, 0.6) \). The trait-heritability was set to 40%. As the sample size decreased from 400 to 100, the true positive rates decreased in ROC curves, indicating that including more samples increased the true positive rates with fixed false positive rates. Next, we evaluated the effect of the number of SNPs (Fig. 4B). The simulation data were
generated with 400 individuals, 1% causal SNPs, and 40% trait-heritability. As the number of SNPs increased from 1,000 to 5,000 (i.e., the corresponding number of causal SNPs increased from 10 to 50), the true positive rates decreased, meaning that the inclusion of more SNPs decreased the true positive rates. We then examined the effect of the proportion of causal SNPs (Fig. 4C). The sample size and number of SNPs were fixed to 400 and 1,000, and the trait-heritability was set to 40%. The true positive rates decreased as the proportion of causal SNPs increased from 1% to 5% (i.e., the corresponding number of causal SNPs increased from 10 to 50) because per-SNP heritability—or the average proportion of phenotypic variation explained by a single SNP—decreased as the proportion of causal SNPs increased while keeping trait-heritability constant. Lastly, to demonstrate the effect of trait-heritability, we considered a setting where the sample size, number of SNPs, and proportion of causal SNPs were 400, 1,000, and 1%, respectively. We then changed trait-heritability from 40% to 10% in Fig. 4D. The true-positive rates decreased as trait-heritability decreased, showing...
that larger heritability increased the true positive rates. Supplementary Tables 1, 2, 3, and 4 summarize the posterior means, medians, standard deviations and 95% HPD intervals of all parameters in the simulations for sample size, number of SNPs, proportion of causal SNPs, and heritability, respectively. The posterior means and medians were close to the true values, and all the 95% HPD intervals contained the true values, indicating that our Bayesian method performed well.

Supplementary Table 5 showed the average DIC and simplified BPIC scores over 100 replications for all simulations. Table 3 summarizes the simulation settings for all simulation setups based on genetic effect terms, the number of grid points, sample size, number of observations, number of SNPs, number of causal SNPs, and trait-heritability.

Discussion

We developed a grid-based Bayesian mixed model for longitudinal genetic data with a built-in variable selection feature. The proposed Bayesian method modeled multiple candidate SNPs simultaneously and allowed SNP-SNP and SNP-time interactions, which enabled us to identify SNPs with time-varying effects. Such
SNPs are of great scientific and medical interest. In addition, we proposed a new grid-based method to model the covariance structure nonparametrically. Not only is the proposed method parsimonious in estimating the covariance matrix, but also by employing a reasonable number of grid-points, it can flexibly approximate any type of covariance structure. The number of grid points was preset, but DIC and simplified BPIC can be used to select the optimal number.

The simulation studies showed that the proposed Bayesian method using all time points outperformed the ordinary Bayesian method with one or all time points included. As expected, the proposed method that utilized the full data was more powerful than the cor-

### Table 1. Posterior means, medians, standard deviations, and 95% HPD intervals of the parameters for random errors and random effects in the simulation study

<table>
<thead>
<tr>
<th>Setup</th>
<th>Par</th>
<th>True</th>
<th>Mean</th>
<th>Med</th>
<th>SD</th>
<th>95% HPD</th>
<th>Setup</th>
<th>Par</th>
<th>True</th>
<th>Mean</th>
<th>Med</th>
<th>SD</th>
<th>95% HPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\sigma^2$</td>
<td>1</td>
<td>1.01</td>
<td>1.01</td>
<td>0.04</td>
<td>0.92 to 1.09</td>
<td>2</td>
<td>$\sigma^2$</td>
<td>1</td>
<td>1.01</td>
<td>1.01</td>
<td>0.04</td>
<td>0.93 to 1.10</td>
</tr>
<tr>
<td></td>
<td>$\delta_1$</td>
<td>1</td>
<td>0.98</td>
<td>0.98</td>
<td>0.11</td>
<td>0.77 to 1.19</td>
<td></td>
<td>$\delta_1$</td>
<td>1</td>
<td>0.98</td>
<td>0.98</td>
<td>0.10</td>
<td>0.78 to 1.19</td>
</tr>
<tr>
<td></td>
<td>$\delta_2$</td>
<td>1.2</td>
<td>1.19</td>
<td>1.19</td>
<td>0.13</td>
<td>0.92 to 1.43</td>
<td></td>
<td>$\delta_2$</td>
<td>1.2</td>
<td>1.19</td>
<td>1.19</td>
<td>0.13</td>
<td>0.92 to 1.43</td>
</tr>
<tr>
<td></td>
<td>$\delta_3$</td>
<td>0.8</td>
<td>0.76</td>
<td>0.76</td>
<td>0.13</td>
<td>0.50 to 1.00</td>
<td></td>
<td>$\delta_3$</td>
<td>0.8</td>
<td>0.76</td>
<td>0.76</td>
<td>0.13</td>
<td>0.48 to 0.97</td>
</tr>
<tr>
<td></td>
<td>$\psi_{31}$</td>
<td>0.6</td>
<td>0.65</td>
<td>0.63</td>
<td>0.21</td>
<td>0.31 to 1.13</td>
<td></td>
<td>$\psi_{31}$</td>
<td>0.6</td>
<td>0.67</td>
<td>0.64</td>
<td>0.21</td>
<td>0.33 to 1.14</td>
</tr>
<tr>
<td></td>
<td>$\psi_{31}$</td>
<td>0.4</td>
<td>0.52</td>
<td>0.51</td>
<td>0.24</td>
<td>0.09 to 1.04</td>
<td></td>
<td>$\psi_{31}$</td>
<td>0.4</td>
<td>0.51</td>
<td>0.49</td>
<td>0.24</td>
<td>0.07 to 1.01</td>
</tr>
<tr>
<td></td>
<td>$\psi_{32}$</td>
<td>0.6</td>
<td>0.56</td>
<td>0.53</td>
<td>0.27</td>
<td>0.11 to 1.18</td>
<td></td>
<td>$\psi_{32}$</td>
<td>0.6</td>
<td>0.67</td>
<td>0.64</td>
<td>0.29</td>
<td>0.20 to 1.34</td>
</tr>
<tr>
<td>3</td>
<td>$\sigma^2$</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>0.04</td>
<td>0.92 to 1.08</td>
<td>4</td>
<td>$\sigma^2$</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>0.04</td>
<td>0.92 to 1.09</td>
</tr>
<tr>
<td></td>
<td>$\delta_1$</td>
<td>1</td>
<td>1.06</td>
<td>1.06</td>
<td>0.10</td>
<td>0.85 to 1.26</td>
<td></td>
<td>$\delta_1$</td>
<td>1</td>
<td>0.99</td>
<td>0.99</td>
<td>0.10</td>
<td>0.78 to 1.19</td>
</tr>
<tr>
<td></td>
<td>$\delta_2$</td>
<td>1.2</td>
<td>1.20</td>
<td>1.20</td>
<td>0.13</td>
<td>0.94 to 1.44</td>
<td></td>
<td>$\delta_2$</td>
<td>1.2</td>
<td>1.18</td>
<td>1.19</td>
<td>0.13</td>
<td>0.92 to 1.42</td>
</tr>
<tr>
<td></td>
<td>$\delta_3$</td>
<td>0.8</td>
<td>0.74</td>
<td>0.74</td>
<td>0.12</td>
<td>0.49 to 0.98</td>
<td></td>
<td>$\delta_3$</td>
<td>0.8</td>
<td>0.74</td>
<td>0.74</td>
<td>0.13</td>
<td>0.48 to 0.98</td>
</tr>
<tr>
<td></td>
<td>$\psi_{31}$</td>
<td>0.6</td>
<td>0.69</td>
<td>0.67</td>
<td>0.20</td>
<td>0.36 to 1.14</td>
<td></td>
<td>$\psi_{31}$</td>
<td>0.6</td>
<td>0.62</td>
<td>0.60</td>
<td>0.20</td>
<td>0.29 to 1.07</td>
</tr>
<tr>
<td></td>
<td>$\psi_{31}$</td>
<td>0.4</td>
<td>0.65</td>
<td>0.64</td>
<td>0.24</td>
<td>0.22 to 1.17</td>
<td></td>
<td>$\psi_{31}$</td>
<td>0.4</td>
<td>0.47</td>
<td>0.45</td>
<td>0.24</td>
<td>0.03 to 0.96</td>
</tr>
<tr>
<td></td>
<td>$\psi_{32}$</td>
<td>0.6</td>
<td>0.65</td>
<td>0.62</td>
<td>0.27</td>
<td>0.19 to 1.27</td>
<td></td>
<td>$\psi_{32}$</td>
<td>0.6</td>
<td>0.65</td>
<td>0.61</td>
<td>0.29</td>
<td>0.18 to 1.31</td>
</tr>
<tr>
<td>5</td>
<td>$\sigma^2$</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>0.04</td>
<td>0.92 to 1.09</td>
<td>6</td>
<td>$\sigma^2$</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>0.04</td>
<td>0.92 to 1.09</td>
</tr>
<tr>
<td></td>
<td>$\delta_1$</td>
<td>1</td>
<td>0.98</td>
<td>0.98</td>
<td>0.10</td>
<td>0.77 to 1.18</td>
<td></td>
<td>$\delta_1$</td>
<td>1</td>
<td>0.96</td>
<td>0.96</td>
<td>0.11</td>
<td>0.75 to 1.17</td>
</tr>
<tr>
<td></td>
<td>$\delta_2$</td>
<td>1.2</td>
<td>1.20</td>
<td>1.20</td>
<td>0.13</td>
<td>0.93 to 1.43</td>
<td></td>
<td>$\delta_2$</td>
<td>1.2</td>
<td>1.18</td>
<td>1.19</td>
<td>0.13</td>
<td>0.92 to 1.41</td>
</tr>
<tr>
<td></td>
<td>$\delta_3$</td>
<td>0.8</td>
<td>0.72</td>
<td>0.72</td>
<td>0.13</td>
<td>0.47 to 0.97</td>
<td></td>
<td>$\delta_3$</td>
<td>0.8</td>
<td>0.75</td>
<td>0.75</td>
<td>0.13</td>
<td>0.48 to 1.01</td>
</tr>
<tr>
<td></td>
<td>$\psi_{31}$</td>
<td>0.6</td>
<td>0.63</td>
<td>0.60</td>
<td>0.20</td>
<td>0.29 to 1.09</td>
<td></td>
<td>$\psi_{31}$</td>
<td>0.6</td>
<td>0.61</td>
<td>0.58</td>
<td>0.20</td>
<td>0.27 to 1.07</td>
</tr>
<tr>
<td></td>
<td>$\psi_{31}$</td>
<td>0.4</td>
<td>0.46</td>
<td>0.45</td>
<td>0.25</td>
<td>0.01 to 0.99</td>
<td></td>
<td>$\psi_{31}$</td>
<td>0.4</td>
<td>0.32</td>
<td>0.31</td>
<td>0.24</td>
<td>-0.14 to 0.81</td>
</tr>
<tr>
<td></td>
<td>$\psi_{32}$</td>
<td>0.6</td>
<td>0.65</td>
<td>0.61</td>
<td>0.29</td>
<td>0.18 to 1.31</td>
<td></td>
<td>$\psi_{32}$</td>
<td>0.6</td>
<td>0.67</td>
<td>0.63</td>
<td>0.30</td>
<td>0.19 to 1.35</td>
</tr>
</tbody>
</table>

HPD, highest posterior density; Par, parameters; True, true values of parameters; Med, median; SD, standard deviation.

### Table 2. Average DIC scores and simplified BPIC scores over 100 replications and the proportion selecting the model with the correct number of grid points using the proposed Bayesian model

<table>
<thead>
<tr>
<th>True k</th>
<th>k</th>
<th>Avg DIC</th>
<th>#Sel (%)</th>
<th>Avg Sim BPIC</th>
<th>#Sel (%)</th>
<th>Avg PD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>6,614.47</td>
<td>79</td>
<td>6,681.78</td>
<td>94</td>
<td>67.32</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6,617.97</td>
<td>15</td>
<td>6,690.51</td>
<td>6</td>
<td>72.54</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6,623.02</td>
<td>6</td>
<td>6,700.99</td>
<td>0</td>
<td>77.98</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>6,745.26</td>
<td>0</td>
<td>6,812.49</td>
<td>0</td>
<td>67.23</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6,696.53</td>
<td>91</td>
<td>6,769.83</td>
<td>98</td>
<td>73.30</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6,707.12</td>
<td>9</td>
<td>6,785.14</td>
<td>2</td>
<td>78.02</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>6,745.07</td>
<td>0</td>
<td>6,814.76</td>
<td>0</td>
<td>69.70</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6,718.66</td>
<td>0</td>
<td>6,792.75</td>
<td>7</td>
<td>74.09</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6,695.41</td>
<td>100</td>
<td>6,775.37</td>
<td>93</td>
<td>79.96</td>
</tr>
</tbody>
</table>

DIC, deviance information criterion; BPIC, Bayesian predictive information criterion; Avg DIC, average deviance information criterion scores over 100 replications; #Sel (%), proportion selecting the model with the correct number of grid points; Avg Sim BPIC, average simplified Bayesian predictive information criterion scores over 100 replications; Avg PD, average $P_D$. 

https://doi.org/10.5808/gi.21080
Fig. 4. Receiving operating characteristic curve analyses in the simulation study for sample size, number of single-nucleotide polymorphisms (SNPs), proportion of causal SNPs, and heritability. (A) We decreased the sample size from \( n = 400 \) (total number of observations, \( N = 2,000 \)) to \( n = 100 \) (\( N = 1,000 \)) for accessing the effect of sample size in receiver operating characteristic curves. The simulation data contained \( p = 1,000 \) SNPs, \( c = 1\% \) causal SNPs and \( h^2 = 40\% \) trait-heritability. (B) We increased the number of SNPs from \( p = 1,000 \) to \( p = 5,000 \) to evaluate the effect of number of SNPs. The simulation data contained \( N = 2,000 \) observations, \( c = 1\% \) causal SNPs, and \( h^2 = 40\% \) trait-heritability. (C) We increased the proportion of causal SNPs from \( c = 1\% \) to 5\%. The simulation data contained \( N = 2,000 \) observations, \( p = 1,000 \) SNPs, and \( h^2 = 40\% \) trait-heritability. (D) We decreased the trait-heritability from \( h^2 = 40\% \) to \( h^2 = 10\% \). The simulation data contained \( N = 2,000 \) observations, \( p = 1,000 \) SNPs, and \( c = 1\% \) causal SNPs.

A responding univariate analysis method that only used a subset of the data. Furthermore, the proposed Bayesian method performed better than the ordinary Bayesian method because our method modeled the within-subject correlation. Further simulation studies showed that statistical power increased as the data had more samples, a smaller number of SNPs, a lower proportion of causal SNPs, and larger trait-heritability. For our simulation studies, we utilized data from the 1000 Genome Project. With only 400 independent samples of EAS ancestry, we restricted out analysis with up to 5,000 SNPs. With a sufficient sample size, our method can be applied to all available SNPs. We are currently developing a parallel computing algorithm based on the message passing interface to execute multiple groups of SNPs simultaneously. This will make it feasible to apply our method to large-sample GWAS data.
Another important issue to mention is Bayesian model identifiability. In the Bayesian community, there is a wide diversity of views on the identifiability issue. Lindley [38] remarked that non-identifiability causes no real difficulty in Bayesian approaches. Poirier [39] and Eberly and Carlin [40] argued that a Bayesian analysis of a non-identifiable model is always possible if priors on all of the parameters are proper, since proper priors yield proper posterior distributions, and hence every parameter can be well-estimated. However, if the priors imposed on any non-identifiable model are not proper, or too close to being improper, ill-behaved posterior distributions may be generated such that the trajectory of the parameters can drift to extreme values, as demonstrated by Gelfand and Sahu [41]. In this paper, we investigated the identifiability of our Bayesian model, which motivated us to utilize only proper priors (see the Methods section). Non-identifiability occurred when the number of the grid points equaled the number of observed time points (see Supplementary Data 1), but we found that the posterior distribution behaved well due to the proper priors employed.

ORCID

Wonil Chung: https://orcid.org/0000-0002-5766-6247
Youngkwang Cho: https://orcid.org/0000-0002-7980-4145

Authors’ Contribution

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

Conflicts of Interest

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

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (2020R1C1A01012657) and Basic Science Research Program through...
the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A1A10044154). This work was supported by the Soongsil University Research Fund.

**Supplementary Materials**

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

**References**

29. Yandell BS, Mehta T, Banerjee S, Shriner D, Venkataraman R,
Introduction

It is important to understand the causality between two phenotypes to uncover the pathogenesis of diseases. Some strategies exist for assessing causality in epidemiological studies. Mendelian randomization (MR) is a technique that uses genetic variants as instrumental variables (IVs) to estimate the causal effect of an exposure on an outcome [1]. In accordance with Mendel’s laws of inheritance, alleles are randomly inherited from parents. Therefore, the genotypes of offspring can be considered independent of confounding factors. Furthermore, the fact that genotypes are fixed and are not affected by phenotypes obviates the reverse causation problem. For these reasons, genetic variants naturally meet many of the basic assumptions of IVs.

Summary statistics released from large genome-wide association studies recently began to facilitate MR by providing exposure effect sizes for multiple genetic variants [2]. The
type of MR analysis using an external dataset for quantifying exposure effect is called a two-sample MR design (2SMR). An advantage of 2SMR is that the statistical power can be increased by merging summary statistics from various sources including large consortia such as the UK Biobank [3]. The causal effect between an exposure and an outcome is estimated by the ratio between the reported genetic effect to the exposure in an external dataset and the observed genetic effect to the outcome in the target dataset. Since there are multiple variants, the ratio estimates over multiple variants are usually combined into a single estimate via the inverse-variance weighted method.

In 2SMR, the standard error of the estimated ratio is conventionally approximated by the first-order term from the delta method. As stated by Thomas et al. [4], however, this approximation can lead to an underestimation of the variance. This underestimation can lead to both increased power and an increased false-positive rate (FPR). An alternative is to use the second-order approximation of the standard error, which can considerably correct for the deviation of the first-order approximation.

In this study, we extensively simulate MR to show the impact of this first-order approximation on the FPR and power of MR. We simulate several different situations to evaluate which study design parameters affect the errors of the first-order approximation, and also compare the errors of the first-order approximation to those of the second-order approximation.

Methods

Genetic variants as instrumental variables

Genetic variants such as single-nucleotide polymorphisms (SNPs) have several properties that make them appropriate as an instrument of exposure. The random inheritance of the alleles makes the genotype distribution independent of socio-economic factors and lifestyle factors such as income [5]. Inherited alleles are not changed from birth by diseases or conditions, except in rare cases of somatic mutations. However, some assumptions still need to be satisfied to ensure the validity of a genetic variant as an IV (Fig. 1).

