Aims and scope

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

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

- Manuscript Editing by InfoLumi Co., Seongnam, Korea.
- E-submission system by Inforang, Seoul, Korea
- PDF layout, XML production, and homepage management by M2Community Co., Seoul, Korea

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

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

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

© Copyright 2021, the Korea Genome Organization
© It is an open access journal. The articles are distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Editorial

Editor's introduction to this issue (G&I 19:1, 2021)
Taesung Park

Original articles

Elucidating molecular mechanisms of acquired resistance to BRAF inhibitors in melanoma using a microfluidic device and deep sequencing
Jiyeon Han, Yeonjoo Jung, Yukyung Jun, Sungsu Park, Sanghyuk Lee

Functional annotation of lung cancer–associated genetic variants by cell type–specific epigenome and long–range chromatin interactome
Andrew J. Lee, Inkyung Jung

Identification of potential candidate genes for lip and oral cavity cancer using network analysis
Sarmilah Mathavan, Chin Siang Kue, Suresh Kumar

Genetic alterations in Wnt family of genes and their putative association with head and neck squamous cell carcinoma
Aditya Jain, Smiline Girija A. S., Paramasivam A., Vijayashree Priyadharsini J.

Screening of novel alkaloid inhibitors for vascular endothelial growth factor in cancer cells: an integrated computational approach

In silico discovery and evaluation of phytochemicals binding mechanism against human catechol-O-methyltransferase as a putative bioenhancer of L-DOPA therapy in Parkinson disease
Surya Narayan Rath, Lingaraja Jena, Rajabrata Bhuyan, Nimai Charan Mahanandia, Manorama Patri

De novo assembly, annotation and gene expression profiles of gonads of Cytorace–3, a hybrid lineage of Drosophila nasuta nasuta and D. n. albomicans
Koushik Ponnanna, Stafny M. DSouza, Nallur B. Ramachandra

Computational evaluation of interactions between olfactory receptor OR2W1 and its ligands
S. June Oh

Machine learning based anti-cancer drug response prediction and search for predictor genes using cancer cell line gene expression
Kexin Qiu, JoongHo Lee, HanByeol Kim, Seokhyun Yoon, Keunsoo Kang

Forecasting of the COVID-19 pandemic situation of Korea
Taewan Goo, Catherine Apiio, Gyujin Heo, Doeun Lee, Jong Hyeok Lee, Jisun Lim, Kyulhee Han, Taesung Park
In this issue, there are 10 Original Articles, of which five are related to cancer research. The first original article, by the group of Han et al. (Ewha Womans University, Korea), focused on elucidating the molecular mechanisms of acquired resistance to BRAF inhibitors in melanoma. A microfluidic chip with a concentration gradient of vemurafenib was utilized to rapidly obtain therapy-resistant clones from two melanoma cell lines with the \textit{BRAF}^{V600E} mutation. Exome and transcriptome data were produced from 13 resistant clones. This study provides an omics-based comprehensive overview of the molecular mechanisms governing acquired resistance to BRAF inhibitor therapy. The second article, by Lee and Jung (KAIST, Korea), reported functional annotation of lung cancer-associated genetic variants based on eight major cell types of human lung tissue. This work showed that approximately 22% of lung cancer-associated risk variants were linked to noncoding regulatory elements. Through integrative analysis of high-resolution long-range chromatin interactome maps and single-cell RNA-sequencing data, the authors uncovered a number of putative target genes of these variants and functionally relevant cell types, which expands the scope of functional annotation of lung cancer-associated genetic risk factors.

The third article, by Mathavan et al. (Management and Science University, Malaysia), identified potential candidate genes for lip and oral cavity cancer using network analysis. Using the DisGeNET database and STRING database, the authors identified several hub genes, such as \textit{VEGFA}, \textit{IL6}, \textit{MAPK3}, \textit{INS}, \textit{TNF}, \textit{MAPK8}, \textit{MMP9}, \textit{CXCL8}, \textit{EGF}, and \textit{PTGS2}, which could provide a new understanding of the underlying molecular mechanisms of lip and oral cavity cancer. The fourth article, by Jain et al. (Saveetha Institute of Medical and Technical Sciences, India), reported genetic alterations in the \textit{WNT} family of genes and their putative association with head and neck squamous cell carcinoma. The \textit{WNT} signaling pathway is known to be involved in crucial mechanisms for cellular maintenance and development. The authors reported a marked difference in the gene expression profile of \textit{WNT11} between grades and when compared with normal samples. The fifth article, by Shahik et al. (AFC Agro Biotech Ltd., Bangladesh), presented the results of screening alkaloid inhibitors for vascular endothelial growth factor (VEGF) in cancer cells. Through an integrated computational approach, the authors proposed five alkaloid candidates for inhibiting VEGF and \textit{VEGFR} receptor-mediated angiogenesis, which can be used as novel lead compounds to design new and effective drugs against cancer.

The sixth article, by Rath et al. (Odisha University of Agriculture and Technology, India), investigated the in silico discovery and evaluation of phytochemicals of the plant \textit{Withania somnifera}. As a putative bioenhancer of levodopa therapy in Parkinson disease, the authors reported nine phytochemicals that had strong binding efficiency against human catechol-O-methyltransferase in comparison to the inhibitory drugs opicapone and entacapone. The seventh article, by Ponnanna et al. (University of Mysore, India) reported de novo assembly, annotation, and gene expression profiles of gonads of Cytorace-3, a
hybrid lineage of *Drosophila nasuta nasuta* and *D. n. albomicans*. This study provided an overview of the expression divergence and inheritance patterns of transcriptomes in an independently evolving distinct hybrid lineage of *Drosophila*. The next article, by Oh (Inje University College of Medicine, Korea), performed a computational evaluation of interactions between olfactory receptor olfactory receptor 2W1 (OR2W1) and its ligands. Through modeling the interaction between an olfactory receptor and its ligands at the molecular level, the author successfully demonstrated the modes of ligands binding to the three-dimensional (3D) model of olfactory receptor OR2W1 and showed a statistically significant difference in the binding affinity to the olfactory receptor between the agonist and the antagonist.

The last two articles are about machine learning and statistical modeling. The article by Qiu et al. (Dankook University, Korea) established machine learning-based prediction models of anti-cancer drug response using cancer cell line gene expression and drug response data. Several statistical methods, such as Pearson correlation analysis and an ElasticNet regression model, were performed to find the model with the best performance. The last article, by Goo et al. (Seoul National University, Korea), is about predicting the future spread of coronavirus disease 2019 (COVID-19) in Korea. Using five mathematical, machine learning, and statistical models, the authors predicted the daily number of COVID-19 confirmed cases. Through comparative studies, the authors showed that machine learning models, a standard susceptible exposed infected recovered model, and a non-linear model tended to provide accurate predictions. These prediction results are expected to help in the pandemic response by informing decisions about planning, resource allocation, and decisions concerning social distancing policies.

**ORCID**

Taesung Park: https://orcid.org/0000-0002-8294-590X
Elucidating molecular mechanisms of acquired resistance to BRAF inhibitors in melanoma using a microfluidic device and deep sequencing

Jiyeon Han¹, Yeonjoo Jung², Yukyung Jun³, Sungsu Park⁴, Sanghyuk Lee¹,²∗

¹Department of Bio-information Science, Ewha Womans University, Seoul 03760, Korea
²Ewha Research Center for Systems Biology (ERCSB), Ewha Womans University, Seoul 03760, Korea
³Center for Supercomputing Application, Division of National Supercomputing, Korea Institute of Science and Technology Information, Daejeon 34141, Korea
⁴School of Mechanical Engineering, Sungkyunkwan University, Suwon 16419, Korea

BRAF inhibitors (e.g., vemurafenib) are widely used to treat metastatic melanoma with the BRAF V600E mutation. The initial response is often dramatic, but treatment resistance leads to disease progression in the majority of cases. Although secondary mutations in the mitogen-activated protein kinase signaling pathway are known to be responsible for this phenomenon, the molecular mechanisms governing acquired resistance are not known in more than half of patients. Here we report a genome- and transcriptome-wide study investigating the molecular mechanisms of acquired resistance to BRAF inhibitors. A microfluidic chip with a concentration gradient of vemurafenib was utilized to rapidly obtain therapy-resistant clones from two melanoma cell lines with the BRAF V600E mutation (A375 and SK-MEL-28). Exome and transcriptome data were produced from 13 resistant clones and analyzed to identify secondary mutations and gene expression changes. Various mechanisms, including phenotype switching and metabolic reprogramming, have been determined to contribute to resistance development differently for each clone. The roles of microphthalmia-associated transcription factor, the master transcription factor in melanocyte differentiation/dedifferentiation, were highlighted in terms of phenotype switching. Our study provides an omics-based comprehensive overview of the molecular mechanisms governing acquired resistance to BRAF inhibitor therapy.

Keywords: cancer drug resistance, melanoma, microfluidic device, RNA sequencing, targeted therapy, whole exome sequencing

Introduction

Melanoma is a malignant skin cancer that is primarily caused by excessive exposure to UV radiation from sunlight. Although melanoma represents a small proportion of skin cancer, its metastatic form is fatal, with a 5-year survival rate between 5%–19% [1]. In 2020, more than 100,000 new melanoma patients and 6,850 deaths from melanoma are expected in the United States [2]. Melanoma has been the focus of modern genomic studies and cancer therapeutics since the development of targeted cancer drugs and immunotherapies.
**BRAF** mutation occurs in more than 80% of melanoma patients, with the V600E mutation being the most frequently observed [3]. It is also responsible for ~40% of papillary thyroid carcinoma and small portion of other tumors (e.g., colon, pancreas, brain, lung, etc.), taking ~8% of human tumors in total [4,5]. Vemurafenib, targeting **BRAF** V600 alterations, is among the most well-known cancer drugs with rapid and dramatic early responses, but the tumor eventually relapses in most cases [6]. Thus, overcoming resistance to targeted therapy is of prime importance and would require a detailed understanding of the molecular mechanisms underlying resistance development. Intrinsic tumor heterogeneity and the evolution of cancer cells are the major causes of therapy resistance [7].

Molecular mechanisms of resistance to **BRAF** inhibitor (**BRAFi**) have been reported by analyzing the genome and transcriptome data from patient samples [8,9]. Reactivation of the mitogen-activated protein kinase (**MAPK**) pathway by secondary mutations in the RAS/RAF/MEK/ERK signaling cascade is the most frequently observed mechanism, occurring in up to 80% of **BRAFi**-resistant tumors [10]. Many additional mechanisms, however, have been found to contribute to the development of therapy resistance, including activation of the phosphoinositide 3-kinase (PI3K)–AKT–mammalian target of rapamycin (mTOR) pathway, tumor microenvironment reprogramming by the Hippo signaling pathway, and phenotype switching by master transcription factors (TFs), such as microphthalmia-associated transcription factor (MITF), and receptor tyrosine kinase (RTK), such as AXL [11].

Although patient tumor samples are highly useful for their clinical relevance, it is difficult to study the details of molecular mechanisms because of various issues in sample quality and quantity. Therefore, patient-derived cell lines are important for studying the complex interplay among various mechanisms. In the case of cell line studies for drug resistance, the most difficult and labor-intensive part usually is obtaining resistant cells. Previously, we demonstrated that a microfluidic chip with a concentration gradient of cancer drugs could induce drug resistance rapidly, thereby being designated as the cancer drug resistance accelerator (CDRA) chip, and that multiple molecular mechanisms underlying erlotinib resistance could be elucidated by analyzing exome and transcriptome sequencing data [12]. In this study, we applied the same principle to investigate the mechanisms governing the resistance of melanoma cells to the **BRAFi** vemurafenib. The aims of our study were not only to identify resistance mechanisms but also to investigate whether different clones or cell lines acquire drug resistance in different ways, i.e., whether the resistance acquisition process is stochastic.

### Methods

**Cell culture and establishment of vemurafenib-resistant cells**

Human melanoma cell lines A375 and SK-MEL-28 were purchased from the American Type Culture Collection (Manassas, VA, USA) and were maintained in Dulbecco’s modified Eagle’s medium (for A375) or minimum essential medium (for SK-MEL-28) supplemented with 10% fetal bovine serum (HyClone, Logan, UT, USA), 100 units/mL penicillin (Invitrogen, Carlsbad, CA, USA) and 100 μg/mL streptomycin (Invitrogen).

Vemurafenib-resistant A375 and SK-MEL-28 cells were established using a microfluidic chip as described previously [12]. Briefly, the interior surface of the chip was sanitized with 70% ethanol and coated with 10 μg/mL fibronectin (Sigma, St. Louis, MO, USA). The cells were seeded carefully onto the surface of a chip and incubated to adhere to the surface. One inlet reservoir was filled with serum-free media containing vemurafenib, and the other inlet reservoir was filled with serum-free media only (i.e., without vemurafenib). Two outlet reservoirs were filled with serum-free media. The reservoirs were replaced every day with freshly prepared serum-free media with or without vemurafenib. After cultivation in the chip, the cells were trypsinized, collected, and transferred to a new culture dish and cultured to obtain enough cells. To examine the effect of exposure to increasing concentrations (0.0001 μM to 10 μM) of vemurafenib (Selleck Chemicals, Houston, TX, USA), cell proliferation was measured at 72 hours using an EZ-Cytox Cell Viability Assay Kit (Daeillab Services, Seoul, Korea). Cells were plated at densities of 4,000 cells or 5,000 cells per well into 96-well plates by hexa-repeat. The results are expressed as a percentage of the cell number in drug-untreated control wells. The IC_{50} values for vemurafenib were calculated by fitting the plot of percentage inhibition versus the log of drug concentration with the nonlinear regression method in GraphPad Prism 6 (GraphPad Inc., La Jolla, CA, USA) software. Error bars represent the standard error of the mean.

**Production and processing of exome and transcriptome sequencing data**

Total genomic DNA was extracted from control cells and vemurafenib-resistant cells by the traditional phenol extraction method. One microgram of genomic DNA was used for exome sequencing. Total RNA was extracted from individual conditions using an RNeasy Mini Kit according to the manufacturer’s protocol (Qiagen, Hilden, Germany), and 1 μg of total RNA was used for RNA sequencing. The exome-seq and RNA-seq data acquisition process have been described previously [12].
Sequencing reads were trimmed with Sickle (ver. 1.33) to remove low-quality reads and adaptor sequences [13]. For the RNA sequencing data, trimmed reads were mapped to the reference human genome (UCSC hg19) using STAR (ver. 2.6.1c) [14], and the gene expression values were quantified with RSEM (ver. 1.3.3) [15] with the default options. Exome sequencing data were mapped with the BWA-MEM alignment tool (ver. 0.7.17) [16] and indexed with Samtools (ver. 1.11) [17]. Strelka2 (ver. 2.9.10) [18], as well as MuTect2 [19], in the Genome Analysis Toolkit (GATK ver. 4.1.8.1) were used for calling somatic variants (single nucleotide variations and indels). We also calculated copy number variations (CNVs) from exome sequencing data using EXCAVATOR2 (ver. 1.1.2) [20]. All resulting variants were annotated with ANNOVAR software (ver. 2019Oct24) [21].

Transcriptome data analysis
Most of our analysis was performed with R (ver. 4.0.0) and several R-based packages. To obtain subgroups from the samples, variable genes were selected within the top 20% in the coefficient of variation (COV20). Hierarchical clustering with these COV20 genes yielded three sample groups, named SK-MEL-28, A375_G1, and A375_G2, according to the cellular origin. Differentially expressed genes were obtained by comparison with the control sample using the edgeR package (ver. 3.30.3) with false discovery rate (FDR) < 0.05, absolute log2FoldChange > 1, and logCPM > 1 [22]. We used the GSVA program (ver. 1.36.3) [23] with the Kyoto Encyclopedia of Genes and Genomes (KEGG) subset of canonical pathways from MSigDB (c2.cp.kegg.v7.2.symbols.gmt) [24] to calculate the gene set activity for each sample. Gene sets with variable activities among three groups were obtained with a t-test supported in R base functions under the threshold of FDR < 0.01.

Public transcriptome data of melanoma patients
We searched the Gene Expression Omnibus (GEO) for transcriptome data before and after vemurafenib treatment in melanoma patients. We identified such patients from three GEO records with the accession numbers GSE141484 [15], GSE99898 [27]. Normalized expression data were downloaded and merged with quantile normalization to reduce the batch effect. The fold change in expression was used for heatmap visualization and calculating pathway activities for each sample using the GSVA program. The correlation with the MITF expression pattern was calculated with MITF and its regulators, RTK genes, and the pathway activities of KEGG pathways. The genes below the correlation coefficient of 0.5 were filtered out. We also performed a principal component analysis using genes highly correlated with MITF expression. The R based package factoextra (ver. 1.0.7) was employed to visualize the principal component analysis (PCA) plot of our samples and public patient data.