Three basic assumptions must hold for a genetic variant to be used as an IV for MR [6].

IV1. The genetic variant is associated with the exposure.

IV2. The genetic variant influences the outcome only through the exposure.

IV3. The genetic variant is independent of confounding factors affecting the exposure-outcome relationship.

Whether these assumptions are satisfied in various conditions has been discussed elsewhere [7]. Herein, we simply accept these assumptions and proceed to the description of MR.

Basic model of MR and the first-order approximation of variance

In this section, we describe the basic model of MR along with the commonly used first-order variance approximation (Fig. 1). Let G be an IV (e.g., a SNP), X be an exposure such as body mass index, and Y be an outcome, such as disease. We can set the relationships between variables (G, X, and Y) via a linear regression model.

\[
X \mid G = \beta_{X0} + \beta_{XG}G + \epsilon_X \\
Y \mid G = \beta_{Y0} + \beta_{YG}G + \epsilon_Y
\]

If we assume that all IV assumptions are satisfied, then \( \beta_{X0} = 0 \) because of IV1 and \( \beta_Y = \beta_X \times \beta \) because of IV2 and IV3. That is, G (Fig. 1) affects Y (outcome) only through X (exposure). It is assumed that the error terms \( \epsilon_X \) and \( \epsilon_Y \) follow normal distributions and are independent in the case of 2SMR of two disjoint samples. Even in the case of two non-overlapping samples, a report has stated the sample correlation between \( \hat{\beta}_X \) and \( \hat{\beta}_Y \) can be ignored [8]. The ratio estimate \( \hat{\beta} = \frac{\hat{\beta}_X}{\hat{\beta}_Y} \) reflects the causal effect between exposure and outcome, and is consistent asymptotically.

To test whether \( \beta \neq 0 \), it is essential to obtain the variance estimate of \( \hat{\beta} \). The commonly used first-order approximation is

\[
\text{Var}(\hat{\beta}) = \frac{\text{Var}(\hat{\beta}_X)}{\hat{\beta}_X^2}
\]

The first-order approximation method involves treating the denominator \( \hat{\beta}_X \) as a constant. However, because of the innate uncertainty in \( \hat{\beta}_0 \), we can expect that \( \frac{\text{Var}(\hat{\beta}_X)}{\hat{\beta}_X^2} \) tends to underestimate the true variance of \( \hat{\beta} \).
The second-order approximation method of variance of estimated causal effects

Thomas et al. [4] suggested a second-order approximation of the variance of $\hat{\beta}$. With the delta method, one can approximate the variance of causality $\hat{\beta}$ as follows.

$$\text{Var}(\hat{\beta}) \approx \frac{\text{Var}(\hat{\beta}_x)}{\hat{\beta}_x^2} - 2 \frac{\hat{\beta}_x}{\hat{\beta}_x^2} \text{Cov}(\hat{\beta}_x, \hat{\beta}_y) + \frac{\hat{\beta}_y}{\hat{\beta}_y^2} \text{Var}(\hat{\beta}_y)$$

Since we use different samples (2SMR), we can set $\text{Cov}(\hat{\beta}_y/\hat{\beta}_x) = 0$, as X and Y are from non-overlapping samples. Therefore, we obtain the following approximation.

$$\text{Var}(\hat{\beta}) \approx \frac{\text{Var}(\hat{\beta}_x)}{\hat{\beta}_x^2} + \frac{\hat{\beta}_y}{\hat{\beta}_y^2} \text{Var}(\hat{\beta}_y)$$

The second term is always positive. Therefore, if researchers use only the first term from this approximation for the variance, this can lead to an underestimation of the standard error.

Simulation design

We designed simulations to evaluate the magnitude of error in the first-order approximation method. We assumed specific true values for $\hat{\beta}$ and $\beta_x$, which also gave us the true value of $\beta_y = \beta_x \hat{\beta}_x$. We assumed the intercepts $\beta_{0x} = 0.03$ and $\beta_{0y} = 0.03$, and the errors $\sqrt{\text{Var}(\epsilon_x)} = \text{sd}(\epsilon_x) = 0.3$ and $\sqrt{\text{Var}(\epsilon_y)} = \text{sd}(\epsilon_y) = 0.3$. We independently generated genotypes (SNP alleles) $G_x$ and $G_y$, which are composed of 0, 1, and 2 from the distribution Binomial(2, MAF), where MAF denotes the minor allele frequency. We generated $(X\mid G_x, Y\mid G_y)$ by adding noise with mean 0 and variance $(\text{Var}(\epsilon_x), \text{Var}(\epsilon_y))$ to $(\hat{\beta}_x \hat{\beta}_y + \beta_x G_x + \beta_y G_y, Y_i)$, via simple linear regression. We can expect $\hat{\beta}_x$ and $\hat{\beta}_y$ to be randomly distributed by

$$\hat{\beta}_x \sim N(\beta_x, \frac{\text{Var}(\epsilon_x)}{\text{SS}_G})$$

$$\hat{\beta}_y \sim N(\beta_y, \frac{\text{Var}(\epsilon_y)}{\text{SS}_G})$$

$$\text{SS}_G = \Sigma G_x^2 \frac{1}{N_x} \text{ and } \text{SS}_G = \Sigma G y^2 \frac{1}{N_y}$$

where $N_x$ is the size of the reference dataset used in 2SMR and $N_y$ is the size of the target sample.

To approximate $\text{Var}(\hat{\beta})$, we can use either the first-order or the second-order approximation:

First-order: $\text{Var}(\hat{\beta}) \approx \frac{\text{Var}(\hat{\beta}_x)}{\hat{\beta}_x^2}$

Second-order: $\text{Var}(\hat{\beta}) \approx \frac{\text{Var}(\hat{\beta}_x)}{\hat{\beta}_x^2} + \frac{\hat{\beta}_y}{\hat{\beta}_y^2} \text{Var}(\hat{\beta}_y)$

Our simulation allowed us to empirically obtain a very accurate estimate of $\text{Var}(\hat{\beta})$ by repeating the simulation many times (we set the number of simulations as 100,000 in our study) with the same assumptions and calculating the observed variance of $\hat{\beta}$. This allowed us to compare the first and second-order approximations to the empirically obtained values.

We provide the R script code to run the entire simulation pipeline as Supplementary Data.

Results

We performed empirical simulations to compare the two types of analytical approximations: the classical way, in which only the first-order term is used, and the recently suggested way [4], which includes up to the second-order term. We also obtained an accurate estimate of the variance by empirically repeating simulations 100,000 times. Assuming that the empirically obtained variance is the gold standard, we calculated the ratio of the estimated variance to the gold standard.

In our simulations, we varied multiple parameters. We varied the N-ratio ($N_r/N_s$), we also varied $\hat{\beta}$ (the magnitude of causal effect) and MAF. Fig. 2 shows that the analytical approximation that contained variance up to the second-order term was almost as accurate as the empirical estimate, whereas the first-order approximation method was often largely inaccurate depending on the situation.

Fig. 2A shows that the error due to the first-order approximation decreased as the number of individuals ($N_x$) decreased from 200,000 to 2,000 (as the N-ratio increased from 1 to 100). The ratio was 0.84 when $N_x$ was 100,000, which is equal to $N_x/2$ (N-ratio = 2). The ratio rose to 0.99 when $N_x$ was 2,000 (N-ratio = 100). The mean of the ratios was 0.98, which translates to a reduced $\text{SE}(\hat{\beta})$ by $\sqrt{0.98} = 0.99$ times in the first-order approximation. Fig. 2B shows that the errors increased when the actual causal effect ($\beta$) between the exposure and outcome increased from 0.01 to 1. Therefore, if there is not a strong causal effect between the exposure and outcome in MR, the error from the first-order approximation would be small. The mean of the ratios of the first-order approximation was 0.93. Fig. 2C shows that, interestingly, the ratio appeared to be independent of the MAF of the variant. The mean of the ratios in this simulation was 0.93 in the first-order case.

We then analyzed the impact of the underestimated variance. If the variance is underestimated, the FPR can increase. We assumed the null hypothesis of no causal effect and generated 100,000 samples under an environment equivalent to that of Fig. 2A. We calculated the FPR based on the significance threshold of $\alpha = 0.05$. Fig.
Fig. 2. The ratio of the approximation of the variance of causal effect estimate to the true value. (A) We varied the N-ratio (N_y/N_x) value from 1 to 100 assuming N_x = 200,000, β_x = 0.02, β = 0.6, and minor allele frequency (MAF) = 0.2. (B) We varied the value β, i.e. the ratio of β_x and β_y from 0.01 to 1 assuming β_x = 0.02, N_x = 200,000, N_y = 10,000 and MAF = 0.2. (C) We varied the MAF from 0.02 to 0.5 assuming β = 1, β_x = 0.02, N_x = 200,000 and N_y = 10,000. The true value was estimated by empirical simulations (N_sim = 100,000).

Fig. 3. The scatter plots of false-positive rate (FPR) and statistical power. (A) The scatter plot of relationship between the N-ratio in the simulation of Fig. 2A and FPR. If N-ratio = 2 (N_y = 100,000 and N_x = 200,000), the FPRs were 0.071 (the dark red colored large dot) for the first-order approximation and 0.049 (the dark blue colored large dot) for the second-order approximation. To calculate FPR values, we generated 100,000 samples from the null hypothesis of no causal effect. (B) The scatter plot of relationship between the N-ratio the power. We generated 100,000 samples with β = 0.6 which is the causal effect of the exposure on the outcome.

3A shows the relationship between the N-ratio and the FPR. Notably, when the variance was underestimated by a factor of 0.84, as shown in Fig. 2A (for the case of an N-ratio = 2—that is, N_y = 100,000 and N_x = 200,000), the FPR of the first-order approximation method increased to 0.071 (the dark red colored large dot in Fig. 3A), while the FPR of the second-order approximation method was 0.049 (the dark blue colored large dot in Fig. 3A), corresponding to approximately 0.7 times that of the first-order case.
The average FPR in the second-order approximation method was 0.049, whereas the average FPR in the first-order approximation was 0.052. These findings indicate that the second-order approximation can be a good choice to prevent inflation of the FPR.

We also analyzed the statistical power (Fig. 3B). Since the variance of $\hat{\beta}$ is underestimated, the first-order approximation method may also tend to increase the power (or underestimate the false-negative rate). To compare the powers of the first and the second-order approximation methods, we generated 100,000 samples under an environment equivalent to that of Fig. 2A, with $\beta = 0.6$, which denotes the causal effect of the exposure on the outcome. Under this setting, the power of the first-order approximation was similar to that of the second-order approximation (on average 1.01 times greater).

Discussion

In this study, we performed simulations to evaluate the errors in the variance estimate of causal effects in 2SMR. We simulated a range of study parameters and showed that the commonly used first-order approximation can be inaccurate depending on the situation, while the second-order approximation is consistently accurate. We then showed that the underestimated variance can lead to a significant increase in the FPR.

In our simulations, the variance errors due to the first-order approximation were dependent on parameters such as the N-ratio and the $\beta$-ratio. When the number of samples in the target study increased while the number of samples in the external dataset for exposure association was fixed, the errors became larger. This suggested that in future studies, a larger study size may correspond to increased error from the first-order approximation method. Furthermore, as the true causal effect increased, so did errors. Interestingly, the errors appeared to be independent of the MAF.

In this study, we simply assumed the use of a single SNP as an IV in 2SMR. The causal effect between an exposure and an outcome is usually obtained by merging the ratio per variant ($\beta$) via the inverse-variance weighted method over a large number of variants. In this extended multi-variant model, we expect that the variance of the final estimate will also be affected by the errors induced by the first-order approximation, because the ratio for all variants is affected regardless of MAF. Then, the standard error of the causal effect, $\hat{\beta}$, would be dependent on the same parameters (N-ratio and the magnitude of beta) as in the extended model. Some other issues, such as linkage disequilibrium and pleiotropy, should also be addressed in the extended multivariate model.

Overall, our study suggests that the use of the second-order approximation is always preferable, since it provides an accurate estimate of the variance regardless of the situation. However, when the IV-exposure association is much greater than the IV-outcome association (i.e., $\beta$ is very small), we observed no significant difference between the first- and second-order approximations. Therefore, we expect that whether one must apply the second-order approximation to avoid an increased FPR will depend on many factors, including the actual range of $\beta$.

ORCID

Hakin Kim: https://orcid.org/0000-0003-2186-015X
Kunhee Kim: https://orcid.org/0000-0001-7899-7961
Buhm Han: https://orcid.org/0000-0002-2266-5164

Authors’ Contribution

Conceptualization: BH. Data curation: HK. Formal analysis: HK. Funding acquisition: BH. Methodology: BH, KK, HK. Writing - original draft: HK, KK, BH. Writing - review & editing: HK, BH.

Conflicts of Interest

Buhm Han is the CTO of Genealogy Inc.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) (grant number 2022R1A2B5B02001897) funded by the Korean government, Ministry of Science and ICT. This work was supported by the Creative-Pioneering Researchers Program funded by Seoul National University (SNU).

Supplementary Materials

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

References


Ghada Tagorti, Bülent Kaya*

Department of Biology, Faculty of Sciences, Akdeniz University, 07058 Campus, Antalya, Turkey

Human exposure to pollutants has been on the rise. Thus, researchers have been focused on understanding the effect of these compounds on human health, especially on the genetic information by using various tests, among them the somatic mutation and recombination tests (SMARTs). It is a sensitive and accurate method applicable to genotoxicity analysis. Here, a comprehensive bibliometric analysis of SMART assays in genotoxicity studies was performed to assess publication trends of this field. Data were extracted from the Web of Science database and analyzed by the bibliometric tools HistCite, Biblioshiny (RStudio), VOSViewer, and CiteSpace. Results have shown an increase in the last 10 years in terms of publication. A total of 392 records were published in 96 sources mainly from Brazil, Spain, and Turkey. Research collaboration networks between countries and authors were performed. Based on document co-citation, five large research clusters were identified and analyzed. The youngest research frontier emphasized on nanoparticles. With this study, how research trends evolve over years was demonstrated. Thus, international collaboration could be enhanced, and a promising field could be developed.

Keywords: bibliometric, cancer, Drosophila, genotoxicity, SMART, Web of Science

Introduction

Anthropogenic activities disseminate large amounts of chemical substances into the environment [1]. Hence, humans are currently exposed to several pollutants with genotoxic potentials such as metals [2], pesticides [3], industrial waste mix [4], and nanomaterials [5]. Genotoxicity is a wide term comprising DNA damage and mutagenicity, where the mutagenic effect is described as an occurred event with irreversible and heritable outcomes affecting the DNA and/or chromosome structure [6]. The genotoxic effect recorded on somatic cells has been associated with pathological endpoints such as premature aging, neuronal diseases, and even carcinogenesis [7,8]. Accordingly, the genotoxicity evaluation is a required component in the human health risk assessment and as one single test is unable to detect all the genotoxic endpoints, a battery of in vitro and in vivo tests was recommended [9]. Among these assays, the somatic mutation and recombination tests (SMARTs) are one of the commonly used tests. The SMART assays are in vivo assay to assess the potential genotoxicity of substances in the somatic cells of Drosophila melanogaster [10]. This assay could target wing cells named as wing-spot test proposed first time by Graf et al. [11], or eyes cells known as eye-spot test defined by Wurgler and Vogel [12]. In both cases, losing heterozygosity by deletions, point mutations, mitotic recombination, and nondisjunction unravels the expression of genetic markers in heterozy-
ous or transheterozygous individuals, ensuring the quantification of the damage by visual scoring [13]. The wing-spot test includes two types of cross using recessive genetic markers on the 3rd chromosome, a standard cross with normal bioactivation between female virgins (flr/s/In (3LR)TM3, ri p/+ se p I(3)89Aa bx<sup>34e</sup> e Bd) and (mwh/mwh) males, and a high metabolic bioactivation cross with high levels of cytochrome P450 between female virgins (ORR; flr/s/In (3LR)TM3, ri p/+ se p I(3)89Aa bx<sup>34e</sup> e Bd) and (mwh/mwh) males. The mutant spots are produced after chemical exposure that induced point mutation, deletion, or mitotic recombination [14,15]. The eye-spot test is based on a cross between wild-type eyed females (w+/w+) and white-eyed males (w/Y). The gene white (w) is a recessive marker found on the X chromosome. During the offspring period, a mutagenic event could occur and cause the formation of white phenotype spots (mutant ommatidia) in the wild-type eyes [16]. Although both tests are accurate, sensitive, and specific, the wing-spot test allows the visual scoring of wings over time, whereas in the eye-spot test, the analysis should be performed quickly since no preserving actions are available on the eyes [17]. In fact, the SMART assays have been applied in the analysis of the genotoxicity and the antimutagenicity of several chemicals and agents such as pesticides [18-20], nanomaterials [21,22], food products [23,24], hormones [25], plant extracts [26], and drugs [27-29]. With this variety of studies, a bibliometric analysis is required to assess the impact of this methodology on the genotoxicity studies. Bibliometric analysis is considered a highly sensitive method to evaluate research outputs based on statistical tools and to study metrological features of information created in a specific field [30]. To the best of our knowledge, no paper using the bibliometric analysis to explore the trends of SMART assays research has been published. Therefore, in the current study, various aspects were examined to evaluate the publications and citation trends in SMART assays from 1984 to 2020. Hence, the following research objectives were considered guiding the study design: (1) to identify the most influential journals and publications, the impactful authors and institutions, and the leading countries in SMARTs literature; (2) to find the patterns of collaboration between countries and authors within this research domain; (3) to explore the emerging keywords and research themes. The findings of the present study will provide a comprehensive overview of the importance of SMART assays, it would provide information to the scholars to easily identify the research profile and enhance collaboration.

**Methods**

This bibliometric study analyzed the published academic studies of SMARTs indexed in the Web of Science (WOS) core collection database. As known, WOS is the most reliable global citation database with a collection of over 21,000 peer-reviewed journals and the most accepted one for analysis of academic papers [31]. A comprehensive four-step approach was framed in this study as shown in Fig. 1. Boolean operators were used with adding all the relevant keywords to retrieve more relevant papers. (TS = (“SMART assays” OR “SMART assay” OR “SMART test” OR “SMART tests” OR “wing-spot test” OR “wing-spot assay” OR “somatic mutation and recombination test” OR “eye-spot test” OR “eye-spot assay” OR “small single spots” OR “large single spots” OR “mutant ommatidia” OR “mutant eye unit” OR “w/w+ SMART-assay”)). The search was performed on 14 January 2021; it is essential to present the date of records collection as the database is constantly updating [32]. Most bibliometric studies are based on academic articles [30], thereby this study was limited to original articles written in English in the strict sense. The search was also limited to 2020. The authors adopted the PRISMA approach, which has been used in bibliometric studies. A total of 460 records were extracted initially, which later were filtered by document types, to exclude the following document types: proceedings papers, book chapters, and early access. The data relevance and accuracy were assured by scanning the title

![Fig. 1. PRISMA flow diagram.](https://doi.org/10.5808/gi.21083)
and abstract of each record, and 49 records were excluded. The irrelevant articles were from the following categories: (1) Internet of things and wireless sensing, (2) communication services, (3) image processing, (4) industry, and (5) single molecule amplification and re-sequencing technology (SMART). Finally, a total of 392 records were remained to be downloaded in plain text to extract the following data: publication year, author, title, abstract, keywords, cited references, journal title, and institution, for further analysis.

Having selected 392 relevant records, data visualization and analysis were conducted by using various bibliometric software and applications, including Microsoft Excel, HistCite, Biblioshiny (RStudio), VOSviewer, and CiteSpace. The analysis was performed in three stages. First, a bibliometric citation analysis was performed (number of publications and citations, relevant journals, productive institutions, and the most impactful articles and authors). Secondly, a network analysis was applied, including collaboration between countries and co-authoring using the Walktrap clustering algorithm. This algorithm has the advantage to be computed effectively and placing the data in a network [33]. Furthermore, a three-fields-plot based on the Sankey diagram to indicate the interrelation between keywords, journals, and countries was visualized. Detection of the emerging research fields is required to outline research areas. Therefore, thirdly, a content analysis was conducted. HistCite (version 12.03.17) was used for the bibliometric citation analysis to sort the collected data by quantitative (number of publications and citations) and qualitative indicators (total global citation score and total local global citation score). Hence, these two types of indicators were applied in the current study. The total global citation score (TGCS) represents the number of citations of a paper included in the collection selected for the analysis in the WOS whereas the total local citation score (TLCS) refers to the total number of citations of a paper included in the collection and has been cited by other papers of the same collection [34]. The RStudio software (version 1.3.1093) with Biblioshiny application was used [35]. Hence, this application was also used to analyze the basic indicators of the search, plus the collaboration between countries and authors, and the three-fields plot. VOSviewer software (version 1.6.16) was used to analyze and visualize the emerging keywords. CiteSpace (version 5.7.R4) was used to perform the co-citation analysis and to investigate the emerging topics. Co-citation analysis is an effective tool to understand the intellectual structure of a research field, with the cited papers intellectual bases are revealed whereas papers in their active state of citation represent the research frontiers which usually display characteristics of a field specificity [36]. One of the important tools found in CiteSpace is the betweenness centrality. A paper with high betweenness centrality was defined as an important research paper in the network [37]. Furthermore, articles with strong citation bursts with time slices were identified. The burst is observed when a publication has an excess in its citation counts compared to its peers. This furthers identify publications that interested the scholars over time and thereby help to explore the research frontiers of a given field [38,39].

Results

The present study analyzed the SMART assays in the genotoxicity studies published during 1984—2020. A total of 392 records have been written by 933 authors from 35 different countries with an average of 17.01 citations per document. Authors of single-authored documents are seven (0.75%) while authors of multi-authored documents represent 926 authors (99.25%).

Basic indicators

Yearly publication and citation

The first publication appeared in 1984, thereafter the number of publications has been gradually increasing at a rate of 14.4 articles per year with a registered bloom in 2013 and 2015 (Recs, 22), whereas the TLCS and TGCS have recorded the bloom in 1984 (TLCS, 319; TGCS, 527) and in 1992 (TLCS, 282; TGCS, 547). The output of the first 25 years (1984–2009) was 206 publications and only in the past 10 years (2010–2020), 186 studies were published. This finding indicates the growing interest in using SMART assays in genotoxicity studies (Fig. 2).

Authors

The top 10 authors collectively contributed to 302 studies. Graf U, Marcos R, and de Andrade HHR are the top three authors in SMART assays field (Table 1). Graf U is the most influential author with the highest number of TLCS (953) and TGCS (1,716).

Institutions and countries

The top 10 institutions published 70.2% (n = 275) studies of the total publications. The Federal University of Uberlândia in Brazil is on the top of the list with 39 articles (TGCS, 425) (Table 2). Interestingly, the University of Zürich ranked in 4th position has the highest TGCS (1,892) and TLCS (1,119). There is only one country having a number of publications in three-digits, Brazil, with 106 publications and 1,103 TGCS, followed by Spain (TGCS, 1,470; TLCS, 338) and Turkey (TGCS, 784; TLCS, 221).

Journals

The 392 studies in SMART assays were published in 96 academic...
Fig. 2. Evolution of the number of articles and citations over the years. Recs, number of publications; TGCS, total global citation score; TLCS, total local citation score.

Table 1. The top 10 impactful authors ranked by Recs

<table>
<thead>
<tr>
<th>Author</th>
<th>Recs</th>
<th>TLCS</th>
<th>TLCS/t</th>
<th>TGCS</th>
<th>TGCS/t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graf U</td>
<td>42</td>
<td>953</td>
<td>33.47</td>
<td>1,716</td>
<td>65.59</td>
</tr>
<tr>
<td>Marcos R</td>
<td>39</td>
<td>235</td>
<td>21.23</td>
<td>775</td>
<td>79.41</td>
</tr>
<tr>
<td>de Andrade HHR</td>
<td>34</td>
<td>176</td>
<td>11.25</td>
<td>494</td>
<td>35.34</td>
</tr>
<tr>
<td>Lehmann M</td>
<td>33</td>
<td>113</td>
<td>8.93</td>
<td>327</td>
<td>29.67</td>
</tr>
<tr>
<td>Kaya B</td>
<td>30</td>
<td>163</td>
<td>14.85</td>
<td>449</td>
<td>41.8</td>
</tr>
<tr>
<td>Dihl RR</td>
<td>26</td>
<td>72</td>
<td>6.82</td>
<td>198</td>
<td>21.72</td>
</tr>
<tr>
<td>Creus A</td>
<td>25</td>
<td>199</td>
<td>15.39</td>
<td>603</td>
<td>51.26</td>
</tr>
<tr>
<td>Reguly ML</td>
<td>25</td>
<td>172</td>
<td>10.43</td>
<td>419</td>
<td>25.68</td>
</tr>
<tr>
<td>Spano MA</td>
<td>25</td>
<td>105</td>
<td>9.57</td>
<td>301</td>
<td>33.59</td>
</tr>
<tr>
<td>Demir E</td>
<td>23</td>
<td>99</td>
<td>10.91</td>
<td>276</td>
<td>35.09</td>
</tr>
</tbody>
</table>

Recs, number of publications; TLCS, total local citation score; TLCS/t, total local citation score per year; TGCS, total global citation score; TGCS/t, total global citation score per year.

Table 2. The most productive institutions and countries ranked by Recs

<table>
<thead>
<tr>
<th>No.</th>
<th>Country</th>
<th>Institution</th>
<th>Recs</th>
<th>TLCS</th>
<th>TGCS</th>
<th>No.</th>
<th>Country</th>
<th>Recs</th>
<th>TLCS</th>
<th>TGCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Brazil</td>
<td>Univ Fed Uberlândia</td>
<td>39</td>
<td>156</td>
<td>425</td>
<td>1</td>
<td>Brazil</td>
<td>106</td>
<td>361</td>
<td>1,103</td>
</tr>
<tr>
<td>2</td>
<td>Spain</td>
<td>Univ Autonoma Barcelona</td>
<td>38</td>
<td>227</td>
<td>763</td>
<td>2</td>
<td>Spain</td>
<td>69</td>
<td>338</td>
<td>1,470</td>
</tr>
<tr>
<td>3</td>
<td>Turkey</td>
<td>Akdeniz Univ</td>
<td>34</td>
<td>170</td>
<td>511</td>
<td>3</td>
<td>Turkey</td>
<td>66</td>
<td>221</td>
<td>784</td>
</tr>
<tr>
<td>4</td>
<td>Switzerland</td>
<td>Univ Zürich</td>
<td>30</td>
<td>1,119</td>
<td>1,892</td>
<td>4</td>
<td>Switzerland</td>
<td>41</td>
<td>446</td>
<td>1,016</td>
</tr>
<tr>
<td>5</td>
<td>Switzerland</td>
<td>Swiss Fed Inst Technol</td>
<td>27</td>
<td>347</td>
<td>770</td>
<td>5</td>
<td>Mexico</td>
<td>39</td>
<td>95</td>
<td>457</td>
</tr>
<tr>
<td>6</td>
<td>Brazil</td>
<td>Univ Fed Goias</td>
<td>24</td>
<td>47</td>
<td>260</td>
<td>6</td>
<td>Japan</td>
<td>30</td>
<td>72</td>
<td>405</td>
</tr>
<tr>
<td>7</td>
<td>Brazil</td>
<td>Univ Luterana Brasil</td>
<td>22</td>
<td>65</td>
<td>285</td>
<td>7</td>
<td>USA</td>
<td>25</td>
<td>489</td>
<td>923</td>
</tr>
<tr>
<td>8</td>
<td>Spain</td>
<td>Univ Cordoba</td>
<td>21</td>
<td>71</td>
<td>517</td>
<td>8</td>
<td>Greece</td>
<td>14</td>
<td>24</td>
<td>264</td>
</tr>
<tr>
<td>9</td>
<td>Brazil</td>
<td>Univ Fed Rio Grande do Sul</td>
<td>21</td>
<td>95</td>
<td>304</td>
<td>9</td>
<td>India</td>
<td>10</td>
<td>23</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>Mexico</td>
<td>Univ Nacl Autonoma Mexico</td>
<td>19</td>
<td>27</td>
<td>178</td>
<td>10</td>
<td>UK</td>
<td>10</td>
<td>91</td>
<td>179</td>
</tr>
</tbody>
</table>

Recs, number of publications; TLCS, total local citation score; TGCS, total global citation score.
sources. More than half of these studies (n = 241, 61.5%) were published in the top 10 journals. The sources 'Food and Chemical Toxicology' and 'Mutation Research Genetic Toxicology and Environmental Mutagenesis' are on the top of the list with 53 publications (Table 3). The journal with the highest impact factor Chemosphere (7.086) has published 10 studies with 177 TGCS.

Articles
The years of the top 10 highly cited articles ranged from 1984 to 1996. There is only one article that obtained over 200 citations. This article has been titled 'Somatic Mutation and Recombination Test in Drosophila melanogaster' by Graf et al. and published in 1984 [11]. Half of the highly cited articles were published in Mutation Research journal, and this journal is on the top list of influential journals (Table 4).

Network Analysis
Co-authorship
Each node represents an author and the edges indicate the research collaboration between them. Six clusters are recorded, the blue and the orange clusters both with five authors are the largest clusters, followed by the purple cluster which includes four authors (Fig. 3). The brown cluster has one separate author.

Table 3. The most productive journals ranked by Recs

<table>
<thead>
<tr>
<th>Journal</th>
<th>Recs</th>
<th>TLCS</th>
<th>TLCS/t</th>
<th>TGCS</th>
<th>TGCS/t</th>
<th>IF (2020)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food and Chemical Toxicology</td>
<td>53</td>
<td>165</td>
<td>15.8</td>
<td>714</td>
<td>69.78</td>
<td>6.025 (Q1)</td>
</tr>
<tr>
<td>Mutation Research–Genetic Toxicology and Environmental Mutagenesis</td>
<td>53</td>
<td>250</td>
<td>17.55</td>
<td>1,085</td>
<td>81.73</td>
<td>2.873 (Q3)</td>
</tr>
<tr>
<td>Environmental and Molecular Mutagenesis</td>
<td>31</td>
<td>213</td>
<td>11.24</td>
<td>511</td>
<td>29.03</td>
<td>3.216 (Q2)</td>
</tr>
<tr>
<td>Mutation Research</td>
<td>30</td>
<td>610</td>
<td>20.28</td>
<td>1,128</td>
<td>37.78</td>
<td>2.11 (Q1)</td>
</tr>
<tr>
<td>Mutagenesis</td>
<td>20</td>
<td>187</td>
<td>9.44</td>
<td>440</td>
<td>18.44</td>
<td>3.000 (Q2)</td>
</tr>
<tr>
<td>Fresenius Environmental Bulletin</td>
<td>17</td>
<td>34</td>
<td>3.35</td>
<td>53</td>
<td>5.78</td>
<td>0.489 (Q4)</td>
</tr>
<tr>
<td>Chemosphere</td>
<td>10</td>
<td>30</td>
<td>3.12</td>
<td>177</td>
<td>19.19</td>
<td>7.086 (Q1)</td>
</tr>
<tr>
<td>Genetics and Molecular Biology</td>
<td>9</td>
<td>28</td>
<td>2.2</td>
<td>63</td>
<td>6.01</td>
<td>1.771 (Q3)</td>
</tr>
<tr>
<td>Mutation Research–Fundamental and Molecular Mechanisms of Mutagenesis</td>
<td>9</td>
<td>28</td>
<td>1.12</td>
<td>114</td>
<td>5.78</td>
<td>2.433 (Q3)</td>
</tr>
<tr>
<td>Toxicology and Industrial Health</td>
<td>9</td>
<td>17</td>
<td>2.85</td>
<td>68</td>
<td>11.73</td>
<td>2.273 (Q4)</td>
</tr>
</tbody>
</table>

Recs, number of publications; TLCS, total local citation score; TLCS/t, total local citation score per year; TGCS, total global citation score; TGCS/t, total global citation score per year; IF, impact factor; Q, quartile.

Table 4. The top 10 highly cited articles ranked by LCS

<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
<th>Source</th>
<th>Year</th>
<th>LCS</th>
<th>GCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal experimental-design and sample-size for the statistical evalu-</td>
<td>Frei and Wurgler [40]</td>
<td>Mutation Research–Environmental Mutagenesis and Related Subjects</td>
<td>1995</td>
<td>133</td>
<td>156</td>
</tr>
<tr>
<td>ation of data from somatic mutation and recombination tests (SMART) in Drosophila</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improved high bioactivation cross for the wing somatic mutation and recom-</td>
<td>Graf and van Schaik [41]</td>
<td>Mutation Research</td>
<td>1992</td>
<td>132</td>
<td>159</td>
</tr>
<tr>
<td>bination test in Drosophila melanogaster</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The genotoxicity of the antitumor drug mitoxantrone in somatic and germ-cells of Drosophila melanogaster</td>
<td>Frei et al. [42]</td>
<td>Mutation Research</td>
<td>1992</td>
<td>85</td>
<td>100</td>
</tr>
<tr>
<td>Metabolism of promutagens catalyzed by Drosophila melanogaster Cyp6a2 enzyme in Saccharomyces cerevisiae</td>
<td>Saner et al. [44]</td>
<td>Environmental and Molecular Mutagenesis</td>
<td>1996</td>
<td>41</td>
<td>91</td>
</tr>
</tbody>
</table>

LCS, local citation score; GCS, global citation score.
Collaboration between countries
A total of 38 collaboration entries are registered worldwide. Turkey, Brazil, and Mexico were the most collaborative countries (Fig. 4).

Co-word
The co-occurrence of keywords represents the relationship between two words that occurred together. Three minimum number of occurrences of a keyword were selected; hence, out of 861 authors’ keywords, 70 meet this criterion to form 11 clusters (Fig. 5). Each color indicates a separate cluster and clusters are organized based on the link strength and occurrence. Thus, the size of the bubble represents the relationship between link strength and occurrence. The first five keywords with the high total link strength are Drosophila melanogaster (link strength: 320), genotoxicity (247), SMART (181), Drosophila (108), and wing-spot test (105).

Three-fields plot
The interconnections among sources (left), countries (middle), and author keywords (right) are analyzed to understand which keywords are preferable to which countries and used to what sources. The three top countries (Brazil, Turkey, and Spain) have a strong connection with the source 'Food and Chemical Toxico-
gy’ and prefer to publish four keywords (*Drosophila* melanogaster, genotoxicity, SMART, and antigenotoxicity). The block length presented in Fig. 6 indicates the level of connection.

**Document co-citation and citation bursts**

A co-citation network was generated with 985 nodes and 3,291 links for a one-year time slice (Fig. 7). The cited references are representing in the form of nodes and the co-citation relationships are visualized in the form of links. The top five co-cited articles are shown in Table 5 [14,41,47-49]. From a total of 392 records and 16,734 references, a list of the top 25 references with the strongest citation bursts was generated (Fig. 8). The citation bursts in the list of the top 25 references have been expanded between 1989 and 2017. Most of the strength bursts ranged between 4 and 7, and most citations have 3 to 4 years expanded duration. However, the citation with the most expanded duration (2011–2016) has a low strength of its citation bursts (5.47) [50]. The reference paper written by Frei and Wurgler (1988) had the highest citation burst (16.12) [47].

**Thematic analysis**

A total of 17 clusters were identified in the SMART assays literature. The largest clusters are shown in Fig. 9. Based on specific metrics, term frequency–inverse document frequency, log-likeli-
hood tests (LLR), and mutual information tests, CiteSpace analyses the title of articles to extract a noun to characterize the cluster type. Generally, LLR covers the best themes (Table 6). A silhouette value (S) > 0.7 denotes the high credibility of a cluster and a value of modularity (Q) > 0.3 reveals the significant structure of the network [51]. As shown in Figs. 7 and 8, the top-ranked item by centrality is Frei and Wurgler (1995) [40] in Cluster #2, with the centrality of 41. The second one is Carmona et al. (2011) [50] in Cluster #3, with the centrality of 38. The third is Demir et al. (2011) [49] in Cluster #2, with the centrality of 34. Therefore, these papers are considered as pivotal points that allow connections between the research area. The clusters #0, #2, and #3 are the most active clusters with the strongest citation bursts (Table 7) [14,41,48-50,52-55]. This implies that these clusters denote where the supreme effort of research in the SMART assay.

Discussion

In the current study, a comprehensive analysis of the emerging trends in the field of SMART assays from 1984 to 2020 was performed. The analysis reveals an increase in the number of publications in this period, with most of these having been published in the last 10 years, confirming the growing interest in SMART assays. Of note, recently published studies have received fewer citations compared to ancient studies, as it requires time for a study to make an impact. The first observed bloom of TLCS and TGCS has been associated with the first paper introducing the SMART assays in 1984. Thus, the highly cited article is the earliest publication. The second bloom in 1992 has been generated from two influential articles written by Graf and van Schaik (1992) [41] and Frei et al. (1992) [42]. The most prolific author in the SMARTs field is Graf U (Switzerland), with 953 TLCS and 1,716 TGCS. Furthermore, four of his research papers are in the top 10 highly cited articles, and five of them in the top 25 strongest citation bursts showing the in-
Fig. 7. Co-citation network of SMART assays field. Each link colors indicate a given time slice. The oldest co-citation relationships are visualized as dark blue, whereas yellow links presented articles that are recently co-cited (created by CiteSpace). SMART, somatic mutation and recombination test.