Results

Acquisition of vemurafenib-resistant cells using a microfluidic CDRA chip
We cultured two melanoma cell lines with the BRAF V600E mutation, A375 and SK-MEL-28 specifically, for 8–47 days on CDRA chips with a concentration gradient of vemurafenib (Fig. 1A). The cultured product was further maintained in dishes with a relatively high dose of vemurafenib to select vemurafenib-resistant cells. We obtained 13 samples, specifically four from SK-MEL-28 and nine from A375 cell lines, and confirmed that those samples were indeed resistant to vemurafenib (Fig. 1B). The extent of resistance, however, was notably different between the two cell lines. The average IC50 values of resistant cells from SK-MEL-28 cells increased by 23.7-fold (from 0.143 to 3.3815 μM), whereas those from A375 cells increased by 6.3-fold (from 0.0655 to 0.4133 μM) (Table 1). Thus, the characteristics of drug resistance were highly dependent on the original cellular identity.

Transcriptome data show different mechanisms of acquired resistance in the two cell lines
In an effort to identify the molecular mechanisms of drug-induced resistance, we performed exome and transcriptome sequencing for two original cell lines and 13 resistant cells obtained from CDRA chips. Exome sequencing data identified only one mutation with known driver potential (HRAS Q61K) in one of the A375-derived resistant cells (Supplementary Fig. 1). Copy number variation profiles were mild in all cases. Thus, we focused on the interpretation of transcriptome data.

Hierarchical clustering of transcriptome data revealed that our resistant cells could be divided into two groups according to the source cell line (Supplementary Fig. 2). Resistant cells from A375 were further divided into two subgroups: A375_G1 and A375_G2. Next, we examined the gene expression of pathway marker genes that are known to be associated with vemurafenib resistance, including the RTK pathway [28], transforming growth factor β (TGF-β) pathway [29], MITF regulation [30], Sonic hedgehog pathway [31], MAPK pathway [32–34], and PI3K-AKT-mTOR pathway [11,35] (Fig. 2). The expression levels of most of these known factors were nearly opposite between the two cell lines. The expression pattern was mild in all cases. Thus, we focused on the interpretation of transcriptome data.

Most marker genes showed a dichotomous expression pattern between the SK-MEL-28-derived cells and A375-derived cells, strongly suggesting that the resistance mechanisms are notably dif-

https://doi.org/10.5808/gi.20074
Table 1. Experimental conditions

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Chip-Conc. (μM)</th>
<th>Duration in chip (day)</th>
<th>Dish culture</th>
<th>IC&lt;sub&gt;50&lt;/sub&gt; (μM)</th>
<th>FC to WT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A375 cell line</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A375-321s-2</td>
<td>25</td>
<td>13</td>
<td>70 nM (3 wk) 350 nM (1 wk) 3.5 μM (1 wk)</td>
<td>0.0655</td>
<td>1.0</td>
</tr>
<tr>
<td>A375-321s-3</td>
<td>25</td>
<td>13</td>
<td>70 nM (3 wk) 350 nM (1 wk) 3.5 μM (1 wk)</td>
<td>0.3623</td>
<td>5.5</td>
</tr>
<tr>
<td>A375-321s-4</td>
<td>25</td>
<td>13</td>
<td>70 nM (3 wk) 350 nM (1 wk) 3.5 μM (1 wk)</td>
<td>0.2412</td>
<td>3.7</td>
</tr>
<tr>
<td>A375-328s-1</td>
<td>10</td>
<td>8</td>
<td>70 nM (3 wk) 350 nM (1 wk) 3.5 μM (1 wk)</td>
<td>0.3329</td>
<td>9.8</td>
</tr>
<tr>
<td>A375-328s-2</td>
<td>10</td>
<td>8</td>
<td>70 nM (3 wk) 350 nM (1 wk) 3.5 μM (1 wk)</td>
<td>0.5420</td>
<td>8.3</td>
</tr>
<tr>
<td>A375-328s-3</td>
<td>10</td>
<td>8</td>
<td>70 nM (3 wk) 350 nM (1 wk) 3.5 μM (1 wk)</td>
<td>0.5076</td>
<td>7.8</td>
</tr>
<tr>
<td>A375-328s-4</td>
<td>10</td>
<td>8</td>
<td>70 nM (3 wk) 350 nM (1 wk) 3.5 μM (1 wk)</td>
<td>0.3190</td>
<td>5.1</td>
</tr>
<tr>
<td>A375-503s-1</td>
<td>10</td>
<td>47</td>
<td>3.5 μM (3 wk) 0.15 μM (6 wk) 1.5 μM (3 wk)</td>
<td>0.1886</td>
<td>2.9</td>
</tr>
<tr>
<td>A375-503s-2</td>
<td>10</td>
<td>47</td>
<td>3.5 μM (3 wk) 0.15 μM (6 wk) 1.5 μM (3 wk)</td>
<td>0.5874</td>
<td>9.0</td>
</tr>
<tr>
<td>SK-MEL-28 cell line</td>
<td></td>
<td></td>
<td></td>
<td>0.143</td>
<td>1.0</td>
</tr>
<tr>
<td>SK-MEL-28-1</td>
<td>4.5</td>
<td>12</td>
<td>0.15 μM (6 wk) 1.5 μM (3 wk)</td>
<td>5.617</td>
<td>39.3</td>
</tr>
<tr>
<td>SK-MEL-28-2</td>
<td>4.5</td>
<td>12</td>
<td>0.15 μM (6 wk) 1.5 μM (3 wk)</td>
<td>4.577</td>
<td>32.1</td>
</tr>
<tr>
<td>SK-MEL-28-3</td>
<td>7.5</td>
<td>12</td>
<td>0.15 μM (6 wk) 1.5 μM (3 wk)</td>
<td>1.966</td>
<td>13.8</td>
</tr>
<tr>
<td>SK-MEL-28-4</td>
<td>7.5</td>
<td>12</td>
<td>0.15 μM (6 wk) 1.5 μM (3 wk)</td>
<td>1.366</td>
<td>9.6</td>
</tr>
</tbody>
</table>

FC, fold change; WT, wild type.

Fig. 1. Acquisition of drug-resistant cells using microfluidic chips. (A) Schematic diagram of microfluidic chips and the experimental setup. (B) Cell viability plots to confirm drug resistance (i.e., increased IC<sub>50</sub> values) of cells obtained from microfluidic chips.
Fig. 2. Expression of key genes in drug-resistant A375 and SK-MEL-28 cells. Expression values are the log2FoldChange values converted into the row-wise Z-score. RTK, receptor tyrosine kinase; TGFB, transforming growth factor β; MITF, microphthalmia-associated transcription factor; SHH, smoothened signaling pathway or sonic hedgehog signaling; MAPK, mitogen-activated protein kinase; PI3K, phosphoinositide 3-kinase; mTOR, mammalian target of rapamycin.

Different in the two cell lines. For example, MITF and its regulators were downregulated in all SK-MEL-28–derived cells, while they were upregulated in all A375-derived cells. We also obtained a few markers (e.g., LEF1, CDK1, E2F1, MAP2K1, MAPK3, AKT3, and MTOR) showing differences between A375_G1 and A375_G2 cells. Thus, the resistance mechanisms may vary even within the same cell origin, implicating the stochastic nature of the resistance acquisition process.

To further elucidate resistance mechanisms at the pathway level, we calculated the pathway activities for KEGG curated pathways from MSigDB using the GSVA algorithm (Supplementary Table 1). The pathway activity pattern again showed the same three groups, and differential pathways were obtained by two t-test comparisons of (1) SK-MEL-28–derived cells and A375-derived cells and (2) A375_G1 cells and A375_G2 cells (Fig. 3).

A number of pathways appeared different between the two cells. SK-MEL-28–derived cells had downregulated metabolism, upregulated signaling (MAPK, WNT, and Hedgehog pathways), and elevated proliferation (cell cycle, DNA replication, and mismatch repair). A375-derived cells mostly exhibited the opposite expression pattern. The two subgroups of A375 cells had differential activities in mTOR and Hedgehog signaling pathways and RNA homeostasis (basal transcription, spliceosome, and RNA degradation).