Table 5. The top five critical articles in SMART assays

<table>
<thead>
<tr>
<th>Cited Frequency</th>
<th>Title</th>
<th>Author</th>
<th>Year</th>
<th>Betweenness Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>Statistical methods to decide whether mutagenicity test data from Drosophila assays indicate a positive, negative, or inconclusive result</td>
<td>Frei and Wurgler [47]</td>
<td>1988</td>
<td>0.08</td>
</tr>
<tr>
<td>21</td>
<td>30 Compounds tested in the Drosophila wing spot-test</td>
<td>Graf et al. [14]</td>
<td>1989</td>
<td>0.13</td>
</tr>
<tr>
<td>19</td>
<td>The genome of Drosophila melanogaster</td>
<td>Lindsley and Zimm [48]</td>
<td>1992</td>
<td>0.08</td>
</tr>
<tr>
<td>19</td>
<td>Genotoxic analysis of silver nanoparticles in Drosophila</td>
<td>Demir et al. [49]</td>
<td>2011</td>
<td>0.12</td>
</tr>
<tr>
<td>14</td>
<td>Improved high bioactivation cross for the wing somatic mutation and recombination test in Drosophila melanogaster</td>
<td>Graf and van Schaik [41]</td>
<td>1992</td>
<td>0.12</td>
</tr>
</tbody>
</table>

SMART, somatic mutation and recombination test.

Interest of researchers in his field. The collaboration between authors could be due to the emergence of interest among researchers. While only one of the prominent clusters is interconnected, it is expected in the future to improve overall collaborative work. The large number of multi-authored documents could be related to various collaborations between countries to expedite the usage of SMART assays. In fact, 35 countries have published about SMART assays. The two influential authors de Andrade HHR and Lehmann M are both from Brazil. Thus, this country had the highest number of publications. The rise in publications from Brazil can also be attributed to the high frequency of collaboration with institutions in Switzerland. Both authors de Andrade HHR and Lehman M were collaborators with Graf U in 2000 to study the genotoxic potential of tannic acid [56]. However, the usage of SMART assays in genotoxicity studies is still ignored in many countries. The most influential journals accounted for 61.5% of all the publication.
tions, and this finding illustrated that the distribution of publication was narrow. These influential journals are in Q1(3), Q2(2), Q3(3), and Q4(2) category. To note, five highly cited articles were published in Mutation Research (Q1), two in Mutagenesis (Q2), and one in Environmental and Molecular Mutagenesis (Q2). The first most-cited article by Graf et al. was published in 1984 [11]. This article was published in Environmental Mutagenesis (Q2), ranked at 55th position with only this publication. Therein, the protocol of SMART assay was presented, and several chemicals such as β-propiolactone, 1,2-dibromoethane, aflatoxin B1, diethyl-nitrosamine, dimethylnitrosamine, mitomycin C, and procarbazine have been identified as mutagens. The second most-cited article by Frei and Wurgler [40] was published in Mutation Research-Environmental Mutagenesis and Related Subjects. This journal ceased publication, to be incorporated in 1997 into Mutation Research-Genetic Toxicology and Environmental Mutagenesis journal [57]. In this study, to reduce the risk of inconclusive results, a new statistical test was proposed. The third most-cit-
ed article by Graf and van Schaik [41] proposed new strains to improve the visual score. This article was published in Mutation Research (Q1 in 1999) currently known as Mutation Research: Fundamental and Molecular Mechanisms of Mutagenesis (Q3) [57]. The most used keywords allow distinguishing the articles relevant to SMART assays. The three-fields plot showed that Drosophila melanogaster, genotoxicity, SMART, and antigenotoxicity were used by authors from Brazil, Turkey, and Spain. These keywords appear to be generic; yet it was used frequently. Based on thematic analysis, five large clusters were observed. The two clusters 'somatic mutation' and 'nitrogen mustard' are the two largest and oldest clusters. The cluster 'copper oxide nanoparticle' is the youngest cluster. The value of the mean silhouette (S) and the modularity (Q) are 0.9501 and 0.3944, respectively, suggesting reliable and robust results. In the largest cluster labeled ‘somatic mutation’ (#0) with 103 members and which is also the most active cluster, the most actively citing article (a research frontier article) identified polycyclic aromatic hydrocarbons and their derivatives as genotox [58]. The most actively cited articles (intellectual-based papers) are Graf et al. (1989) [14], Lindsley and Zimm (1992) [48], and Graf and van Schaik (1992) [41]. Graf et al. [14] is the second top-ranked paper with the strongest citation burst (11.22) and therein the efficiency of the SMART assay was proven by applying it on 30 compounds to evaluate their genotoxic potential. Lindsley and Zimm [48] described and identified the genome of Drosophila. Graf and van Schaik [41] improved new strain to assure better cross. In brief, cluster #0 mainly concentrated on enhancing the SMART protocol to be more practical and accurate. In the second cluster labeled ‘nitrogen mustard’ (#1), the most actively citing paper focused on the effect of tannic acid on nitrogen mustard, mitomycin C, and methylmehanesulfonate [56], whereas the most actively cited paper are Frei and Wurgler [40] and Graf [59]. These two papers focused on the Drosophila model by identifying the sample size required as well as the appropriate age of larvae, depending on the mutagens to avoid inconclusive results. It can be concluded that cluster #1 focused on the incorporation of a new control positive (nitrogen mustard) into SMART assay. Not surprisingly that the two largest clusters cover the most interests as both focused on the improvement of SMART assays protocol. In the third cluster labeled ‘vivo model’ (#2) the most actively citing paper was written by Carmona et al. [50] to evaluate the genotoxic potential of titanium dioxide anatase nanoparticles. Demir et al. [49] and Vales et al. [52] are the most-cited paper. Demir et al. [49] analyzed the genotoxic potential of silver nanoparticles

Table 6. The top five clusters in the literature

<table>
<thead>
<tr>
<th>ID</th>
<th>Size</th>
<th>Silhouette</th>
<th>Label (TF-IDF)</th>
<th>Label (LLR)</th>
<th>Label (MI)</th>
<th>Cited year</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>103</td>
<td>0.910</td>
<td>Drosophila melanogaster</td>
<td>Somatic mutation</td>
<td>Inhibitory activity</td>
<td>1989</td>
</tr>
<tr>
<td>1</td>
<td>97</td>
<td>0.864</td>
<td>Drosophila melanogaster</td>
<td>Nitrogen mustard</td>
<td>Pyrrolizidine alkaloid</td>
<td>1995</td>
</tr>
<tr>
<td>2</td>
<td>79</td>
<td>0.932</td>
<td>Drosophila melanogaster</td>
<td>Vivo model</td>
<td>Grifola gargal singer</td>
<td>2010</td>
</tr>
<tr>
<td>3</td>
<td>77</td>
<td>0.961</td>
<td>Drosophila melanogaster</td>
<td>Doxorubicin-induced somatic mutation</td>
<td>Sage tea</td>
<td>2005</td>
</tr>
<tr>
<td>4</td>
<td>74</td>
<td>0.977</td>
<td>Drosophila melanogaster</td>
<td>Copper oxide nanoparticle</td>
<td>Grifola gargal singer</td>
<td>2015</td>
</tr>
</tbody>
</table>

TF-IDF, term frequency–inverse document frequency; LLR, log-likelihood tests; MI, mutual information.

Table 7. The top three articles in clusters #0, #2, and #3 with the strongest citation bursts

<table>
<thead>
<tr>
<th>ID</th>
<th>Burst</th>
<th>Title</th>
<th>Author</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11.22</td>
<td>30 Compounds tested in the Drosophila wing spot-test</td>
<td>Graf et al. [14]</td>
<td>1989</td>
</tr>
<tr>
<td>9.31</td>
<td>Genotoxicity of cohabitant nanoparticles and ions in Drosophila</td>
<td>Demir et al. [49]</td>
<td>2011</td>
<td></td>
</tr>
<tr>
<td>7.21</td>
<td>Genotoxicity of cohabitant nanoparticles and ions in Drosophila</td>
<td>Vales et al. [52]</td>
<td>2013</td>
<td></td>
</tr>
<tr>
<td>5.73</td>
<td>Role of homologous recombination in carcinogenesis</td>
<td>Bishop and Schiestl [54]</td>
<td>2003</td>
<td></td>
</tr>
<tr>
<td>5.70</td>
<td>Protective effects of a mixture of antioxidant vitamins and minerals on the genotoxicity of doxorubicin in somatic cells of Drosophila melanogaster</td>
<td>Costa and Nepomuceno [55]</td>
<td>2006</td>
<td></td>
</tr>
</tbody>
</table>
whereas Vales et al. [52] evaluated the genotoxic effect of cobalt nanoparticles. It seems that in cluster #2 nanoparticles have gradually attracted the attention of scholars. Although the most-cited and citing papers were mainly dedicated to the genotoxic effect of nanoparticles, most of members treated various compounds, justifying thereby the label ‘vivo model’ instead of nanoparticles. In the fourth cluster labeled ‘doxorubicin-induced somatic mutation’ (#3), all the intellectual bases are related to cancer. Fragiorge et al. [53] analyzed the antigenotoxicity effect of ascorbic acid on doxorubicin. Doxorubicin is an antibiotic to treat human cancers which generally induces genotoxicity by oxidative damage [61]. Bishop and Schiestl [54] studied the function of homologous recombination in cancer. Similarly, the citing paper focused on the use of the herbal extract of ginseng to inhibit the genotoxic effect of doxorubicin [62]. In brief, cluster #3 was concentrated on resolving the doxorubicin-induced genotoxicity. The fifth cluster labeled ‘copper oxide nanoparticle’ (#4) is considered a persistent cluster denoting the continuity of an existing trend. In this cluster, the research frontier assessed the effect of copper oxide nanoparticles [63]. The metal oxide nanoparticles possess a redox property suggesting an antigenotoxic and anticarcinogenic potential [21]. The intellectual base emphasis on nanomaterials. Alaraby et al. [64] discussed the side effect of nanomaterials. Carmona et al. [60] presented the titanium dioxide anatase nanoparticles, being generally used in pharmaceuticals and cosmetics. This nanomaterial can generate oxidative stress, and subsequently genotoxicity [65]. Interestingly, this paper was a research frontier in cluster #2. It is worth mentioning, the first top-ranked burst item Frei and Wurgler (16.12) [47] is in a small and old cluster (1986) labeled ‘antiparasitic nitrofuran’ (#6) with 45 members. In sum, the research fronts share sometimes the same theme with the intellectual bases since most often the continuation and the growth of intellectual base are the research fronts. The present study has certain limitations. Data were extracted from the WOS database; other databases were not considered. The database WOS is always updating, even with recently published papers, top publications are high enough that including recent paper would not have an influence on this research. The proceedings papers, book chapters, and early access are not included. Even though book chapters were excluded, the book by Wurgler and Vogel (1986) [12] describing the eye-spot assay was not indexed in the WOS database. The researchers tried carefully to include a maximum of relevant keywords; however, two studies Martínez-Valdivieso et al. (2017) [66] and Fernández-Bedmar and Alonso-Moraga (2016) [67] have been missed. The absence of these papers is due to keyword search where none of the relevant used keywords were present in the title, neither in the abstract and/or the keywords section. The number of authors may differ since some authors published articles with a different initial of the first name. This led to some authors being separated into two authors. Even with these limitations, the current study on SMART assays can provide a directive to researchers to find influential articles, journals, and authors to assure collaboration and to reduce the research gaps.

In this study, a comprehensive bibliometric analysis was performed on SMART assays literature. Most of the publications were published in the last 10 years, mainly from Brazil, especially from the Federal University of Uberlândia. Notably, multiple authors were publishing papers on this field and mostly in ‘Food and Chemical Toxicology’ and ‘Mutation Research Genetic Toxicology’ and Environmental Mutagenesis’ journals. By analyzing the emerging trends, the nanoparticle area represents a persistent cluster. This area is expected to draw more attention. Finally, further studies should assess the records present on other databases to find out if the same trends for SMART assays are present.

**ORCID**

Ghada Tagorti: https://orcid.org/0000-0003-4597-8320

Bülent Kaya: https://orcid.org/0000-0002-0491-9781

**Authors’ Contribution**

Conceptualization: GT. Data curation: GT, BK. Formal analysis: GT. Methodology: GT, BK. Writing - original draft: GT. Writing - review & editing: BK.

**Conflicts of Interest**

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

**References**


https://doi.org/10.5808/gi.21083


47. Frei H, Wurgler FE. Statistical methods to decide whether mutagenicity test data from Drosophila assays indicate a positive, negative, or inconclusive result. Mutat Res 1988;203:297-308.


Designing a novel mRNA vaccine against *Vibrio harveyi* infection in fish: an immunoinformatics approach

Sk Injamamul Islam\(^1\)^, Moslema Jahan Mou\(^3\), Saloa Sanjida\(^4\), Muhammad Tariq\(^5\), Saad Nasir\(^6\), Sarower Mahfuj\(^1\)

\(^1\)Department of Fisheries and Marine Bioscience, Faculty of Biological Science, Jashore University of Science and Technology, Jashore 7408, Bangladesh
\(^2\)Chulalongkorn University, Department of Veterinary Microbiology, Faculty of Veterinary Science and Technology, Bangkok 10330, Thailand
\(^3\)Department of Genetic Engineering & Biotechnology, Faculty of Earth and Life Science, University of Rajshahi, Rajshahi 6205, Bangladesh
\(^4\)Department of Environmental Science and Technology, Faculty of Applied Science and Technology, Jashore University of Science and Technology, Jashore 7408, Bangladesh
\(^5\)Department of Biotechnology, Faculty of Biological Sciences, University of Malakand, Chakdara 18800, Pakistan
\(^6\)Department of Clinical Medicine and Surgery, Faculty of Veterinary Medicine, University of Veterinary and Animal Sciences, Lahore 54000, Pakistan

*Corresponding author:*
E-mail: 6378506331@student.chula.ac.th

\textit{Vibrio harveyi} belongs to the family Vibrionaceae of class Gammaproteobacteria. Around 12 Vibrio species can cause gastroenteritis (gastrointestinal illness) in humans. A large number of bacterial particles can be found in the infected cells, which may cause death. Despite these devastating complications, there is still no cure or vaccine for the bacteria. As a result, we used an immunoinformatics approach to develop a multi-epitope vaccine against the most pathogenic hemolysin gene of *V. harveyi*. The immunodominant T- and B-cell epitopes were identified using the hemolysin protein. We developed a vaccine employing three possible epitopes: cytotoxic T-lymphocytes, helper T-lymphocytes, and linear B-lymphocyte epitopes, after thorough testing. The vaccine was developed to be antigenic, immunogenic, and non-allergenic, as well as have a better solubility. Molecular dynamics simulation revealed significant structural stiffness and binding stability. In addition, the immunological simulation generated by computers revealed that the vaccination might elicit immune reactions *Escherichia coli* K12 as a model, codon optimization yielded ideal GC content and a higher codon adaptation index value, which was then included in the cloning vector pET2\(^{+}\) (a). Altogether, our experiment implies that the proposed peptide vaccine might be a good option for vibriosis prophylaxis.

**Keywords:** immune simulation, molecular dynamics simulation, T-cell epitopes, vaccine design, *Vibrio harveyi*

---

**Introduction**

The disease, called vibriosis, affects both farmed and marine fish across the globe. The pathogenicity of *Vibrio* species and their antibiotic resistance is, however, poorly understood. The virulence components of *Vibrio* spp. that have been associated to animal and human illnesses are often not accessible or common in the environment [1]. Because
Vibrio has a highly plastic genome, there is a substantial likelihood that pathogenic and ambient Vibrio will share genes for virulence. As a result, there has been an increase in the number of pathogenic Vibrio strains in the aquatic environment [2]. Vibrio harveyi is one of the most serious infections affecting farmed fishes, and Vibrio alginolyticus, Vibrio parahaemolyticus, and Vibrio campbellii have also been found in numerous tropical nations [3-6]. Among temperate waters of Asia, southern Europe, and South America, this pathogen occurs naturally in marine habitats and has become an important pathogen of wild and cultured fish and invertebrates. The symptoms of V. harveyi infection include anemia, necrosis of the intestine, ascitic fluid, petechial hemorrhages, tail erosion, infection of the eye, mucous secretions, and frequent mortality in fish [7]. Moreover, resistance to bactericidal processes is one of the most important aspects in the pathogenicity of fish infections. Overuse of antibiotics in human medicine, agriculture, and aquaculture systems has resulted in the emergence and evolution of antimicrobial resistance in Vibrio spp. over the last few decades [8]. V. harveyi is highly pathogenic to salmonids, sea bass, and tilapia, and generates an extracellular product with a high titer of hemolytic activity against fish erythrocytes. V. harveyi has a single chromosome with a length of 6,374,398 base pairs [9]. In spite of several genes associated with V. harveyi causing disease in fish, hemolysin is well-known as a virulence factor linked to both fish and human diseases [10]. The hemolysin protein is a crucial protein that allows viruses to penetrate the host cell wall, making it a suitable target antigen for vaccine development [11]. Cytotoxic T lymphocyte (CTL) epitopes [12] and CD4+ T cell epitopes were found to diverge in fish species [13] by epitopes mapping using fish with experimentally infected disease and a library of overlapping peptides of viruses [14]. Immunizations are intended to elicit an immune response to a potentially lethal foreign pathogen and to prepare the body to infiltrate those particles, limit toxicity, or initiate assassination activities against the bacteria. A vaccination, according to prior research, can prevent future outbreaks of bacteria-associated natural microorganisms such as bacteria [15]. The prompt discovery of safe, efficient, uncomplicated, economical, dependable, and fast production of antibody against the guided antigen is made possible by in-silico design of multi-epitope vaccines against pathogens. Epitope-based vaccines have been successfully created in the postgenomic period to stimulate responsiveness against some of the worst human viruses, including influenza, nipah, chikunguniya, zika, ebola, Middle East respiratory syndrome coronavirus, rota, and others [16-20]. Previously, the in-silico technique in fish had not been developed due to a lack of understanding of the differences between major histocompatibility complexes (MHC class I and II) and human leukocyte antigen (HLA) [21,22] but recent research on fish species has generated data to enable in-silico techniques [23-25]. Both MHC class I and class II molecules were found in the experimental data of cord and tilapia for starting immune responses against infections. As a result, the peptide with excellent binding capacities to HLA-A*0201, HLA-B*3501, and HLA-B*3508 might be employed as efficient vaccinations against certain fish diseases [21,26]. Lately, an in-silico technique was effective in predicting epitopes and multiepitopes with significant responsiveness against Streptococcus agalactiae, Edwardsiella tarda, and Flavobacterium columnare, three harmful bacteria that induce streptococcosis, edwardsiellosis, and columnaris in fish, separately [27-29]. Experts expect that in the coming days, computer-assisted techniques will be increasingly successful in controlling fish diseases [30,31]. As a result, the main objective of this research was to identify multi-epitope from the best antigenic protein to fight against V. harveyi infection.