Comparison with clinical samples highlights the roles of MITF and accompanying pathways

Comparison with clinical sample data is critical to in vitro experiments, such as the CDRA chip. We collected public expression data of melanoma patients who received vemurafenib treatment because of the V600 mutation and whose transcriptome data were available before and after vemurafenib treatment. We identified 19 such pairs from three studies, enabling multiple pairs to be exam-
The regulatory markers in Fig. 2 can be merged into a unifying model where RTK-mediated MAPK and PI3K-AKT pathways affect the expression and nuclear export of MITF [36]. Since MITF is the master regulator of melanocyte proliferation and differentiation as well, we searched for regulatory genes and/or pathways that would play roles within the context of MITF regulation. Candidate genes were selected among the RTK genes and known TFs of MITF [30,37-39]. Then, correlation coefficients with MITF expression were obtained for gene expression or pathway activities using the patient sample pairs (Fig. 4A). Positively correlated genes included MYC, HIF1A, RYK, KIT, and LEF1, whereas negatively correlated genes were AATK, NFKB2, ROR2, CREB3, STYK1, FGFR2, and AXL. We also examined the correlation of pathway activities for KEGG pathways. We used the pathways differentially scored in our data to patient data. As a result, the MAPK pathway and hedgehog pathway showed a high negative correlation tendency, whereas several metabolic pathways had a strong positive correlation. Additionally, cell adhesion-related pathways, such as gap junctions and extracellular matrix (ECM) receptor interactions, showed slight negative correlations.

Since our resistant clones from cell lines and patient samples exhibited dichotomous patterns of MITF expression, we performed a PCA of samples using regulator genes. The PCA plot again showed that the two cell lines formed their own clusters, with patient samples scattered more dispersely but with a specific association with each cell line (Fig. 4B). Thus, our results with cell lines may represent two different classes of patient samples in the mo-

Fig. 3. Pathway activities for Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The heatmap shows pathway activities for 68 KEGG pathways obtained from the differential test with false discovery rate < 0.05. The full list is shown in Supplementary Table 1 with detailed numbers. mTOR, mammalian target of rapamycin; MAPK, mitogen-activated protein kinase; ECM, extracellular matrix.
Fig. 4. Gene expression and pathway activity from patient samples. (A) Microarray expression data are in log2FoldChange. The expression and pathway activity heatmaps are shown at the top and bottom, respectively, in different colors. Note that the genes and pathways are sorted according to the correlation coefficient with microphthalmia-associated transcription factor (MITF) expression. (B) Principal component analysis plot of both patient samples (circles) and cell line-derived resistant cells (triangles) using 20 correlated genes in A. All expression values are in log2FoldChange.
Discussion

Deciphering the molecular mechanisms of resistance is an important step in the development of new treatment methods in targeted cancer therapy. In this study, we adopted a microfluidic chip to rapidly induce drug resistance and applied whole exome and transcriptome sequencing to investigate the molecular mechanisms governing resistance to vemurafenib. Chemo-resistant cell lines are usually acquired by exposing cells to stepwise increasing concentration of chemo-agents, which is labor-intensive and time-consuming. Notably, we obtained resistant cells in 1–7 weeks using the CDRA chips in contrast to several months by conventional methods.

In the case of vemurafenib resistance, genomic variations, such as mutation and CNVs, were relatively mild, probably due to a relatively short time of drug exposure compared to the actual patients. On the other hand, the transcriptome signatures demonstrated that the two cell lines acquired vemurafenib resistance in different ways that are highly contrasted in MITF expression.

MITF is the master TF of melanocytes responsible for regulating the proliferation, differentiation, and metabolic reprogramming of melanocytes. BRAFi treatment causes melanoma cells to change their MITF expression to low or high levels, both of which lead to slowly proliferating resistant cells. AXL, an important regulator of apoptosis and epithelial-mesenchymal transition (EMT), is usually inversely correlated with MITF expression. Although MITF and AXL have been characterized as two primary regulators of cellular phenotype switching, many other pathways are also relevant to MITF regulation, including the MAPK, Hippo, TGF-β, and autophagy signaling pathways. Our pathway activity data showed that many of these pathways were coordinately associated with MITF expression in both cell lines and patient tumors, but it was difficult to identify consistent behavior. This finding implies that MITF regulation is highly complex and dependent on cellular contexts, such as the tumor microenvironment.

We also observed that A375-derived resistant cells were divided into two subgroups with differential activities in the cell cycle and ECM interactions. These subgroups further highlight the heterogeneity of acquired resistance. Further work, probably based on the systems biology discipline, is necessary to elucidate the roles played by stochastic factors responsible for subgrouping.

Among the regulators and pathways implicated, the MAPK pathway and hedgehog pathways were negatively correlated with MITF expression in patients. Hedgehog signaling is known as the master regulator of EMT [40]. AXL and ECM-related pathways further support their role in phenotype transition to induce drug resistance.

Another important factor in phenotype switching is metabolic rewiring. Our data showed that some of the metabolic pathways that were differentially expressed between the A375-derived samples and SK-MEL-28-derived samples were also positively correlated with MITF expression in the patient data. This result is consistent with the findings of a previous report, which asserted that metabolic distractions are the main driver of drug resistance [41,42].

In conclusion, resistance development to vemurafenib treatment is a complex process in which various factors and pathways are involved, including cellular differentiation and dedifferentiation, the tumor microenvironment, and metabolic reprogramming. Further studies to investigate the interplay of these factors and pathways with the master regulator MITF may facilitate the development of new therapies to overcome drug resistance problems in melanoma.
Supplementary Materials

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

References


https://doi.org/10.5808/gi.20074
Functional annotation of lung cancer-associated genetic variants by cell type-specific epigenome and long-range chromatin interactome

Andrew J. Lee, Inkyung Jung*

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

Functional interpretation of noncoding genetic variants associated with complex human diseases and traits remains a challenge. In an effort to enhance our understanding of common germline variants associated with lung cancer, we categorize regulatory elements based on eight major cell types of human lung tissue. Our results show that 21.68% of lung cancer–associated risk variants are linked to noncoding regulatory elements, nearly half of which are cell type–specific. Integrative analysis of high-resolution long-range chromatin interactome maps and single-cell RNA-sequencing data of lung tumors uncovers number of putative target genes of these variants and functionally relevant cell types, which display a potential biological link to cancer susceptibility. The present study greatly expands the scope of functional annotation of lung cancer–associated genetic risk factors and dictates probable cell types involved in lung carcinogenesis.

Keywords: 3D chromatin interaction, cis-regulatory element, genome-wide association study, lung cancer, single-cell RNA sequencing, single nucleotide polymorphism

Introduction

Gene regulation is a critical biological process that determines cellular identity and function. Systematic investigation of chromatin architecture has shown that the spatiotemporal gene regulation process is tightly controlled by cis-regulatory elements (cREs), which modulate the activities of spatially distant promoters [1-4]. Previous studies have reported that cREs are highly dynamic genomic entities whose dysregulation is associated with various human disorders, including genetic and complex diseases [5-7]. This is well described in the 2019 professional release of the human gene mutation database (HGMD), where more than 4,500 disease-associated mutations are in regulatory sequences [8]. The distal target genes of cREs harboring genetic mutations have been identified as causal elements in human diseases. The representative example is polydactyly syndrome, a congenital limb malformation that results from point mutations in a Shh regulatory element [9]. However, the identification of such long-range regulation is difficult since cREs may regulate genes located beyond large genomic distances. In light of this, Hi-C is a novel technique that enables the investigation of genome-wide, all-to-all chromatin interactions and has substantially improved our view of cREs on long-range gene expression control through 3D chromatin organization [10-14].

The pathogenesis of complex diseases is an outcome of a heterogeneous cell population and various causal genetic variants. The causal genetic variants and the functional
cell type in which these disease-associated variants may be active are often unclear. Although the recent development of single-cell RNA sequencing (scRNA-seq) technology has allowed the assessment of the cell type-specific transcriptome, gene regulation that underlies the functional properties of complex diseases is not well understood. Therefore, we performed a comprehensive analysis of cell type-specific gene regulatory mechanisms by integrating publicly available data on cell type-specific epigenome, single-cell transcriptome, and 3D chromatin interactome surrounding human lung cancer. We explored the dynamics of epigenomic profiles of eight major cell types that exist in the tumor environment and the relationships of cell type-specific regulatory elements with genetic risk variants. By integrating high-resolution chromatin contact maps and scRNA-seq data, we characterize cell type dependent expression profile of putative interaction target genes of lung cancer-associated variant-harboring cREs, and identify a list of potential candidate genes associated with lung carcinoma and their relevant cell type.

Methods

Epigenome profiling of regulatory genome elements

Nineteen H3K27ac chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) datasets were downloaded from the Encyclopedia of DNA Elements (ENCODE) database, representing seven cell populations: epithelial (2 primary epithelial cells from mammary gland, 1 primary epithelial cell from prostate), fibroblast (2 primary fibroblast cells from lung and IMR90 cell line), endothelial (2 primary cells from umbilical vein, 1 primary cell from brain microvasculature), T lymphocyte (1 primary T cell, 1 naïve thymus-derived CD4-positive primary cell, and 1 CD8-positive primary cell), natural killer (NK) cell (1 natural killer primary cell), B lymphocyte (2 primary B cells, and GM12989 cell line), myeloid (3 CD14-positive primary monocyte cells) [2,3]. H3K27ac ChIP-seq data for 3 lung cancer cell lines (A549, A427, and H322) were downloaded from the DBTSS database [15]. H3K27ac ChIP-seq data for two alveolar lung epithelial cells representing squamous type 1 (AT1) cells and cuboidal type 2 (AT2) cells were downloaded from NCBI Gene Expression Omnibus (GEO) database (accession code: GSE84273) [16]. Detailed sample information, biosample ID, and library ID for raw ChIP-seq data are described in Supplementary Table 2. The ChIP-seq reads were aligned against the human genome (hg19) using BWA-mem with default parameters. Non-uniquely mapped, low-quality (MAPQ < 10), and PCR duplicate reads were removed. Peak calling of individual ChIP-seq experiments was performed with MACS2 callpeak with a p-value threshold of 1e-5 and with a respective input control used as the background [17]. cREs were obtained by selecting H3K27ac peaks located distal to promoters (> 2.5 kb from transcription start site [TSS]), merging the peaks across the samples, and stitching peaks within a 3 kb distance. For the quantification of cRE activity, reads per million (RPM) values were calculated and quantile normalized across the samples for comparative analysis.

Collection of genome-wide association study‒single nucleotide polymorphisms associated with lung cancer

A total of 286 genome-wide association study (GWAS)‒single nucleotide polymorphism (SNPs) related to lung cancer were obtained from the NHGRI-EBI GWAS catalog (downloaded on August 20, 2019) [18], targeting nine traits as follows: familial lung adenocarcinoma, familial lung cancer, familial squamous cell lung carcinoma, non-small cell lung cancer, non-small cell lung cancer (recurrence rate), non-small cell lung cancer (survival), small cell lung cancer (survival), small cell lung carcinoma, and squamous cell lung carcinoma. We expanded these tag SNPs by using linkage disequilibrium (LD) information ($r^2 > 0.8$). The LD scores were calculated using PLINK for five ethnic populations obtained from 1000 Genomes Phase 3 data. Tight LD associations ($r^2 > 0.8$) recurrent in at least three ethnic groups were used for LD expansion. The number of total LD-expanded SNPs was 2,128, and these SNPs were stored in a manner that each of them was traceable back to its parental tag SNP.

Mapping 3D long-range chromatin interactions

To obtain high-resolution chromatin contact maps in lung tissue, we downloaded in situ Hi-C data for A549 (lung carcinoma cell line), IMR90 (lung fibroblast cell line), GM12878 (lymphoblastoid cell line), and HMEC (human mammary epithelial cell line) cells from the ENCODE database [2,3]. Detailed sample information, biosample ID, and library ID for the raw in situ HiC data are described in Supplementary Table 2. Raw Hi-C sequenced reads were mapped to the human reference genome (hg19) using BWA-mem (-M option). An in-house script was used to remove low-quality reads (MAPQ < 10), the reads that span ligation sites, chimeric reads, and self-interacting reads in which two fragments are located within 15 kb. The read pairs were merged together as paired-end aligned BAM files, and PCR duplicates were removed with Picard. Statistically significant contacts in Hi-C data were identified at a 5 kb resolution using FitHi-C [19]. We used the default Fit-Hi-C code to calculate the Q-value for each bin pair within a 1 Mb genomic window. A Q-value threshold (Q < 0.01) was used to define significant chromatin contacts.
Identification of cell type–specific gene expression using scRNA-seq data

The raw UMI count matrix of scRNA-seq data for lung adenocarcinoma was downloaded from NCBI GEO with accession code GSE131907 [20]. Data covering 11 tumors and 11 distant normal lungs were selected and processed by using the Seurat R package v3.2.2 [21]. We discarded cells that expressed < 200 genes. To exclude low-quality cells from our data, we filtered out the cells that expressed mitochondrial genes in > 20% of their total gene expression. In each cell, the gene expression was normalized on the basis of the total read count and log transformed. To align the cells originating from different samples, 5,000 highly variable genes from each sample were identified by the vst method. We found anchors and aligned the samples based on the top 10 canonical correlation vectors. The aligned samples were scaled, and principal component analysis was conducted. Then, the cells were clustered by unsupervised clustering (0.5 resolution) and visualized by tSNE using the top 40 principal components (PCs). Known markers were used to assign each subcluster to a corresponding cell population: EP CAM and KRT19 for epithelial cells, DCN and COL1A1 for fibroblasts, PECAM1 and CLDN5 for endothelial cells, CD3D and TRAC for T lymphocytes, NKG7 and GNLY for NK cells, CD79A and IGHM for B lymphocytes, LYZ and CD68 for myeloid cells, and KIT and MS4A2 for mast cells.

Results

Identification of cell type–specific cREs associated with human lung tissue

To characterize distal regulatory elements surrounding human lung tissue, we obtained 24 H3K27ac ChIP-seq datasets from the ENCODE, DBTSS, and GEO databases, representing eight major cell types (myeloid cells, T cells, B cells, NK cells, endothelial cells, epithelial cells, lung cancer cells, and fibroblasts) [2,3,15] and identified 86,312 distal cREs. The quantification of cRE activities in RPM indicated that the samples clustered according to cell type of origin (Fig. 1A), largely into two groups consisting of stromal (epithelial cells, endothelial cells, and fibroblasts) and immune cells (myeloid, NK, T, and B cells). The cRE profiles of lung cancer cell lines (A549, A427, and H322) were highly correlated with the epithelial cell type, reminiscent of the cell type of their origin. We identified cell type enriched cREs using a quasi-likelihood F test (false discovery rate [FDR] < 0.05) in Bioconductor package EdgeR by contrasting each cell type to the other cell types [22], which resulted in 45,706 cell type–specific cREs (9,707 for lung cancer cells, 14,135 for epithelial cells, 6,952 for myeloid cells, 5,486 for endothelial cells, 4,891 for B cells, 4,447 for fibroblasts, 2,194 for T cells, and 193 for NK cells) (Fig. 1B). This large portion of cell type-specific cREs (52.95%) suggests a dynamic role of regulatory genomic elements in determining cellular identity.

Characterization of lung cancer–associated genetic variations with cell type–specific cREs

To assess the association of common genetic risk variants for lung cancer with cREs with cell type annotation, we collected 286 tag SNPs from the NHGRI-EBI GWAS catalog (downloaded on 2019.08.20) [18] covering nine lung cancer–related traits. Our LD-based association analysis ($r^2 > 0.8$) showed that 21.33% (62 of 286) of lung cancer SNPs were linked to cREs (Fig. 1C). Among them, 28 tag SNPs were associated with cell type-specific cREs (a full list with the corresponding cell type is provided in Table 1), and 34 tag SNPs were linked to constitutive cREs (Supplementary Table 3). SNP-harboring cREs exhibited cell type–specific activities, and myeloid ($n = 11$) and epithelial ($n = 10$) cell types were recognized by having the most SNP-harboring cREs, highlighting a potential role of these cell types in lung carcinogenesis (Fig. 1D).