Methods

Architectural flow chart is being given in Fig. 1.

Retrieval of proteome and antigen selection
We used the NCBI (https://www.ncbi.nlm.nih.gov/) database to find accessible Vibrio harveyi proteomes for antigen selection. Hemolysin is a crucial protein that allows bacteria to penetrate the host cell wall, making it a suitable target antigen for V. harveyi vaccine development [32,33]. The hemolysin is a vibriosis component in fish that is thought to be responsible for causing mortality in fish [34]. We examined the hemolysin protein of the V. harveyi for multi-epitope vaccine design because of its direct role in pathogenesis. After the hemolysin was isolated, the chosen amino acid sequences of the bacteria were obtained as FASTA files (GenBank: ACF32997.1). VaxiJen v2.0 (http://www.ddg-pharmfac.net/ vaxi- jen/) server was used to assess the protective antigens of hemolysin [35] and for each of them, a threshold value of 0.4 was chosen on the ANTIGENpro (http://scratch.proteomics.ics.uci.edu/) server [36]. Subsequently, the hemolysin with the highest antigenic score was chosen for further research.

Prediction and assessment of cytotoxic T-lymphocyte epitope
CTLs are basic kinds of immune responsive cells that have the ability to directly destroy other infectious cells [37]. They immediately enter the infected cell and contribute to the host’s defensive
response. The sequence of the chosen protein was entered into a server named NetCTL v1.2 (http://www.cbs.dtu.dk/services/NetCTL/) to predict CTLs epitope [38]. It integrates information about proteasomal C terminal cleavage affinity (C-score), TAP transport efficiency, and MHC class I affinity to deliver its output for a given protein. The threshold parameter for prediction was set to 0.4 to obtain 0.89 sensitivity and 0.940 specificity. VaxiJen v2.0 was used to further evaluate the predicted epitopes [35], followed by MHC class I immunogenicity (http://tools.iedb.org/immunogenicity/) [39], ToxinPred (http://crdd.osdd.net/raghava/toxinpred/) [40], and AllerTop v2.0 (https://ddg-pharmfac.net/OP/) [41] online servers. All of the forecasts were made using the default parameters of each server.

**Epitopes of helper T-lymphocytes prediction and assessment**

In response to external antigens, helper T-lymphocytes (HTLs) detect and activate B lymphocyte and CTL and causing the infectious pathogen to be destroyed [42]. The HTL epitopes were defined using the IEDB’s MHC class II binding allele prediction tool, which can be found at http://tools.iedb.org/mhcii/. With a percentile rank of 5%, the HTL epitopes were chosen using the CONSENSUS technique [43]. The IEDB currently recommends making selections based on a percentile rank of ≤ 1% for each (MHC allele, length) combination to cover most of the immune responses. Alternatively, a binding affinity (IC50) threshold of 500 nM identifies peptide binders recognized by T cells and this threshold can be used to select peptides [44]. This tool employs different methods to predict MHC class II epitopes, including a consensus approach which combines NN-align, SMM-align, and combinatorial library methods. The antigenicity and cytokine-inducing properties of the anticipated epitopes, namely interferon-γ (IFN-γ), interleukin-4 (IL-4), and interleukin-10 (IL-10), were assessed further. Antigenicity was predicted using the VaxiJen v2.0 server, while IFN-γ, IL-4, and IL-10 features were anticipated employing IFNepitope (http://crdd.osdd.net/raghava/ifnepitope/) [45], IL-4pred (http://crdd.osdd.net/raghava/il4pred/) [45], and IL-10pred (http://crdd.osdd.net/raghava/IL-10pred/) [46] servers, respectively, with default parameters.

**Prediction and assessment of linear B-lymphocyte epitopes**

To promote humoral or antibody-mediated immunity, B-cell epi-
topes are required. B-cells are made up of amino acid groups that bind with secreted antibodies and stimulate the immune system to fight infections [47]. As a consequence, we utilized the ibCE-EL server (http://www.thegleelab.org/ibCE-EL/) to identify the linear B-lymphocyte (LBL) epitopes using default settings [48]. It is an ensemble method that combined extremely randomized tree and gradient boosting algorithms, which respectively utilizes a combination of amino acid composition and physicochemical properties and a combination of dipeptide and physicochemical properties as an input feature. For a given peptide, ibCE-EL predicts its calss and probability values [48]. This server also can give 12-25 mer sequence as output. The Vaxijen v2.0, ToxinPred, and AllerTop v2.0 servers were used to test the anticipated LBL epitopes.

**Peptide modeling and molecular docking**

PEP-FOLD v3.0 (https://bioserv.rpbs.univ-parisdidier.fr/services/PEP-FOLD3/) server was used to simulate the chosen CTL and HTL epitopes. For the procedure, the sOPEP sorting scheme with 200 simulations was employed [49]. HLA-B*3508, HLA-A*0201, and HLA-B*3501 were chosen for selected CTL epitopes, whereas DRB1*07:01, DRB1*04:01, and DRB1*11:01 were chosen for HTL epitopes, based on epitope-wise HLA binding allele analysis. The HLA protein crystal structures were obtained from the Protein Data Bank (PDB) (https://www.rcsb.org/) [50] followed by processing with BIOVIA Discovery Studio 2017. The AutoDock program was used to construct a grid-box around the active site of each HLA allele for molecular docking. Furthermore, the AutoDock Vina script was used to perform molecular docking between the epitopes and their associated HLA alleles [51]. To compare epitope binding effectiveness, the corresponding co-crystal ligands were utilized as a positive control. BIOVIA Discovery Studio 2017 and PBDSum were used to visualize the docked complex.

**Formulating of multi-epitope vaccine**

The vaccine was created by combining the chosen CTL, HTL, and LBL epitopes with a suitable adjuvant and linking them with the proper linkers [52,53]. Because Toll-like receptor 4 (TLR4) is recognized by viral glycoproteins, and the adjuvant is essential for optimum translation and synthesis of the target vaccine candidate, we employed TLR4 agonist as the adjuvant [54,55]. As a result, the adjuvant 50S ribosomal protein L7/L12 (NCBI ID: P9WHE3) was evaluated to boost the vaccine candidate’s immunogenicity. With the EAAAK bi-functional linker, which can break apart two b domains with weakly interacting interactions over a wide range of peptide lengths, the adjuvant was attached to the vaccination front. In contrast, the selected CTL was linked with the help of Ala-Ala-Tyr (AAY) linkers, the HTL was linked with GlyPro-Gly-Pro-Gly (GPFPGP) linkers and the LBL was linked with Lys-Lys (KK) linker [47,52]. The AAY linker is a proteasome cleavage site that has been exploited to modify protein stability, decrease immunogenicity, and improve epitope presentation [56]. With GPGPG, a ‘functional epitope’ is avoided, which simplifies immune processing, while the bi-lysine KK linker helps to maintain the separate immunogenic properties of the vaccine construct.

**Physicochemical and immunological evaluation**

The physiochemistry of a protein describes its fundamental characteristics. The ProtParam server, which can be found at https://web.expasy.org/protparam/, was used to predict the vaccine’s physicochemical properties to comprehend the vaccine’s essential essence [57]. We also evaluated the immunological properties through Vaxijen v2.0 [35], MHC-I immunogenicity [39], AllerTop [41], and SOLpro [36] servers.

**Secondary structure prediction**

The two-dimensional (2D) structural features such as alpha-helix, beta-turn, and random coils of the construct were identified by SOPMA (Self-Optimized Prediction Method with Alignment) server at https://npsa-prabi.ibcp.fr/NPSA/npsa_seccons.html [58] and PSIPRED v4.0 (PSI-blast based secondary structure prediction) server at http://bioinf.cs.ucl.ac.uk/psipred/ [59] with default parameters. SOPMA has a prediction accuracy of above 80% [58]. To further understand the vaccine’s composition quality, 2D structural characteristics were retrieved and assessed.

**Homology modeling, 3D structure refinement, and validation**

The constructed vaccine was submitted into the RaptorX server (http://raptorx.uchicago.edu/) [60]. Using a cutting-edge algorithm and a 3D structure, the RaptorX server produces the most precise structure of the protein and its activities [60]. The C-score, TM-score value, root mean square deviation (RMSD), and top five models of a particular protein sequence may all be predicted and determined using this web service. The generated 3D structure was saved as a PDB file, which was chosen based on the C-score. The C-score on the server ranges from −5 to 2, with a higher number indicating a more confident protein model. For the refinement of the vaccine structure, the discovered 3D structure was uploaded to the GalaxyRefine (http://galaxy.seoklab.org/refine) online web-based server. The CASP10 refining approach was used to operate this webserver [61]. The RMSD, energy score, and overall quality score are all available on the GalaxyRefine website.
The improved structure was downloaded, and the chosen structure was determined using the energy scores of the lowest and maximum RMSD values. PyMOL v2.3.4 was used to show the refined and discovered structure [62]. The Ramachandran plot score (vaccine structure validity) and Z-score value, which identify the standard deviations from the mean value, were used to analyze the final 3D structure. The Ramachandran plot was analyzed by the Rampage server (http://mordred.bioc.cam.ac.uk/rapper/rampage.php), which runs considering allowed and disallowed regions of amino acid [63]; and Z-score plot was analyzed by the ProSA-web (https://prosa.services.came.sbg.ac.at/prosa.php) tool [64].

**Molecular docking studies**

The binding interactions between modeled proteins and receptor molecules can be revealed through molecular docking experiments. For this, we used the ClusPro v2.0 server, which can be found at https://cluspro.bu.edu/, to submit the refined vaccine model as a ligand and the TLR4 protein as an immunological receptor for molecular docking [65]. The TLR4 receptor was chosen and downloaded from the PDB server (PDB ID: 4G8A). Separating the associated ligand from the protein was the first step in preparing the receptor, which was followed by the removal of water and other chemicals. All of these procedures were carried out using the PyMOL v2.3.4 program [62]. Discovery Studio 2017 and PBDSum were used to investigate binding interactions and residues in the interacting surface.

**Molecular dynamics simulation**

The complex structure of the selected candidate compounds was evaluated using 50 ns molecular dynamic simulations (MDS) to evaluate their binding stability to the desired protein to the active site cavity of the protein [66]. The MDS of the receptor-ligand complex was performed using the 'Desmond v6.3 Program' in Schrödinger 2020-3 under Linux framework to evaluate the thermodynamic stability of the receptor-ligand complex [67]. To solve the system, a predetermined TIP3P water model was used, with an orthorhombic periodic boundary box form with a box distance of 10 Å assigned to both sides to retain a specific volume. After constructing the solvated system containing protein in complex with the ligand, the system has been minimized and relaxed using the default protocol introduced within the Desmond module with OPLS_2005 force field parameters [67]. In protein preparation wizard: Initially, protein preprocesses by adding hydrogens, create disulfide bonds, fill in the missing side chains, and delete waters using Epik (pH: 7.0 ± 2.0) and optimize by PROPKA pH: 7.0. In model system for simulation run, simulation time = 50 ns, trajectory intervals = 50 ps, total number of frames = 1,000, Ensemble class = NPT, temperature = 300 K, and one atmospheric (1.01325 bar) pressure. Finally, the simulation was carried out for 100 ns, and root mean square fluctuation (RMSF), RMSD, and protein secondary structure elements from the trajectories were analyzed to reveal the stability of the vaccine complex.

**Immune response simulation**

Using the C-IMMSIM v10.1 server (http://www.cbs.dtu.dk/services/C-ImmSim-10.1/) , the entire construct was uploaded for assessment of the vaccine’s potential immunological response [68]. As previously stated, we used a minimum gap of 30 days between two dosages in this situation [69]. Three injections were administered in silico with time steps of 1, 84, and 168, respectively, where one-time step equals 8 h in real life. With the maximum simulation step value set to 300, all other stimulation parameters were left at their default settings.

**Codon adaptation and in-silico cloning**

Codon optimization is required for the expression of a foreign gene in a host organism [70]. As a result, the construct was uploaded to the JCat service for codon adaptation (http://jcat.de/). We employed the commonly used E. coli K12 as the host in this study, and the entire procedure was carried out while avoiding the following three criteria: Sites of restriction enzyme cleavage, binding sites of prokaryotic ribosomes, and rho-independent transcription termination. The codon adaptation index (CAI) value and guanine–cytosine (GC) concentration of the modified sequence were used to evaluate it [70]. Lastly, the in-silico cloning of the adapted nucleotide sequence into the pET28a (+) expression vector was performed using the modified nucleotide sequence. SnapGene v4.2 software was used to carry out the entire in-silico cloning procedure [71].

**Results**

**Highest antigenic protein selection**

The retrieved V. harveyi proteomes featured hemolysin protein. We chose a hemolysin protein with the highest antigenic score of 0.4070 (VaxiJen) and 0.617 (ANTIGENpro) from all examined proteins based on antigenicity. The chosen hemolysin had a length of 418 amino acids and a GenBank accession number of ACF32997.1. For subsequent investigation, the main sequence of the chosen protein was employed.
Potential CTL epitopes
To design a rational vaccine, accurate predictions of CTL epitopes are crucial. Furthermore, they can minimize the amount of experimental effort needed to identify epitopes. From the chosen hemolysin protein, a total of 52 CTL epitopes with a length of nine amino acids were predicted by using NetCTL v1.2 server. 22 CTL epitopes were shown to be antigenic, immunogenic, non-toxic, and non-allergenic. We chose the top three CTL epitopes for the final vaccine design based on the antigenicity score due to the large number of possible epitopes (Table 1). C-score is the combined score provided by the NetCTL server.

Potential HTL epitopes
Initially, the IEDB server was used to identify 358 HTL epitopes, each with a length of 15 amino acids. Only 14 HTL epitopes were able to trigger the three kinds of cytokines tested, including IFN-γ, IL-4, and IL-10. Similarly, based on the antigenic score, we examined the top three HTL epitopes for incorporation into the final vaccine design (Table 2).

Potential LBL epitopes
To develop epitope-based vaccines, produce antibodies, and prevent and diagnose diseases, B-cell epitopes must be identified. In this study, a preliminary investigation found 10 LBL epitopes, each of which is 12 amino acids long. Later with further evaluation, two epitopes were found as antigenic, non-toxic non-allergic (Table 3).

Docking studies of epitope and alleles
The docking approach was utilized to confirm the efficiency of chosen epitopes in binding their HLA alleles. Table 4 lists the epitopes, as well as their corresponding docking alleles, binding affinities, interactions, and hydrogen-bonding residues. CTL epitopes had binding affinities of between −6.1 and −8.4 kcal/mol, while HTL epitopes had binding affinities of between −5.9 and −6.8 kcal/mol. In addition to the tabulated details, we presented the best interacting CTL (AQAKQTYTY) and HTL (DATRAPQFTYSTQEE) epitopes in Fig. 2. Herein, the best CTL epitope produced a total of nine hydrogen bonds, in which eight were classical interactions involved with the active site residue Tyr9, Leu8, Thr7, Glu166, Lys66, Arg170, Tyr4, Trp167, and Ala1. On the other hand, the best HTL epitope showed nine hydrogen bonds, including seven classical interactions while it interacted with Asp29, Lys58, Thr8, Asp30, Thr233, Ser57, Gln5, Glu212, and Lys4 residues.

Vaccine construct and basic properties
The vaccine was created utilizing eight epitopes from three distinct classes that had previously been chosen (3 CTL, 3 HTL, and 2 LBL). As illustrated in Fig. 3, the epitopes were linked together us-

---

Table 1. The selected CTL epitopes for the final vaccine construction

<table>
<thead>
<tr>
<th>Epitope</th>
<th>C-score</th>
<th>Antigenicity</th>
<th>Immunogenicity</th>
<th>Toxicity</th>
<th>Allergenicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQGYNVTLY</td>
<td>1.0098</td>
<td>1.5077</td>
<td>Positive</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>WAENPDSGY</td>
<td>1.8227</td>
<td>0.6526</td>
<td>Positive</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>AQAKQTYTY</td>
<td>1.2614</td>
<td>1.0896</td>
<td>Positive</td>
<td>Negative</td>
<td>Negative</td>
</tr>
</tbody>
</table>

CTL, cytotoxic T lymphocyte.

Table 2. The selected HTL epitopes for the final vaccine construction

<table>
<thead>
<tr>
<th>Epitope</th>
<th>Antigenicity</th>
<th>IFN-γ</th>
<th>IL-4</th>
<th>IL-10</th>
<th>Toxicity</th>
<th>Allergenicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITLLSALLPLSLAH</td>
<td>0.6989</td>
<td>Positive</td>
<td>Inducer</td>
<td>Inducer</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>QAKQTYTYVRCWRYRT</td>
<td>0.4670</td>
<td>Positive</td>
<td>Inducer</td>
<td>Inducer</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>DATRAPQFTYSTQEE</td>
<td>0.8194</td>
<td>Positive</td>
<td>Inducer</td>
<td>Inducer</td>
<td>Negative</td>
<td>Negative</td>
</tr>
</tbody>
</table>

HTL, helper T-lymphocyte; IFN-γ, interferon γ; IL, interleukin.