Identification of distal target genes of SNP-harboring cREs by high-resolution chromatin contact maps

Despite the enrichment of lung cancer-associated genetic variants in cREs, their biological function in lung cancer is largely unknown due to the lack of information about their functional target genes. We hypothesized that investigating the physical chromatin contacts between promoters and SNP-harboring cREs may substantially advance our current knowledge regarding the possible regulatory role of noncoding genetic variants. To this end, we downloaded in situ Hi-C data for the A549, IMR90, GM12878, and HMEC cell lines from the ENCODE database [2,3], representing lung cancer, fibroblasts, myeloid cells, and epithelial cells, respectively. We defined long-range chromatin interactions at 5 kb resolution, implementing the Fit-Hi-C algorithm (FDR < 0.01) [19], which resulted in a total of 3,785,594 unbiased, all-to-all chromatin interactions (1,905,639 for A549, 1,207,580 for IMR90, 1,729,755 for GM12878, and 220,529 for HMEC). Focusing on the chromatin interactions connected to the well-annotated protein-coding gene promoters, we identified chromatin interactions anchored within 2.5 kb of a TSS. Comparison of these promoter-centered interactions across the cell types presented a highly dynamic pattern, with lung cancer (A549) and myeloid (GM12878) cells having a large number of unique chromatin interactions, implicating differential spatial arrangements between cREs and promoters in those cell types (Fig. 2A). The average distance for long-range chromatin interactions was similar across cell
types (234 kb for A549, 222 kb for GM12878, 269 kb for IMR90, and 221 kb for HMEC) (Fig. 2B). Finally, we predicted putative target genes of cell type–specific cREs harboring lung cancer-associated variants by using a union set of chromatin interactions (Fig. 2C). To assess the cell type–specificity of the inferred target genes, we integrated a scRNA-seq dataset generated from lung tissues of 11 healthy individuals and 11 lung cancer patients [20]. We found that the inferred target genes showed a higher gene expression in the corresponding cell type compared to the randomly selected controls (empirical p =  0.0016 from 100,000 iterations) (Fig. 2D). Although the statistical testing for cell type specificity in target gene expression indicated insignificance in enrichment for epithelial and myeloid cell types (Fisher’s exact test; p =  0.1591 for epithelial and p =  0.607 for myeloid), we identified 20 and 5 putative target genes in epithelial and myeloid cells, respectively, with the highest gene expression in the corresponding cell type (Fig. 2E).
<table>
<thead>
<tr>
<th>Tag SNP</th>
<th>rSID</th>
<th>SNP (P)</th>
<th>Trait</th>
<th>Journal</th>
<th>cRE coordinates</th>
<th>Cell type specificity (FDR)</th>
<th>Cell type</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr2:17784157</td>
<td>rs13031455</td>
<td>2.00.E-06</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017) [23]</td>
<td>chr2:17801191-17803176</td>
<td>1.15.E-06</td>
<td>Epithelial</td>
</tr>
<tr>
<td>chr5:72305846</td>
<td>rs258892</td>
<td>5.00.E-06</td>
<td>Small cell lung carcinoma</td>
<td>Nat Genet (2017) [23]</td>
<td>chr5:72083963-72085536</td>
<td>1.41.E-03</td>
<td>Epithelial</td>
</tr>
</tbody>
</table>

Continued

https://doi.org/10.5808/gi.20073

5 / 12
Table 1. Continued

<table>
<thead>
<tr>
<th>Tag SNP</th>
<th>rsID</th>
<th>SNP (P)</th>
<th>Trait</th>
<th>Journal</th>
<th>cRE coordinates</th>
<th>Cell type specificity (FDR)</th>
<th>Cell type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>chr6:32599939-32607692</td>
<td>6.99.E-04</td>
<td>B cell</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>chr6:32652218-32660026</td>
<td>2.53.E-03</td>
<td>B cell</td>
</tr>
</tbody>
</table>

GWAS, genome-wide association study; SNP, single nucleotide polymorphism; cRE, cis-regulatory element; FDR, false discovery rate.

In addition, we found that HYLS1, IL1R1, and PPP1R18 were inferred target genes in endothelial, fibroblast, and T cell, respectively, which also presented a cell type-specific gene expression. The list of inferred target genes with cell type-specific gene expression is provided in Table 2. The statistical insignificance in cell type specificity may be the result of the limited number of tested target genes and the insufficient transcripts detected in the corresponding scRNA-seq technique. The improvement of scRNA-seq techniques covering a higher number of transcripts will help make a clear conclusion in this matter. Altogether, the inference of target genes of cell type-specific, SNP-harboring cREs by using 3D chromatin interaction profile provided considerable insights into functionality of lung cancer-associated GWAS-SNPs.

Characterization of DDR1 and CD84 as potential risk factors associated with lung carcinogenesis

Since the number of putative target genes with cell type dependent expression was the most prevalent in epithelial and myeloid cells, we sought to pin-point specific risk candidate genes by taking a close examination of the epigenomic landscape surrounding the major cell types in the human lung. First, we found that an epithelial-specific cRE (chr6:30,889,573-30,895,783; FDR = 1.61E-05) contains a SNP at chr6:30,894,965, which is linked to a tag SNP at chr6:30,882,415 (rs114274879) based on LD structure. The SNP was significantly associated with squamous cell lung carcinoma (p = 3.0E-16) [23]. The SNP-containing cRE was linked to the promoter of DDR1 by significant long-range chromatin interaction stretching as far as 230 kb in distance (Fig. 3B). The recently demonstrated function of CD84 in chronic lymphocytic leukemia cells and their microenvironment may support the potential functional implication of CD84 in lung carcinogenesis [30]. Furthermore, single-cell transcriptome data indicated that the expression of CD84 is exclusive in myeloid cells (Fig. 3B). Our results highlight the potential role of DDR1 and CD84 in lung carcinogenesis within epithelial and myeloid populations, respectively. The current work involving the functional annotation of lung cancer GWAS-SNPs and the inference of their putative target genes using 3D chromatin contact information effectively expands potential risk candidate genes and their relevant cell types, and offers a rationale for a further investigation of its function in the designated cell type.

Discussion

We used a comprehensive multi-omics approach that integrates cell type-specific epigenome, 3D chromatin interactome, and transcriptome data to conduct a functional characterization of lung cancer-associated GWAS risk variants. It is worth noting that the cellular identity in the global cRE landscape is well replicated regardless of tissue origin, genetic background, and culture condition (e.g., primary cells and immortalized cell line), evidenced by the use of publicly available ChIP-seq data representing various cell types in the current study. This led us to effectively find a considerable portion of risk genetic variants associated with cREs, taking into consideration the genetic LD. The categorization of lung cancer GWAS-SNPs into corresponding cell types provides additional insights into specific cellular populations responsible for lung carcinogenesis. Our results provide evidence that the identification of the cell type-specific promoter-cRE interactome substantially advances the interpretation of GWAS risk variants and broadens the scope for disease risk candidates for lung cancer.

The recent advent of single-nucleus accessible chromatin profil-
Fig. 2. Target gene identification of cell type-specific cREs harboring lung cancer-associated SNPs based on long-range chromatin interactions. (A) Chow–Ruskey plot with a 5 kb resolution promoter-centered chromatin interactions for GM12878, A549, HMEC, and IMR90 cell lines, representing myeloid, lung cancer, endothelial, and fibroblast cell types, respectively. (B) Density plots illustrating the genomic distance of long-range chromatin interactions obtained from Hi-C data. The dashed line represents the mean distance. (C) A description of the functional link between SNP-harboring cREs and inferred target genes through a long-range chromatin contact. (D) Histograms illustrating distribution of the relative expression of randomly selected gene sets based on iterative tests (n = 100,000). Yellow dotted arrows indicate the observed expression of inferred target genes. (E) Gene expression (z-transformed normalized single-cell RNA sequencing counts) of putative target genes of cell type-specific cREs harboring lung cancer-related GWAS-SNPs across the cell types. Genes highlighted in translucent green, brown, purple, orange, yellow, and blue indicates putative targets of cell type-specific cREs in epithelial, endothelial, fibroblast, myeloid, B cells, and T cells, respectively. cRE, cis-regulatory elements; SNP, single nucleotide polymorphism; GWAS, genome-wide association study.
Fig. 3. Epigenome landscape of putative target genes of cell type-specific cRE harboring lung cancer risk variants. (A) Left: Epigenome browser visualization of the DDR1 locus (chr6:30,985,829-30,985,829) showing the localization of lung cancer-related GWAS-SNPs, H3K27ac signals over seven individual cell types associated with the human lung, and 5 kb-resolution chromatin loops. The bars in dark orange indicate the location of cell type-specific cREs. The region of epithelial-specific cREs sharing lung cancer–related genetic variants is highlighted in translucent yellow. Right: DDR1 gene expression level across 7 major lung tissue cell types from scRNA-seq data. (B) Left: Epigenome browser visualization of the CD84 locus (chr1:160,197,000-160,668,000) showing the localization of lung cancer-related SNPs, H3K27ac signals over seven individual cell types associated with the human lung, and 5 kb-resolution chromatin loops. The bars in dark orange indicate the location of cell type dependent cREs. The region of myeloid-specific cREs sharing lung cancer risk variants is highlighted in translucent yellow. Right: CD84 gene expression level across 7 major lung tissue cell types from single-cell RNA-seq data. cRE, cis-regulatory elements; GWAS, genome-wide association study; SNP, single nucleotide polymorphism; scRNA-seq, single-cell RNA-sequencing; LD, linkage disequilibrium.
Table 2. A list of inferred target genes of SNP-harboring cREs with cell-type specific expression

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Ensemble ID</th>
<th>cRE coordinates</th>
<th>Cell type</th>
<th>Cell type specificity (FDR)</th>
<th>Tag SNP</th>
<th>rsID</th>
<th>SNP (P)</th>
<th>Trait</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSORS1C2</td>
<td>ENSG000000204533.8</td>
<td>chr6:30848819-30857882</td>
<td>Epithelial</td>
<td>chr6:30882415</td>
<td>rs114274879</td>
<td>3.00E-16</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
<tr>
<td>STRF2</td>
<td>ENSG000000196260.3</td>
<td>chr6:30848819-30857882</td>
<td>Epithelial</td>
<td>chr6:30882415</td>
<td>rs114274879</td>
<td>3.00E-16</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
<tr>
<td>DPCR1</td>
<td>ENSG000000168631.7</td>
<td>chr6:30848819-30857882</td>
<td>Epithelial</td>
<td>chr6:30882415</td>
<td>rs114274879</td>
<td>3.00E-16</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
<tr>
<td>MUC22</td>
<td>ENSG000000026127.1</td>
<td>chr6:30889573-30895783</td>
<td>Epithelial</td>
<td>chr6:30882415</td>
<td>rs114274879</td>
<td>3.00E-16</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
<tr>
<td>DGKD</td>
<td>ENSG000000077044.5</td>
<td>chr2:233453904-233458464</td>
<td>Epithelial</td>
<td>chr2:233453904-233458464</td>
<td>rs1656402</td>
<td>8.00E-08</td>
<td>Non-small cell lung cancer (survival)</td>
<td>J Thorac Oncol (2010)</td>
<td></td>
</tr>
<tr>
<td>POUSF1</td>
<td>ENSG000000204531.1</td>
<td>chr6:30889573-30895783</td>
<td>Epithelial</td>
<td>chr6:30882415</td>
<td>rs114274879</td>
<td>3.00E-16</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
<tr>
<td>CCHCR1</td>
<td>ENSG000000204536.9</td>
<td>chr6:30848819-30857882</td>
<td>Epithelial</td>
<td>chr6:30882415</td>
<td>rs114274879</td>
<td>3.00E-16</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
<tr>
<td>PSORS1C1</td>
<td>ENSG000000204540.6</td>
<td>chr6:30889573-30895783</td>
<td>Epithelial</td>
<td>chr6:30882415</td>
<td>rs114274879</td>
<td>3.00E-16</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
<tr>
<td>VAR2</td>
<td>ENSG000000137411.12</td>
<td>chr6:30889573-30895783</td>
<td>Epithelial</td>
<td>chr6:30882415</td>
<td>rs114274879</td>
<td>3.00E-16</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
<tr>
<td>TIGD1</td>
<td>ENSG00000021944.3</td>
<td>chr2:233453904-233458464</td>
<td>Epithelial</td>
<td>chr2:233453904-233458464</td>
<td>rs1656402</td>
<td>8.00E-08</td>
<td>Non-small cell lung cancer (survival)</td>
<td>J Thorac Oncol (2010)</td>
<td></td>
</tr>
<tr>
<td>KLHC10</td>
<td>ENSG000000128607.9</td>
<td>chr7:130666628-130685063</td>
<td>Epithelial</td>
<td>chr7:130666818</td>
<td>rs6957511</td>
<td>1.00E-06</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
<tr>
<td>CHRNA1</td>
<td>ENSG000000135902.5</td>
<td>chr2:233453904-233458464</td>
<td>Epithelial</td>
<td>chr2:233453904-233458464</td>
<td>rs1656402</td>
<td>8.00E-08</td>
<td>Non-small cell lung cancer (survival)</td>
<td>J Thorac Oncol (2010)</td>
<td></td>
</tr>
<tr>
<td>C6orf15</td>
<td>ENSG000000204542.2</td>
<td>chr6:30889573-30895783</td>
<td>Epithelial</td>
<td>chr6:30882415</td>
<td>rs114274879</td>
<td>3.00E-16</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
<tr>
<td>PRR</td>
<td>ENSG000000204576.7</td>
<td>chr6:30889573-30895783</td>
<td>Epithelial</td>
<td>chr6:30882415</td>
<td>rs114274879</td>
<td>3.00E-16</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
<tr>
<td>ZNF322</td>
<td>ENSG000000181315.6</td>
<td>chr6:27094480-27096156</td>
<td>Epithelial</td>
<td>chr6:2668613</td>
<td>rs10456332</td>
<td>7.00E-06</td>
<td>Small cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
<tr>
<td>MDC1</td>
<td>ENSG000000137337.10</td>
<td>chr6:30889573-30895783</td>
<td>Epithelial</td>
<td>chr6:30882415</td>
<td>rs114274879</td>
<td>3.00E-16</td>
<td>Squamous cell lung carcinoma</td>
<td>Nat Genet (2017)</td>
<td></td>
</tr>
</tbody>
</table>
ing allows effective identification and characterization of cell populations within human tissues. For example, the single-nucleus ATAC-seq (snATAC-seq) data from the human lung generated by Wang et al. [31] identified six sub-clusters in the epithelial population (AT1/AT2, PNEC, club, basal, and ciliated cells) and cell type-specific gene regulation associated with viral entry. However, the read-depth and coverage obtained in snATAC-seq data are considerably low for each cell type when compared with bulk ChIP-seq data utilizing primary cells and cell lines. The collection of individual ChIP-seq samples representing major cell types in the human lung, as conducted in this study, may allow a more discrete cell type-specific investigation of regulatory dynamics. Moreover, the development of scRNA-seq allowed a population-based analysis of transcriptome data. However, it is worth noting that the number of genes detected by scRNA-seq is limited to a few thousand, which largely restricted our scope of investigating the inferred target genes of cell type-specific cREs. Finally, the current work involving the functional annotation of lung cancer GWAS-SNPs and inference of their putative target genes is a notable endeavor, which will provide qualitative insights into disease mechanisms that may be of value in identifying new risk factors developing new approaches for prevention and treatment.

**ORCID**

Andrew J. Lee: https://orcid.org/0000-0001-9917-8834
Inkyung Jung: https://orcid.org/0000-0002-5885-2754

**Authors’ Contribution**

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

**Conflicts of Interest**

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

**Acknowledgments**

This work was supported by the Ministry of Science and ICT through the National Research Foundation in Republic of Korea under grant number NRF-2020R1A2C4001464 to IJ. A portion of the data used for this study was obtained from the Genome-InfraNet (IDs: 1711056526 and 1711065396) of the Korean Bioinformation Center.
Supplementary Materials

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

References


https://doi.org/10.5808/gi.20073


Identification of potential candidate genes for lip and oral cavity cancer using network analysis

Sarmilah Mathavan, Chin Siang Kue, Suresh Kumar*

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

Lip and oral cavity cancer, which can occur in any part of the mouth, is the 11th most common type of cancer worldwide. The major obstacles to patients’ survival are the poor prognosis, lack of specific biomarkers, and expensive therapeutic alternatives. This study aimed to identify the main genes and pathways associated with lip and oral cavity carcinoma using network analysis and to analyze its molecular mechanism and prognostic significance further. In this study, 472 genes causing lip and oral cavity carcinoma were retrieved from the DisGeNET database. A protein–protein interaction network was developed for network analysis using the STRING database. VEGFA, IL6, MAPK3, INS, TNF, MAPK8, MMP9, CXCL8, EGF, and PTGS2 were recognized as network hub genes using the maximum clique centrality algorithm available in cytoHubba, and nine potential drug candidates (ranibizumab, siltuximab, sulindac, pomalidomide, dexrazoxane, endostatin, pamidronic acid, cetuximab, and apricoxib) for lip and oral cavity cancer were identified from the DGIdb database. Gene enrichment analysis was also performed to identify the gene ontology categorization of cellular components, biological processes, molecular functions, and biological pathways. The genes identified in this study could furnish a new understanding of the underlying molecular mechanisms of carcinogenesis and provide more reliable biomarkers for early diagnosis, prognostication, and treatment of lip and oral cavity cancer.

Keywords: biomarkers, carcinogenesis, lip neoplasms, mouth neoplasms, prognosis

Introduction

Human head and neck cancers begin in the mouth, nose, throat, larynx, sinuses, or salivary glands. Lip and oral cancer is a subgroup of head and neck cancers that cause lip or oral carcinoma. Oral cancers are often referred to as those occurring in a particular anatomical region, including the lip, gum, tongue, mouth (including the floor of the mouth), and palate, corresponding to the International Classification of Diseases, 10th revision C00-06 code. The vast majority of these cancers (up to 85%–95%) are squamous cell carcinomas, often resulting from pre-existing precancerous lesions. Oral cancer stands out among head and neck tumors due to its frequent occurrence and mortality rate, as well as its common association with a late diagnosis [1].