Table 3. The selected LBL epitopes for the final vaccine construction

<table>
<thead>
<tr>
<th>Epitope</th>
<th>Probability</th>
<th>Antigenicity</th>
<th>Allergenicity</th>
<th>Toxicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>KERCQTLDLAN</td>
<td>0.327</td>
<td>0.8769</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>NTLTLEFGLND</td>
<td>0.405</td>
<td>0.8716</td>
<td>Negative</td>
<td>Negative</td>
</tr>
</tbody>
</table>

LBL, linear B-lymphocyte.
Table 4. Binding affinities and interaction between selected epitopes and HLA alleles

<table>
<thead>
<tr>
<th>T-cell epitope</th>
<th>HLA allele</th>
<th>Epitope affinity (kcal/mol)</th>
<th>Control affinity (kcal/mol)</th>
<th>No. of hydrogens bonds (CHB)</th>
<th>Residues involved in CHB networks (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQGYNVTLY</td>
<td>HLA-B*3508</td>
<td>–7.2</td>
<td>–9.2</td>
<td>7 (6)</td>
<td>Gln69, Trp149, Thr7, Ile8, Met19, Ala2, Ile7 (7)</td>
</tr>
<tr>
<td>WAENPDGSY</td>
<td>HLA-A*0201</td>
<td>–6.1</td>
<td>–8.2</td>
<td>7 (5)</td>
<td>Tyr84, Lys146, Val2, Thr7, Val9, Asn77, Thr143 (7)</td>
</tr>
<tr>
<td>AQAKQTYTY</td>
<td>HLA-B*3501</td>
<td>–8.4</td>
<td>–8.2</td>
<td>9 (8)</td>
<td>Tyr9, Leu8, Thr7, Glu166, Lys66, Arg170, Tyr4, Trp167, Ala1 (9)</td>
</tr>
<tr>
<td>ITLLALLLPLSLAH</td>
<td>DRB1*07:01</td>
<td>–5.9</td>
<td>–6.9</td>
<td>9 (7)</td>
<td>Arg71, Thr77, Asn82, Ala12, Thr13, Val14, Val1, Glu6, Ser4 (9)</td>
</tr>
<tr>
<td>QAQKQTYTYVRCWYRT</td>
<td>DRB1*11:01</td>
<td>–6.1</td>
<td>–6.7</td>
<td>12 (10)</td>
<td>Tyr7, Asp9, Asp9, Ser24, Glu63, Lys66, Arg69, Arg69, Tyr99, Glu152, Glu152, Gln155 (12)</td>
</tr>
<tr>
<td>DATRAPQFTSTQEE</td>
<td>DRB1*11:01</td>
<td>–6.8</td>
<td>–7.3</td>
<td>9 (7)</td>
<td>Asp29, Lys58, Thr8, Asp30, Thr233, Ser57, Gln5, Glu212, Lys4 (9)</td>
</tr>
</tbody>
</table>

HLA, human leukocyte antigen.

Fig. 2. Interaction between epitopes and their respective binding alleles. As a representation of all-selected epitopes, we offer the docking interactions of the best HTL and CTL epitopes, where interaction between the HLA-B*3501 alleles and CTL epitope QAQKQTYTY (A, B) and docking between the DRB1*11:01 alleles and HTL epitope DATRAPQFTSTQEE (C, D). HTL, helper T-lymphocyte; CTL, cytotoxic T lymphocyte.
ing AAY, GPGPG, and KK linkers, respectively. To enhance immunogenicity, an adjuvant was applied before the construct. Using the EAAAK linker, the TLR4 agonist 50S ribosomal protein L7/L12 was connected to the initial CTL epitope as an adjuvant. The final vaccination had a length of 268 amino acids (Fig. 4).

**Physicochemical properties and immunological evaluation**

Table 5 shows the physicochemical parameters of the vaccine construct. The construct was discovered to have a molecular weight of 27,044.64 Da. Other features such as the theoretical isoelectric point (pI) of 4.95, the chemical formula of C₁₂₁₄H₁₹₁₄N₃₁₀O₃₇₉S₄, the instability index of 20.25, the aliphatic index of 82.05, and the grand average of hydropathicity of −0.237 were also present. The construct’s physicochemical properties and immunological efficacy were also assessed. For example, the construct’s antigenicity was 0.7017, whereas its immunogenicity was 1.59238. Furthermore, the vaccine was non-allergenic and soluble, with a score of 0.891723 out of 1 (Table 5). α-helix, β-strand, and random coils were examined utilizing two distinct servers as secondary structural characteristics. The SOPMA server predicted 33.07% α-helix, 16.93% β-strand, and 50% random coils in the construct (Table 6). On the other hand, the PSIPRED server anticipated the features as 42.91% α-helix, 20.47% β-strand, and 36.61% random coils (Table 6, Fig. 5).

**Tertiary structure, refinement, and validation**

The RaptorX server was utilized as the best template to build the top five models in homology modeling. We chose the model with

---

**Fig. 3.** Graphical map of the formulated multi-epitope vaccine construct. Herein, the adjuvant and the first CTL epitope were linked by EAAAK linker, CTL epitopes were added together by AYY linkers, HTL epitopes by GPGPG linkers and LBL epitopes by KK linkers. CTL, cytotoxic T lymphocyte; HTL, helper T-lymphocyte; LBL, linear B-lymphocyte; TLR4, Toll-like receptor 4.

**Fig. 4.** Constructed vaccine sequence.
the lowest C-score (–4.87), as advised by the server, out of the five. With GDT-HA score 0.8176, RMSD value 0.519, MolProbity 2.993, Clash score 25.7, and Poor rotamers score 0.8, the vaccine (model 1) exhibited 87.7% residues in the favorable area in the Ramachandran plot after refinement. The ProSA-web servers were used to further evaluate the refined 3D vaccine model. The vaccine's Ramachandran plot showed 78.5% residues in the favorable zone, 18.7% in approved regions, and 0.5% residues in prohibited regions before refining. The Ramachandran plot of the refined vaccine model showed 87.7% residues in the favorable region and 10.5% in allowed regions, while 0.5% residues in disallowed regions (Fig. 6B). Similarly, the crude model had a Z-score of –5.69, but the refined model had a Z-score of –6.01 (Fig. 6D). Fig. 7 shows a structural depiction of the developed vaccine.

Molecular docking studies
To predict their binding affinity and interactions, the vaccine (ligand) and TLR4 (receptor) were docked. As a result, the ClusPro v2.0 server produced ten docked complexes in various positions. We chose the complex with the lowest energy score and the binding posture with functional interactions from among them. As a result, model 1 met the inclination criterion. As a result, it was chosen as the best vaccine–TLR4 complex, with a –937.6 energy score. Binding interactions and residues implicated in active site residues were investigated in the chosen complex. A total of eight hydrogen bonds were found in the interaction surface. There were eight classical hydrogen bonds among the hydrogen bonds. The interacting residues in the CHB from the vaccine were Lys39, Lys20, Ser45, Asn47, His62, Arg67, and Asn44. Moreover, associated TLR4 active site residues are shown in Fig. 8. Other hydrogen bond interactions were as follows: three were electrostatic salt bridges, zero were disulphide bonds and nine single non-banded contact.

Molecular dynamics simulation
We calculated the RMSD for the vaccine complex and the vaccine. The vaccine complex had an average RMSD of 4.76Å, which indicated structural stability during the interaction. Fig. 9 shows that the vaccination complex has an early rise in RMSD characteristics until 5 ns, following which it becomes stable until 15 ns. From 15

---

Table 5. Antigenic, allergenic and physicochemical characteristics of the construct

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Finding</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of amino acids</td>
<td>254</td>
<td>Suitable</td>
</tr>
<tr>
<td>Molecular weight</td>
<td>27,044.64</td>
<td>Average</td>
</tr>
<tr>
<td>Theoretical pl</td>
<td>4.95</td>
<td>Acidic</td>
</tr>
<tr>
<td>Chemical formula</td>
<td>C1214H1914N310O379S4</td>
<td>-</td>
</tr>
<tr>
<td>Extinction coefficient (at 280 nm in H2O)</td>
<td>27515</td>
<td>-</td>
</tr>
<tr>
<td>Estimated half-life (mammalian reticulocytes, in vitro)</td>
<td>30 h</td>
<td>-</td>
</tr>
<tr>
<td>Estimated half-life (yeast-cells, in vivo)</td>
<td>&gt; 20 h</td>
<td>-</td>
</tr>
<tr>
<td>Estimated half-life (Escherichia coli, in vivo)</td>
<td>&gt; 10 h</td>
<td>-</td>
</tr>
<tr>
<td>Instability index of vaccine</td>
<td>20.25</td>
<td>Stable</td>
</tr>
<tr>
<td>Aliphatic index of vaccine</td>
<td>82.05</td>
<td>Thermostable</td>
</tr>
<tr>
<td>Grand average of hydrophaticity (GRAVY)</td>
<td>–0.237</td>
<td>Hydrophobic</td>
</tr>
<tr>
<td>Antigenicity</td>
<td>0.7017</td>
<td>Antigenic</td>
</tr>
<tr>
<td>Immunogenicity</td>
<td>0.7122</td>
<td>Immunogenic</td>
</tr>
<tr>
<td>Allergenicity</td>
<td>No</td>
<td>Non-allergen</td>
</tr>
<tr>
<td>Solubility</td>
<td>0.891723</td>
<td>Soluble</td>
</tr>
</tbody>
</table>

Table 6. The secondary structural features of the vaccine construct

<table>
<thead>
<tr>
<th>Feature</th>
<th>SOPMA server</th>
<th>PSIPRED server</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Amino acid</td>
<td>Percentage</td>
</tr>
<tr>
<td>Alpha helix</td>
<td>84</td>
<td>33.07</td>
</tr>
<tr>
<td>Beta strand</td>
<td>43</td>
<td>16.93</td>
</tr>
<tr>
<td>Random coil</td>
<td>127</td>
<td>50</td>
</tr>
</tbody>
</table>

SOPMA, Self-Optimized Prediction Method with Alignment; PSIPRED, PSI-blast based secondary structure prediction.
to 25 ns, there was a decreased degree of fluctuation, which might be responsible for structural integrity and/or allowing solid binding. Moreover, the protein flexibility across the amino acid residues was evaluated through the RMSF score. The RMSF profile of the vaccine complex indicates maximum amino acid residues from complexes that an RMSF profile below 4.0 Å and greater change was observed for fewer residues. This result from Fig. 10 defines the vaccine complex stability and stiffness.

### Immune response simulation

The simulated immune response in Fig. 11 mimicked the immune response induced by certain infections. Secondary and tertiary immune responses, for instance, were greater than primary immunological responses (Fig. 11A). Secondary and tertiary responses revealed larger levels of antibodies (IgG1 + IgG2, IgM, and IgG + IgM), which correlated with an antigen extenuation showing the establishment of memory cells, resulting in increased antigen clearance after subsequent exposures (Fig. 11A). Furthermore, B-cells, cytotoxic T cells, and helper T cells had a longer time of survival, indicating class flipping between immune cells and IgM memory development (Fig. 11B–11D). The Th0 type immune reaction had a lower proportion (%) and number (cells/mm³) than the Th1 type immune reaction (Fig. 11I). Expanded macrophage mobility was seen during the presentation, but dendritic cell movement was predicted (Fig. 11F and 11G).

### Codon adaptation and in-silico cloning

To improve the translation efficiency of the vaccine design, we ad-
justed the codons according to the *E. coli* K12 on the JCat service. The nucleotide sequences created by the peptide vaccine construct (254 amino acid residues) totaled 761 lengths (Fig. 12). Furthermore, the modified nucleotide sequence has a GC content of 59.39% and a CAI value of 0.62, respectively. We used *XhoI* and *BamHI* restriction sites as the start and end cut points, accordingly, to insert the modified sequence into the pET28a (+) vector. Using the SnapGene program, the optimized vaccine design was cloned into the pET28a (+) cloning vector (Fig. 13).

**Fig. 6.** (A, B) Analysis of Ramachandran plot PROCHECK server. The MFR, AAR, GAR, and DR was represented the most favored, additional allowed, generously allowed, and disallowed regions of vaccine. (C, D) 3D structure validation with a Z-score by Pro-SA server.
Discussion

The current diabolical emergence of *Vibrio harveyi* causing vibriosis poses a serious danger to the worldwide aquaculture industry [72], which influences us to use an immunoinformatics method to build this multi-epitope vaccination. The vaccination based on the hemolysin protein displayed outstanding relevance as predicted by immunoinformatics, proving our effort to be reliable. A vaccine protects against infectious illnesses in a safe and effective manner [73]. Acquired immunity against contagious diseases should be
**Fig. 9.** Molecular dynamic simulation of the multi-epitope vaccine complex. The root mean square deviation (RMSD) plot of the backbone atoms of the complexes.

**Fig. 10.** Molecular dynamic simulation of the multi-epitope vaccine complex. The root mean square fluctuation (RMSF) plot of the multi-epitope docked vaccine candidate. $\alpha$-helical and $\beta$-strand regions are highlighted in red and blue backgrounds, respectively. These regions are defined by helices or strands that persist over 70% of the entire simulation.
Fig. 11. Immune response triggered by the designed vaccine. The graph shows primary, secondary and tertiary immune responses (A), B-cell population (B), cytotoxic T-cell population (C), helper T-cell population (D), induction of cytokines and interleukins (E), dendritic cell population per state (F), macrophage (MA) population per state (G), natural killer (NK) cells (total count) (H), and percentage (%) and amount (cells/mm$^3$) of Th1 mediated (I).
Fig. 12. Codon adaptation of Epstein-Barr virus to Escherichia coli K12 strain.

Fig. 13. The proposed vaccine was cloned into the pET-28a (+) vector in silico.
possible with it [74]. As a result of this study, we designed a vaccine based on epitopes that would provide a strong immune response to *V. harveyi*. *V. harveyi* infection and transmission are difficult to control and prevent in the absence of an effective vaccine. Furthermore, in order to regulate the current situation, effective immunization has yet to be produced. As a result, a novel vaccine development strategy is critical to finding a solution to the current economically threatening aquaculture problem. Because the hemolysin of *V. harveyi* is important for immunological invasion and fish-to-fish transmission, our goal was to develop an epitope vaccination that targeted the hemolysin. In order to enable cellular and humoral immune systems to recognize this protein, the hemolysin protein surface was evaluated for its antigenic region. Previously, scientists develop an *in-silico* designing of epitope-based vaccine against the seven banded grouper nervous necrosis virus affecting fish species [31]. As vibriosis is very common in tilapia species and in the past studies MHC class I and class II molecules were found in the experimental data of cod and tilapia for starting immune responses against infections. So, this multi-epitope vaccine targeting HLA-A*0201*, HLA-B*3501*, and HLA-B*3508* might be an efficient vaccination against certain fish diseases [21,23]. The first step was identifying all possible CTL, HTL, and LBL epitopes. Next, vaccines were designed with three antigenic epitopes—CTL, HTL, and LBL—since the linkers below corresponded to the top three epitopes. They were used in vaccine development as an important component that improves the stability, folding, and transcriptional regulation of our peptide vaccine [75]. The adjuvant was attached to the CTL epitope by EAAAK linker, which helps to induce high levels of both cellular and immunogenic humoral responses for particular antigens, and amplify the vaccine’s stability and longevity [76]. A total of 254 amino acid residues were found in the vaccine construction. An essential characteristic of a recombinant vaccine is its solubility, a type of physico-chemical property [77]. A solubility assessing tool was used to determine whether the vaccine construct was soluble inside the host *E. coli*, and the results showed that it was soluble. The vaccine’s nature, as indicated by the theoretical PI value, was acidic. The protein’s stability index, as recommended by server tools, indicates that it will be stable following synthesis. The GRAVY (grand average of hydropathicity index) value and aliphatic index, on the other hand, indicated that the vaccine was hydrophilic and thermostable, respectively. According to the prediction of physicochemical properties and scores on all parameters, there is a high probability for this vaccine to be a valid candidate against hemolysin protein of *V. harveyi*. The detected models were revised and the best model (based on the lowest energy score) was chosen after the 3D structure prediction (based on C-score). We observed a reasonable number of Z-score (–6.01) and superior features of most favored, acceptable, and prohibited areas for the Ramachandran plot in the validation test of 3D structure. It was suggested by the lowest energy score of 937.6 for a molecular docking between the peptide vaccine and virus glycoprotein binding convenient receptor of TLR4 that the vaccine could have infection-inhibiting activity and might interact tightly with TLR4 receptor. The molecular dynamics simulation is a potentially useful tool for understanding how proteins function and how their structure is derived. Anatomical movement can be simulated by protein dynamic simulations as a function of time. We have performed dynamic simulations of the vaccine candidate for 50 ns, and analyzed the results using the RMSD and RMSF scores. When comparing distinct atomic conformations of a molecular system, the RMSD value is employed. A significant flexibility and departure of vaccine candidates from receptor structure was determined using the RMSD value, whereas the displacement of our particular vaccine candidate’s atoms from receptor structure was determined using RMSF of the complex structure. The calculated average RMSD and RMSF value was 4.77 Å and 4.0 Å, respectively. The fluctuation was not observed to be larger in the vaccine section, but it smoothed out after 5 ns, suggesting that the modeled vaccine and receptor are stable. Lastly, we examined the optimal target clearance and cell density parameters for the best immunologic response against the pathogen by constructing an immune simulation. As a result of the upgraded vaccine doses, the immune system created memory B-cells (with a half-life of several months) and T cells. The vaccination efficiently imitated a humoral immune response to increased immunoglobulin production in this way. In order to optimize the multi-epitope vaccine production, the MD simulation was done to evaluate stability of the vaccine candidate with the receptor, in which codon optimization was done for stability of the construct vaccine within the host. Eventually, the codon was adjusted, and *in-silico* cloning of the intended vaccine candidate into the *E. coli* K12 expression host pET28a (+) vector was successful.

A range of computational techniques were used in this work to find possible T- and B-cell epitopes in *V. harveyi* hemolysin protein, which were finally stitched into a multi-epitope mRNA vaccine. The newly developed vaccine possesses the immunodominant qualities that are sought. Significantly, it was capable of binding to the immunological receptor TLR4 and induce a substantial immune response in regard to *V. harveyi* infection. Based on our findings, we believe that developing a vaccine against the etiological agent of the *V. harveyi* outbreak in fish should begin with the
vaccine candidate. In addition, the possible epitopes discovered in this study can be employed in future research. Nevertheless, more testing is needed to show that our designed vaccine is an effective preventive against *V. harveyi* infection in fish species.

**ORCID**

Sk Injamamul Islam: https://orcid.org/0000-0002-0888-6075
Moslema Jahan Mou: https://orcid.org/0000-0002-2203-4207
Saloa Sanjida: https://orcid.org/0000-0003-0982-0876
Muhammad Tanq: https://orcid.org/0000-0001-7516-1579
Saad Nasir: https://orcid.org/0000-0003-4953-2131
Sarower Mahfuj: https://orcid.org/0000-0002-1014-8565

**Authors’ Contribution**

Conceptualization: SII, MJM. Data curation: SII, SS. Formal analysis: SII, MJM, SS. Funding acquisition: SM. Methodology: SII, SN, SS. Writing - original draft: MJM, SS. Writing - review & editing: SII, MJM, MT.

**Conflicts of Interest**

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

**Acknowledgments**

The author thanks Dr. Foysal Ahmed Sagore and Dr. Kazi Abdus Samad for helpful comments.

**References**


69. Castiglione F, Mantile F, De Berardinis P, Prisco A. How the in-


Characterization of the first mitogenomes of the smallest fish in the world, *Paedocypris progenetica*, from peat swamp of Peninsular Malaysia, Selangor, and Perak

NorJasmin Hussin¹², Izzati Adilah Azmir²*, Yuzine Esa³, Amirrudin Ahmad⁴⁵, Faezah Mohd Salleh⁶, Puteri Nur Syahzanani Jahari⁶, Kaviarasu Munian⁷, Han Ming Gan⁸

¹Faculty of Applied Sciences, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Malaysia
²School of Biology, Faculty of Applied Sciences, Universiti Teknologi MARA (UiTM) Negeri Sembilan, Kampus Kuala Pilah, Pekan Parit Tinggi, 72000 Kuala Pilah, Malaysia
³Department of Aquaculture, Faculty of Agriculture, Universiti Putra Malaysia, 43400 Serdang, Malaysia
⁴School of Marine and Environmental Sciences, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Malaysia
⁵Institute of Tropical Biodiversity and Sustainable Development, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Malaysia
⁶School of Biosciences, Faculty of Science, Universiti Teknologi Malaysia, Johor, Johor Bahru 81310, Malaysia
⁷Forest Biodiversity Division, Forest Research Institute Malaysia, Kepong 52109, Malaysia
⁸GeneSEQ Sdn Bhd, Bukit Beruntung, 48300 Rawang, Malaysia

The two complete mitochondrial genomes (mitogenomes) of *Paedocypris progenetica*, the smallest fish in the world which belonged to the Cyprinidae family, were sequenced and assembled. The circular DNA molecules of mitogenomes P1-*P. progenetica* and S3-*P. progenetica* were 16,827 and 16,616 bp in length, respectively, and encoded 13 protein-coding genes, 22 transfer RNA genes, two ribosomal RNA genes, and one control region. The gene arrangements of *P. progenetica* were identical to those of other *Paedocypris* species. BLAST and phylogenetic analyses revealed variations in the mitogenome sequences of two *Paedocypris* species from Perak and Selangor. The circular DNA molecule of *P. progenetica* yield a standard vertebrate gene arrangement and an overall nucleotide composition of A 33.0%, T 27.2%, C 23.5%, and G 15.5%. The overall AT content of this species was consistent with that of other species in other genera. The negative GC-skew and positive AT-skew of the control region in *P. progenetica* indicated rich genetic variability and AT nucleotide bias, respectively. The results of this study provide genomic variation information and enhance the understanding of the mitogenome of *P. progenetica*. They could later deliver highly valuable new insight into data for phylogenetic analysis and population genetics.

**Keywords:** mitogenome, *Paedocypris progenetica*, Peninsular Malaysia
Introduction

The Paedocypris populations are rapidly declining worldwide due to anthropogenic and environmental actions that pose a threat to their survival [1]. According to Sam et al. [2], the evolution of small sizes, or miniaturization, is extensively seen in vertebrate species and is most commonly documented in amphibians and fishes. Southeast Asia harbors highly acidic blackwater peat swamps that serve as habitats for miniature fish, which are nearly all endemic to these habitats. The features of miniature phenotypes exhibit morphological novelty and increased morphological variability and are mostly unique combinations of ancestral phenotypes that are derived through structural simplification and reduction [3].

Interestingly, the smallest fish in the world, Paedocypris progenetica, is found in Peninsular Malaysia. However, the lack of its genomic data in GenBank could hinder the extensive study of this remarkable species. The mitochondrial genome (mitogenome) contains multiple genes that are noteworthy for ecological and evolutionary studies to investigate the phylogeny and biodiversity of complex species by using high-throughput sequencing technologies [4]. Hence, this study provided the whole mitogenome of P. progenetica from Peninsular Malaysia for the first time.

Methods

DNA sampling and sequencing

The samples of P. progenetica were collected from North Peat Swamp Selangor (3.39°N, 101.15°E) and Pondok Tanjung Perak (5.04°N, 100.4°E), Peninsular Malaysia in February 2021. Genomic DNA was extracted from the tissue of P. progenetica specimens by using a ReliaPrep gDNA Tissue Miniprep system (Promega, Madison, WI, USA), fragmented with a Bioruptor system, and the remaining tissue is currently deposited at University Putra Malaysia (UPM). The library was prepared by using a NEBNext Ultra II DNA Library Prep Kit for Illumina in accordance with the manufacturer’s protocol. The sample was then sequenced by using an Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA) with 150 paired-end modes (PE150) [5].

Mitogenome assembly, annotation, and sequence analysis

Sequencing adapters, low-quality stretches, and leading/tailing Ns were trimmed from the raw reads of the sequences by using AdapterRemoval V2.2.2 [6]. Forward and reverse reads were interleaved into single file and the assembly were carried out using two different softwares, NOVOPlasty v4.2 [5] and Megahit v1.2.9 [7], both using default k-mer sizes. For the assembly using NOVOPlas-
**Table 1.** The reported mitogenome of *Paedocypris progenetica* from Peninsular Malaysia

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Genbank accession number</th>
<th>Origin</th>
<th>Length (bp)</th>
<th>Sex</th>
<th>GC%</th>
<th>AT%</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>OK356905</td>
<td>Pondok Tanjung</td>
<td>16,827</td>
<td>Male</td>
<td>38.3</td>
<td>61.7</td>
</tr>
<tr>
<td>S3</td>
<td>OK413207</td>
<td>North Peat Swamp</td>
<td>16,616</td>
<td>Female</td>
<td>38.5</td>
<td>61.5</td>
</tr>
</tbody>
</table>

**Table 2.** Gene features of the mitochondrial genome of *Paedocypris progenetica*

<table>
<thead>
<tr>
<th>Gene</th>
<th>Position</th>
<th>Size (bp)</th>
<th>Start/Stop codon</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>tRNA-Phe</td>
<td>1-69</td>
<td>69</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>12S rRNA</td>
<td>70-1018</td>
<td>949</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Val</td>
<td>1019-1090</td>
<td>72</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>16S rRNA</td>
<td>1091-2762</td>
<td>1,672</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Leu</td>
<td>2763-2837</td>
<td>75</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>ND1</td>
<td>2838-3812</td>
<td>975</td>
<td>ATG/TAA</td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Ile</td>
<td>3813-3882</td>
<td>70</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Gln</td>
<td>3881-3951</td>
<td>71</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>tRNA-Met</td>
<td>3952-4020</td>
<td>69</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>ND2</td>
<td>4021-5059</td>
<td>1,039</td>
<td>ATG/T</td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Trp</td>
<td>5060-5130</td>
<td>71</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Ala</td>
<td>5133-5201</td>
<td>69</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>tRNA-Asn</td>
<td>5203-5275</td>
<td>73</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>tRNA-Cys</td>
<td>5310-5374</td>
<td>65</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>tRNA-Tyr</td>
<td>5375-5444</td>
<td>70</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>COI</td>
<td>5446-6990</td>
<td>1,545</td>
<td>GTG/TAG</td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Ser</td>
<td>6994-7064</td>
<td>71</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>tRNA-Asp</td>
<td>7066-7138</td>
<td>73</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>COII</td>
<td>7152-7842</td>
<td>691</td>
<td>ATG/T</td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Lys</td>
<td>7843-7915</td>
<td>73</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>ATPase-8</td>
<td>7917-8084</td>
<td>168</td>
<td>ATG/TAA</td>
<td>+</td>
</tr>
<tr>
<td>ATPase 6</td>
<td>8075-8757</td>
<td>683</td>
<td>ATG/TAA(A)</td>
<td>+</td>
</tr>
<tr>
<td>COIII</td>
<td>8758-9541</td>
<td>784</td>
<td>ATG/T</td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Gly</td>
<td>9542-9613</td>
<td>72</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>ND3</td>
<td>9614-9959</td>
<td>346</td>
<td>ATG/TAG</td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Arg</td>
<td>9960-10030</td>
<td>71</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>ND4L</td>
<td>10031-10327</td>
<td>297</td>
<td>ATG/TAA</td>
<td>+</td>
</tr>
<tr>
<td>ND4</td>
<td>10321-11702</td>
<td>1,382</td>
<td>ATG/TAA(A)</td>
<td>+</td>
</tr>
<tr>
<td>tRNA-His</td>
<td>11703-11771</td>
<td>69</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Ser</td>
<td>11772-11840</td>
<td>69</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Leu</td>
<td>11842-11914</td>
<td>73</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>ND5</td>
<td>11915-13750</td>
<td>1,836</td>
<td>ATG/TAG</td>
<td>+</td>
</tr>
<tr>
<td>ND6</td>
<td>13747-14268</td>
<td>522</td>
<td>ATG/TAA</td>
<td>–</td>
</tr>
<tr>
<td>tRNA-Glu</td>
<td>14270-14338</td>
<td>69</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>Cyt b</td>
<td>14341-15477</td>
<td>1,137</td>
<td>ATG/TAA</td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Thr</td>
<td>15480-15550</td>
<td>71</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>tRNA-Pro</td>
<td>15549-15618</td>
<td>70</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>Control Region</td>
<td>15619-16728</td>
<td>1,209</td>
<td></td>
<td>–</td>
</tr>
</tbody>
</table>

https://doi.org/10.5808/gi.21081
tionships between fish species. The stipulated data in Table 2 indicated that the ND6 and seven tRNA genes (tRNA^Glu, tRNA^Ala, tRNA^Asn, tRNA^Cys, tRNA^Ser, tRNA^Glu, and tRNA^Pro) were encoded on the L-strand, whereby, most of *P. progenetica* mitochondrial genes were encoded on the H-strand. This finding was consistent with the result reported by Sam et al. [2] on the mitogenomes of *P. micromegethes* and *P. carbunculus* as there were no significant changes found between the populations of *P. progenetica* between the conserved genes (PCGs, tRNAs and rRNAs). However, the mitogenome length of *P. micromegethes* and *P. carbunculus* were clearly different from those of *P. progenetica*, presumably because of the variations in the control region (D-loop).

**Phylogenetic relationship**

The mitogenomic phylogeny analysis clustered the two mitogenomes of *P. progenetica* (OK356905 and OK413207) with the mitogenome of *P. progenetica* from Indonesia (AP011287) [17] and rooted them with the mitogenomes of other *Paedocypris* species [2,18] (Fig. 1) with the high support of 100% bootstrap and 1.00 posterior probability. GenBank revealed that the closest match (> 96% similarity) was between the newly sequenced mitogenomes of *P. progenetica* from Peninsular Malaysia and the mitogenomes of *P. progenetica* (AP011287) from a peat swamp in Sumatera, Indonesia [18] (Table 1). Moreover, < 96% similarity was found among *Paedocypris* species. The molecular evidence strongly indicated that Clade 1, which included the *P. micromegethes* (NC_051487.1) subclade-1, comprised a stable monophyletic group. The latest research has found identical ancestral patterns for *Cirrhinus reba*, which aligned in the same clade containing the same species [19].

**Protein-coding genes**

The prominent features of *Paedocypris* mitochondrial genes are listed in Table 3, which indicates that all PCGs, except for the COI gene that began with GTG, began with the start codon (ATG). The seven PCGs including nad1, cox1, atp8, atp6, nad4l, nad4, and nad5 were terminated by a complete and canonical stop codon (TAA or TAG). However, the genes encoding cox2, cox3, nad2, and nad3 had a truncated stop codon. Similar to the finding reported by Sam et al. [2], except for the COI gene that was terminated by GTG, most PCGs in the mitogenomes of *P. carbunculus* and *P. micromegethes* were terminated by the codon TAR (TAA/TAG) or an incomplete codon (TA/-/T--). According to Zhong et al. [20], a truncated stop codon (T) is commonly found in the mitochondrial gene of metazoans, such as the spider *Habronattus ore-gonensis*, and does not affect mitochondrial gene transcription or

---

*Fig. 1.* The phylogenetic tree of *Paedocypris progenetica* mitogenomes (OK356905 and OK413207) and other *Paedocypris* species available in GenBank. The bootstrap values were indicated in each branch of the tree representing the result of neighbor-joining probability. *Schimatorhynchos nukta* was selected as an outgroup.
The composition and skewness between *Paedocypris progenetica* from P.M, Malaysia and Sumatera, Indonesia

<table>
<thead>
<tr>
<th>Feature</th>
<th>A + T%</th>
<th>AT skew</th>
<th>GC skew</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Indonesia</td>
<td>PM</td>
<td>Indonesia</td>
</tr>
<tr>
<td>Whole genome</td>
<td>61.22</td>
<td>61.1</td>
<td>0.1</td>
</tr>
<tr>
<td>PCGs</td>
<td>61.3</td>
<td>61.7</td>
<td>0.04</td>
</tr>
<tr>
<td>nad1</td>
<td>59.1</td>
<td>59.3</td>
<td>0.05</td>
</tr>
<tr>
<td>nad2</td>
<td>65.1</td>
<td>64.5</td>
<td>0.18</td>
</tr>
<tr>
<td>cox1</td>
<td>59.9</td>
<td>59.4</td>
<td>-0.05</td>
</tr>
<tr>
<td>cox2</td>
<td>62.1</td>
<td>61.2</td>
<td>0.09</td>
</tr>
<tr>
<td>atp8</td>
<td>64.3</td>
<td>65.5</td>
<td>0.11</td>
</tr>
<tr>
<td>atp6</td>
<td>64.3</td>
<td>63.1</td>
<td>-0.06</td>
</tr>
<tr>
<td>cox3</td>
<td>57.7</td>
<td>57.9</td>
<td>0.03</td>
</tr>
<tr>
<td>nad3</td>
<td>61.3</td>
<td>61.2</td>
<td>-0.02</td>
</tr>
<tr>
<td>nad4t</td>
<td>61</td>
<td>59.9</td>
<td>-0.01</td>
</tr>
<tr>
<td>nad4</td>
<td>63.1</td>
<td>64.2</td>
<td>0.10</td>
</tr>
<tr>
<td>nad5</td>
<td>62.5</td>
<td>62.6</td>
<td>0.11</td>
</tr>
<tr>
<td>nad6</td>
<td>59.4</td>
<td>60.6</td>
<td>-0.52</td>
</tr>
<tr>
<td>cytB</td>
<td>62.9</td>
<td>62.5</td>
<td>-0.02</td>
</tr>
<tr>
<td>tRNAs</td>
<td>59.5</td>
<td>60.36</td>
<td>0.02</td>
</tr>
<tr>
<td>mtDNA</td>
<td>59.9</td>
<td>59.8</td>
<td>0.25</td>
</tr>
</tbody>
</table>

P.M, Peninsular Malaysia.

translation because the complete stop codon is presumably obtained through post-transcriptional polyadenylation [21].

**Gene arrangements**

The remarkable species *P. progenetica* of both samples (S3 and P1) from Peninsular Malaysia were aligned with 96% similarity of *P. progenetica* collected from Indonesia (AP011287) retrieved in GenBank entry. The overall nucleotide composition of *P. progenetica* was 33.0% A, 27.2% T, 23.5% C, and 15.5% G and showed a slightly AT-rich region (60.25%); these results were consistent with the patterns found in most fish mitogenomes [22]. The nucleotide composition of the *P. progenetica* mitogenome was highly biased toward A + T and had similar values as other *Paedocypris* species, such as *P. carbunculus* from Banka. The PCGs had a slightly higher A + T content (61.7%) compared to ribosomal RNA genes (59.8%). Based on Table 3, the AT and GC skew of *P. progenetica* showed 0.10 and −0.20, respectively. The GC skews of all genes, except for those of NAD6 and tRNA, which were positive for both populations, were negative and indicated a regular pattern of base composition behavior in the *P. progenetica* mitogenome. This result agreed well with that of Sam et al. [2], who discovered the AT-skew was mainly positive and the GC-skew were mostly negative values in distinct gene regions of the *P. micromegethes* and *P. carbunculus* mitogenomes.

Meanwhile, a vast difference in nucleotide composition in the control region (D-loop region) located between trnP (tRNA<sub>Phe</sub>) and trnF (tRNA<sub>Pro</sub>) can be seen in this genus. The lengths of PCGs, tRNAs, and rRNAs were conserved, and the variations were mainly attributed to the control region. The lengths of the D-loop region in *P. micromegethes* and *P. carbunculus* were 1,590 and 1,662 bp, respectively, whereas we found the considerably shorter D-loop region length of 1,209 bp (OK356905 and OK413207). These results differed because the D-loop region exhibits a rapid evolutionary rate and tends to possess lower purifying selection compared to PCGs that amass variations in length [23]. According to Li et al. [24], noncoding regions in metazoan mitogenomes frequently vary in length from species to species. However, the D-loop region of the reference species collected from Indonesia (AP011287) was not recorded in GenBank. Eventually, future studies on the non-coding region may contribute genetic data and enhance studies on the genomic data of *P. progenetica*.

**Conclusion**

The full mitogenome sequence of *P. progenetica* was analyzed and compared with that of other *Paedocypris* species mainly focused on *P. micromegethes* and *P. carbunculus* in the Cyprinidae family. The mitogenome length of *P. progenetica* was shorter compared to other *Paedocypris* species predominantly due to variations in the D-loop region. The comparison of the complete mitogenome data
and phylogenetic relationships of *Paedocypris* species provided fundamental information for evolutionary biology and are particularly important for future studies using the D-loop region and whole-genome sequences to resolve the relationship among *Paedocypris* species fully.

**ORCID**

NorJasmin Hussin: https://orcid.org/0000-0002-8355-728X  
Izzatine Izzatine Azmir: https://orcid.org/0000-0002-2829-2741  
Yuzine Esa: https://orcid.org/0000-0002-4774-5574  
Amirrudin Ahmad: https://orcid.org/0000-0002-7775-1289  
Faezah Mohd Salleh: https://orcid.org/0000-0001-7492-583X  
Puteri Nur Syahzanani Jahari: https://orcid.org/0000-0002-5485-6088  
Kaviarasu Munian: https://orcid.org/0000-0001-8425-2675  
Han Ming Gan: https://orcid.org/0000-0001-7987-738X

**Authors’ Contribution**

Conceptualization: IAA, YE, AA. Formal analysis: FMS, HMG, KM. Methodology: PNSJ, HMG. Writing - original draft: NH, PNSJ. Writing - review & editing: IAA, FMS.

**Conflicts of Interest**

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

**Acknowledgments**

This study was fully funded by the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme (Ref No: RACER/1/2019/STG05/UiTM//5) through Universiti Teknologi Mara (UiTM) (Ref No: 600-IRMI/FRGS-RACER S/3 (058/2019). The authors would like to thank everyone from UPM, UiTM, UMT, Selangor Forestry Department, and local communities who were involved in sample collection. The authors would also like to thank the Fish Genetics and Breeding Laboratory, Department of Aquaculture, Faculty of Agriculture, Universiti Putra Malaysia (UPM) for providing facilities and chemicals for molecular work as well as University Technology Malaysia (UTM) aided in conducting the analysis for mitogenome. Lastly, the first author would like to thank the Jabatan Perkhidmatan Awam (JPA) for providing a scholarship for her postgraduate study.

**References**


**TcellInflamedDetector: an R package to distinguish T cell inflamed tumor types from non–T cell inflamed tumor types**

San-Duk Yang¹, Hyun-Seok Park²,³*

¹School of Integrated Software and Design, Kyung Hee Cyber University, Seoul 02447, Korea
²Bioinformatics Laboratory, ELTEC College of Engineering, Ewha Womans University, Seoul 03760, Korea
³Center for Convergence Research of Advanced Technologies, Ewha Womans University, Seoul 03760, Korea

A major issue in the use of immune checkpoint inhibitors is their lack of efficacy in many patients. Previous studies have reported that the T cell inflamed signature can help predict the response to immunotherapy. Thus, many studies have investigated mechanisms of immunotherapy resistance by defining the tumor microenvironment based on T cell inflamed and non–T cell inflamed subsets. Although methods of calculating T cell inflamed subsets have been developed, valid screening tools for distinguishing T cell inflamed from non–T cell inflamed subsets using gene expression data are still needed, since general researchers who are unfamiliar with the details of the equations can experience difficulties using extant scoring formulas to conduct analyses. Thus, we introduce TcellInflamedDetector, an R package for distinguishing T cell inflamed from non–T cell inflamed samples using cancer gene expression data via bulk RNA sequencing.

**Keywords:** gene expression, immune checkpoint inhibitors, immunotherapy, prognosis, RNA-seq, software

**Availability:** The R package TcellInflamedDetector is available on the Comprehensive Archive Network and on GitHub. It is distributed under the GNU General Public License. GitHub: [https://github.com/sandukyang/Tcellinflamed/blob/main/TcellInflamedDetector.tar.gz](https://github.com/sandukyang/Tcellinflamed/blob/main/TcellInflamedDetector.tar.gz).

---

**Introduction**

Cancer cells express programmed death ligand 1 as a signal related to T cell unresponsive-ness. Immunotherapies targeting immune checkpoints (e.g., anti–cytotoxic T lymphocyte associated antigen-4 and anti–programmed death-1 antibodies) are a standard component of care for patients with advanced cancers. Immune checkpoint inhibitors (ICIs) have led to improvements in the survival rate, but only a subset of patients respond to ICIs. Recent studies have reported that the efficacy of ICIs in cancer patients is determined by the T cell inflamed tumor microenvironment [1-3]. The molecular mechanisms of resistance have not yet been elucidated in detail. Nevertheless, previous studies have reported scoring methods for distinguishing non–T cell inflamed tumors based on gene expression data [4,5].

Unfortunately, general researchers who are unfamiliar with the detailed calculations involved in the equations can experience difficulties using these scoring formulas to con-
duct analyses. For this reason, we recently developed TcellInflamedDetector, an R package that predicts T cell inflamed tumors when given RNA-sequencing expression data. This package will be beneficial to optimize the selection of patients predicted to benefit from ICIs. TcellInflamedDetector implements the equation developed by Spranger et al. [5] to differentiate non–T cell inflamed and T cell inflamed tumor subtypes.

### Input Data and Processing

As shown in Fig. 1, TcellInflamedDetector requires RNA-sequencing count input data with genes and sample identifiers. Users follow the steps for data processing that are summarize in the TcellInflamedDetector manual on GitHub [6]. The input CSV file is RNA sequencing log count per million (CPM) data. The count matrix file is converted by EdgeR `aveLogCPM()` and the calcNormFactor function using the trimmed mean of the m-values method. Users can extract previously established gene signatures indicative of a T cell inflamed tumor microenvironment, which include the cytotoxic T lymphocyte (CTL) signature genes CD8A, CD8B, GZMA, GZMB, and PRF1 using R code [7-10]. The established gene signatures were referenced with the Gajewski T cell-inflamed signature, interferon-gamma related signature, T cell effector signature, and immune cytolytic activity signature [4,5].

#### Estimating T Cell Inflamed and Non–T Cell Inflamed Samples

As shown in Fig. 2, gene expression values were converted to a score $S_i = \mu_i \pm \beta_i \sigma_i (i = 1, 2, \ldots, n)$, where $\mu$ and $\sigma$ represent the mean and standard deviation (SD) of the $i^{th}$ gene's expression across all samples, $n$ is the total number of genes, $\beta$ represents the distance between the $i^{th}$ gene's expression in a sample and the mean in units of the SD (equivalent to a z-score). The threshold for non–T cell inflamed and T cell inflamed tumors was $\beta_0 = 0.1$. The algorithm is described in detail below:

- If the z-score value $\beta_i$ is greater than the threshold ($\beta_0 = 0.1$), then $+1$ is assigned.
- Otherwise, if the z-score value $\beta_i$ is less than the threshold ($\beta_0 = 0.1$), then $-1$ is assigned.
- If the sum of the column of genes with assigned values is greater than half of the number of CTL genes, then the output is a classification of “T cell inflamed.”
- If the sum of a column of genes with assigned values is less than half of the number of CTL genes, then the classification is “non–T cell inflamed.”
- Otherwise, the sample is classified as “intermediate.”

Users of the R package can obtain results in the format of a .csv file that contains data on the classification of samples as T cell inflamed and non–T cell inflamed.

---

**Fig. 1.** An exemplary code for extraction of T cell effector gene subset.

**Fig. 2.** An exemplary usage of R code.
Fig. 3. (A) An input file and generated output files. (B) Processing diagram for T cell inflamed function prediction. (C) T cell inflamed annotation of Heatmap expression in The Cancer Genome Atlas (TCGA) lung adenocarcinoma (LUAD) samples.
flamed, non–T cell inflamed, and intermediate. If users want to modify the CTL gene list when running the R package, they do not have to modify the complex R code. Instead, they can simply revise the gene list contained in the CTL.csv file.

Output

Five output formats are available: CTL_Selected_Inputfile.csv, Tcell_NonTcell_Result.csv, zscore_convert.csv, Zscore_convert_sum.csv, and zscore_data.csv. Fig. 3 presents examples of the prediction results of T cell inflamed, intermediate, and non–T cell inflamed groups. Users can check the expression patterns of specific genes through a heatmap. We also confirmed that T cell inflamed samples showed high expression of T cell effector gene signatures [10].

Finally, we conducted a test to demonstrate our tool’s flexibility; we tested it on The Cancer Genome Atlas (TCGA) lung adenocarcinoma RNA-sequencing dataset available through the TCGA Research Network [11]. Each sample was labeled according to the TCGA barcode, which contained gene names. Our package successfully selected subsets of gene expression data from the raw count data. Thus, TcellInflamedDetector can be beneficial for future cancer immunotherapy vaccine developers and researchers.

ORCID

San-Duk Yang: https://orcid.org/0000-0002-1276-365X
Hyun-Seok Park: https://orcid.org/0000-0002-1237-8831

Authors’ Contribution

Conceptualization: SDY, HSP. Data curation: SDY, HSP. Methodology: SDY, HSP. Writing - original draft: SDY, HSP.

Conflicts of Interest

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

Acknowledgments

This research was partially supported by Kyung Hee Cyber University.

References

**Introduction**

Permutation testing is a nonparametric randomization procedure that provides a robust and powerful method of testing statistical hypotheses. In the original form of the permutation test, the response is shuffled \( n \) times, and it is commonly used when the standard distributional assumptions are violated. Although this random permutation is an attractive method in genome-wide association studies (GWAS), there is little guidance to achieve calculation efficiency for adjusting p-values via \( N \)-permutations using computing architecture.

A GWAS is an approach used in genetics to associate specific genetic variations with a particular trait, such as hair color. For example, using a GWAS approach, Ozaki et al. [1] discovered that a single nucleotide polymorphism (SNP) associated to myocardial infarction. With larger data sets in plant breeding, GWAS becomes very powerful in analyzing genetic contributions for traits, such as weight of each grain in spring wheat [2]. However, control for multiple testing and repeated observation of longitudinal traits are known
limitations of GWAS. Although the permutation test can adjust type 1 errors of p-values in multiple tests, adaptive permutation approaches have been suggested, due to the computational cost of permutation, as an alternative solution [3]. Here, we present a software to solve this computational burden in GWAS.

Methods

This note describes an open-source application, message-passing interface (MPI)-GWAS, which was designed for the adjustment of p-values of association analysis, such as Fisher’s exact test, of GWAS via N-permutation for each locus of the trait of interest. To optimize the time cost, a MPI-based work-stealing parallel scheme was applied, where MPI is a portable message-passing standard designed for development of parallel applications leveraging the power of manycore architecture at a supercomputer scale. Fig. 1A depicts the algorithm of MPI-GWAS. In a work-stealing scheme, each MPI process has a queue of computational tasks to perform. Each task fits models of variants by dividing by the number of permutations sequentially, but during its execution, a task may also spawn new tasks that can feasibly be executed in parallel. These new tasks are initially put in the queue. When an MPI task finishes work, it looks at the queues of the other MPI tasks and steals tasks. In effect, work stealing distributes the scheduled work over idle computer central processing units (CPUs), and while all CPU resources are being computed, scheduling overhead does not occur. Thus, the calculation time is greatly reduced. Moreover, MPI-GWAS effectively conducts permutation using Julia, an open-source project for high performance computing (https://julialang.org/). Using Julia, MPI-GWAS demonstrated a ~2–3-fold decrease in elapsed time of R.

The definition of the problem can be summarized into the following equations. We model the data of GWAS as follows.

\[ G = \begin{bmatrix}
    g_{1,1} & \ldots & g_{1,k} \\
    \vdots & \ddots & \vdots \\
    g_{m,1} & \ldots & g_{m,k}
\end{bmatrix}, \quad T^r = [t_1, t_2, \ldots, t_M]^T
\]

where \( g_{m,k} \) = genotype of k-th locus in m-th subject (total subject = M), \( t_m \) = trait value of m-th subject

Using the given observation, the genetic association between the k-th locus (i.e., k-th SNP) and Trait T is determined as \( P^T = [p_1, \ldots, p_k]^T \). \( P_k \) indicates the original p-value of Fisher’s exact test for the k-th locus. In general, the expected scale of G (i.e., number of SNPs) would be \( 10^{-1} - 10^0 \). When the total number of permutations is N, MPI-GWAS shuffled the matrix G and T to generate background distributions of genetic diversity across N/b MPI ranks (where b = subtasks of shuffling numbers per MPI rank). Using these permuted matrices, each subtask per MPI rank calculates \( P'^r = [p'_1, \ldots, p'_k]^T \) as p-values of random distributions (Fig. 1A, red pseudo-code). Thereafter, each MPI rank adds the result value of function f while looping the number of subtasks, where function f is defined as 1 if \( p'_k < p_k \), otherwise 0; then, the master MPI rank yields the adjusted p-value of the k-th locus (\( p^r_k \)) by reducing the result value of other MPI tasks with the sum function and dividing by N (Fig. 1A, red pseudo-code). Specifically, \( p^r_k < p_k \) indicates that the number of observations where random p-values of the k-th locus is less (i.e., more significant) than the calculated p-value using real observations. Therefore, the adjusted p-value means a probability of type 1 error occurrence for the k-th locus under N-time permuted distributions.

Results

The strong and weak scaling performance of MPI-GWAS is presented in Fig. 1B and 1C, respectively. In summary, the strong scale (Fig. 1B) indicates that \( 10^7 \) permutations of one locus can be calculated in 600 s using 2,720 CPU cores, which is 7.2 times faster than 272 cores. The weak scale (Fig. 1C) indicates that even if the number of permutations per one locus is increased according to the number of computation nodes, it performs well. Two cohorts of actual data were used to verify the performance of MPI-GWAS: the Korean Genome and Epidemiology Study (KoGES) [4] and the UK biobank (UKBB) [5]. The repeated observation of longitudinal traits, such as alteration of blood pressures along traced assessments for decades, is a representative example of the violation of the normal distribution of phenotypes. Thus, we utilized the traced phenotype of type 2 diabetes mellitus (T2DM) in the KoGES and UKBB, respectively. The phenotype of T2DM was measured repeatedly seven times every 2 years in the KoGES. Likewise, the participants of the UKBB have been assessed for the phenotype of T2DM up to three times across 10 years. The adjusted p-values via \( 10^2 \) permutations using the KoGES and UKBB are displayed in Fig. 1D. In the case of the KoGES, covering 31,437 loci per assessment, a total computing time with 171,360 CPU cores was ~4 days using 2,500 nodes (25% of Nuron). With the UKBB data, covering 52,858 loci per assessment, the total elapsed time was similar. The selection of SNPs for KoGES is based on the traced loci using genotype array. To achieve a similar scale of validation, we used a subset of loci from the UKBB data. For the selection of 52,858 loci from the UKBB, we utilized the linkage disequilibrium pruning process via the PLINK. As depicted in Fig. 1D, type 1 errors of p-values were adjusted via large-scale N-permutations. In
conclusion, MPI-GWAS enables us to feasibly compute the permutation-based GWAS within a reasonable time and to harness the power of supercomputing resources.

**Discussion**

The parallel computing of MPI-GWAS solves the computational burden in the permutation approach for GWAS on an acceptable scale. To our best knowledge, MPI-GWAS is the first attempt for an acceleration of GWAS permutation using distributed memory system, including the Nurion. Moreover, the MPI is an interface standard for distributed memory parallelization. Although the partitioned global address space (PGAS) programming model has been suggested as the next step of MPI technology, practical usage of PGAS is still pending. The computing time of MPI-GWAS is mainly depending on the network bandwidth under shared memory system. For instance, the bandwidth of the Nurion is 100 GB/sec. Because we released our source code via GitHub for researchers (https://github.com/hypaik/proj_MPIGWAS), we expect the computing performance can be compared via many users. Howev-
er, MPI-GWAS displays linearly improved performance with additional computing nodes. Because it was confirmed that the weak-scaling performance of MPI-GWAS comes close to ideal, it is predicted that it can perform well on large-scale cluster machines for calculations for more loci.

In addition, the utilized computing infra, Nurion system is a national supercomputing infra. Many computational research including astrophysics, nanoscience and protein docking simulations have been utilized this national supercomputing infra for over decades. The detail of web application for the use of the Nurion system is available at [www.ksc.re.kr](http://www.ksc.re.kr). Because we validated the performance of MPI-GWAS using the Nurion system, researchers can utilize directly to the Nurion system as well as other shared memory systems. In conclusion, our application could contribute broadly to GWAS globally, using diverse machines, including supercomputers.

**ORCID**

Hyojung Paik: https://orcid.org/0000-0002-5803-0670  
Yongseong Cho: https://orcid.org/0000-0001-7732-2524  
Seong Beom Cho: https://orcid.org/0000-0002-5734-1645  
Oh-Kyoung Kwon: https://orcid.org/0000-0001-9734-9257

**Authors’ Contribution**

Conceptualization: HP, YC. Data curation: HP, SBC. Methodology: HP, YC. Model development and validation: YC. Visualization: HP. Project administration: OKK. Writing - original draft: HP, YC. Writing - review & editing: OKK.

**Conflicts of Interest**

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

**Acknowledgments**

The Korea Institute of Science and Technology Information (KISTI) (K-21-L02-C10, K-20-L02-C10-S01, K-21-L02-C10-S01), and the Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (2021M3H9A203052011). HP and YC were also supported by the Ministry of Science and ICT (N-21-NM-CA08-S01). This work was also supported by the National Supercomputing Center with supercomputing resources including technical support (TS-2021-RG-0006).

**References**


Instructions for authors

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

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

**Research and publication ethics**

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

**Statement of Informed Consent**

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

**Statement of Human and Animal Rights**

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

**Authorship**

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

**Corresponding author and first author**

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

**Correction of authorship after publication**

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

**Conflict of Interest Statement**

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

**Originality and Duplicate Publication**

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

**Secondary Publication**

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

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

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

**Preparation of manuscripts**

**General requirement**

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

**Manuscript type**

*Original articles*

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

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

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

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

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

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

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

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

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

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

Main text:
All papers should be divided into the following sections and appear in this order:
(1) Introduction: The paper begins with an introduction without subheadings that reviews the literature and states and justifies the purpose of the research.
(2) Methods: This section should contain sufficient detail so that all procedures can be repeated, in conjunction with the cited references. The manufacturer and model number should be stated in this section—for example, as Sigma Chemical Co. (St. Louis, MO, USA).
(3) Results: This section should describe the results of the experiments. Extensive interpretation should be reserved for the Discussion section. The results should be presented as concisely as possible. Footnotes should not be used and will be transferred to the text. Gene symbols should be italicized; protein products are not italicized.
(4) Discussion: This section should provide an interpretation of the results in relation to previously published work and to the experimental system at hand. The Results and Discussion may be combined.
(5) Acknowledgments: Information concerning the sources of financial support should be included in the acknowledgments.

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

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

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

Examples of references are given below:

Journal article

Books

Book sections

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

Conference paper

Dissertation/Thesis

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

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

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

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

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

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

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

File sizes must be as small as possible, for quick downloading. Recommended specifics are:

• Videos
  - File size: <150 MB
  - Frame rate: 30 frames per second
  - Field order: none (progressive, not interlaced)
  - Aspect ratio: widescreen 16:9
  - Video codec: H.264
  - Video bitrate: 2 Mbps
  - Audio codec: AAC
  - Audio bitrate: 128 kbps

• Images
  - Frame size: 300 dpi in resolution
  - Frame rate: 300 dpi in resolution and 10-15cm in width

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

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

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

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

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

Peer review and revision of manuscripts

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

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

**Copyrights, open access policy and open data policy**

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

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

**Archiving policy**
It is accessible without barrier from Korea Citation Index (https://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-
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 (“learned intermediary”) 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.

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

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

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

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

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

**Contact address**

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

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

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

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

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

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

Statement of Informed Consent

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

Statement of Human and Animal Rights

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

Authorship

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

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

Corresponding author and first author

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

Correction of authorship after publication

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

Conflict of Interest Statement

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

Originality and Duplicate Publication

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

Secondary Publication

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

• The authors have received approval from the editors of both journals (the editor concerned with the secondary publication must have access to the primary version).
• The priority for the primary publication is respected by a publication interval negotiated by editors of both journals and the authors.
• The paper for secondary publication is intended for a different group of readers; an abbreviated version could be sufficient.
• The secondary version faithfully reflects the data and interpretations of the primary version.

• The secondary version informs readers, peers, and documenting agencies that the paper has been published in whole or in part elsewhere—for example, with a note that might read, “This article is based on a study first reported in the [journal title, with full reference]”—and the secondary version cites the primary reference.
• The title of the secondary publication should indicate that it is a secondary publication (complete or abridged republication or translation) of a primary publication. Of note, the United States National Library of Medicine (NLM) does not consider translations to be “republications” and does not cite or index them when the original article was published in a journal that is indexed in MEDLINE.

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


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

☐ 3. Abstract in unstructured format within 250 words.

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

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

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

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

☐ 8. Figures as 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.