The low survival rate of oral cancer can be significantly increased if it is detected early or in the pre-cancer stage. Most oral carcinomas are squamous cell carcinomas of the tongue, buccal mucosa, or gums. Lip cancer is the most common tumor in the head and neck of the body, and constitutes 25%–30% of all mouth cancers [2]. Lip carcinomas are usually basal or squamous cell carcinomas [3]. The oral cavity begins from the blood-red...
boundary of the lips and extends to the circumvallate papillae of the tongue and the intersection of the soft and hard palate. The oral cavity is divided into the lip, oral tongue, mouth floor, buccal mucosa, upper and lower gum, retromolar trigon, and hard palate [4]. Benign oral cavity lesions include those affecting the anterior tongue, mouth floor, buccal mucosa, retromolar trigone, hard palate, and gingiva.

An estimated 200,000 cases of oral cancer and 100,000 deaths occur every year worldwide. The worldwide age-standardized prevalence of lip cancer was reported to be 0.3 per 100,000 in 2012 (0.4 in males and 0.2 in females) [5]. Smoking tobacco and excessive alcohol intake account for 75% of cases of lip cancer and oral cancer. Other risk factors include chewing of betel juice (paan) with or without tobacco and the consumption of nitrosamine-rich foods and salted fish. Another significant risk indicator for oral cancer is the overconsumption of cigarettes [6]. The onset of intake, period, and extent of regular use of chewing tobacco or use of bidis as a form of smoked tobacco were closely associated with oral cancer [7]. Smoked tobacco is a major contributor to carcinogenesis in the upper airway, and a positive association exists between tobacco smoke consumption and oral cavity cancer, as documented in numerous studies. The risk of oral cancer is 1.4–1.7 times higher in those who consume tobacco than in those who do not consume tobacco [8]. Paan includes areca nut, betel leaves, and slaked lime, sometimes mixed with tobacco, and certain items like spices, sweets, and essences may be added to paan, depending mostly on taste. In Asia, smoking paan with or without tobacco is a major risk factor for oral cancer [9], but owing to a lack of knowledge and literacy, many people who routinely use paan are unaware of its adverse consequences for health [8].

Other predictors of oral cancer are environmental pollutants such as ultraviolet radiation (lip cancer) or nutritional intake deficits such as fruit and non-starchy vegetables (oral cavity cancer) [5]. Various mutations and genetic mechanisms have also been reported to contribute to lip and oral cancer development and growth. Surgery is usually performed to treat oral squamous cell carcinoma. Surgery facilitates an accurate assessment of the anatomical location, margins, invasive status, and histopathological characteristics of the tumor, and the corresponding advantages and disadvantages may determine the strategy.

To date, the ability to treat advanced oral cancer has been constrained by a lack of understanding of the specific key genes that underlie the growth of this cancer. The aim of this study was to identify key genes through a gene enrichment analysis of lip and oral cavity carcinoma using bioinformatics and to identify novel potential diagnostic biomarkers for oral carcinoma.

Methods

Retrieval of disease genes

The genes associated with lip and oral cavity carcinoma (C0220641) were retrieved from the DisGeNET database (accessed December 2019) available at http://www.disgenet.org/home/. DisGeNET is a wide-ranging platform that integrates genes and variants involved in human disease [10]. A total of 472 disease-associated genes were obtained and downloaded for further analysis.

Network analysis in Cytoscape

The UniProt IDs of associated disease genes recorded from the summary of disease-gene associations were uploaded using the STRING protein query in Cytoscape 3.7.2, which is a precomputed global resource designed to evaluate protein-protein interaction (PPI) information [11]. The confidence score (cutoff) was set to 0.4 and the maximum additional interactors remained the default parameter to obtain more closely related genes to the targeted protein. The STRING network of PPI was constructed and displayed a hierarchical layout for a better view.

Hub gene retrieval in cytoHubba

The cytoHubba taskbar is a convenient tool for extracting a subnetwork that contains hub genes from an entire large PPI collection. As the scoring method, maximum clique centrality (MCC) was selected to identify featured nodes. In the cytoHubba plugin, the MCC algorithm has been reported to be the most effective method of finding hub nodes. In this study, the top 10 genes with the highest MCC values were considered as hub genes [12].

Functional and pathway enrichment analysis of hub genes

The WEB-based GEnE SeT AnaLysis Toolkit (WebGestalt) (functional enrichment analysis web) integrates management, data retrieval, organization, visualization, and statistical studies of functional enrichment and data visualization, as well as the analysis of large gene sets, and is available at http://www.webgestalt.org/ [13]. The WebGestalt database is unique in that it includes information from various biological contexts, including gene ontology (GO), the Reactome pathway, network module, gene-phenotype and gene-disease interactions, gene-drug associations, and chromosome position. WebGestalt has greatly expanded the scope of functional domains, contributing to a total of 78,612 functional categories. It provides a graphical representation of the data with over-representation enrichment analysis, mainly involving cellular components, biological processes, molecular functions, and biological pathways [14]. A false discovery rate (FDR) \( \leq 0.05 \) was considered to indicate statistical significance.
Identification of drug candidates based on hub genes

The online tool DGIdb (http://www.dgidb.org/), an available resource containing drug-gene interaction data from more than 30 databases, was used to screen antineoplastic drugs targeting hub genes.

Results and Discussion

In total, 472 genes associated with lip and oral cavity carcinoma (C0220641) were identified using the DisGeNET database (Supplementary Table 1). DisGeNET combines text-mined databases with expert-curated databases, provides one of the most extensive sets of associations of human genetic diseases, and is a valuable module for the study of molecular mechanisms underlying genetic diseases [10]. The PPI network was constructed using all 472 genes. PPI analyses help to explain protein roles at the molecular level and to discover the process of cell regulation. The STRING database is often used to evaluate and pre-calculate global-view protein associations comprising 89 full genome sequence datasets, including 261,033 orthologous genes [15]. The visual representation of the predicted, ranked protein interactions network in the STRING database offers a high-level overview of functional associations, enabling extensible analyses of biological systems [15]. In total, 444 nodes and 8,573 edges (interactions) were identified with the STRING network based on a confidence score of 0.4 and the maximum additional interactors as the default parameter (Fig. 1).

Nodes and edges are particularly important because they can be linked to data on gene expression and protein structure information. CytoHubba, which has been widely used to explore important nodes in biological networks, was then applied to identify the lip and oral cavity carcinoma hub proteins in the PPI network. First, scores for all 11 methods are given to each node across the pre-loaded PPI network by selecting the options for cytoHubba from the network panel. The MCC score method, which is a local-based method, was chosen for this study. The PPI network and hub genes were visualized using a hierarchical layout. A local rank approach will only recognize the relationship between the node and its significant neighbors to measure the node’s score within the network. The MCC method was used to explore the features of the nodes to enhance their efficacy. According to the MCC scores, the top 10 ranked nodes for each specific score system were then obtained from the cytoHubba (Table 1) column in the CytoHubba taskbar and shown in the results column, and the subgraph of all the identified nodes can be seen in the taskbar windows with a large (red) to basic (yellow) color palette (Fig. 2). The top 10 genes associated with lip and oral cavity carcinoma according to MCC scores from the cytoHubba plugin in Cytoscape were screened [12]. The identified key genes responsible for lip and oral cavity carcinoma were vascular endothelial growth factor A (VEGFA), followed by interleukin-6 (IL6), mitogen-activated protein kinase 3 (MAPK3), insulin (INS), tumor necrosis factor (TNF), mitogen-activated protein kinase 8 (MAPK8), matrix

Fig. 1. Protein-protein interaction network overview built using STRING in Cytoscape. The network consists of 8,573 edges (interactions) between 444 nodes based on a confidence score of 0.4 and the maximum additional interactors default parameter. Nodes represent proteins, edges represent the interaction between two nodes (proteins).

Table 1. The top 10 ranked nodes were selected using the MCC method in the cytoHubba app in Cytoscape 3.7.2

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gene name</th>
<th>MCC score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VEGFA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.88148120178673E+37</td>
</tr>
<tr>
<td>2</td>
<td>IL6</td>
<td>1.88148120178002E+37</td>
</tr>
<tr>
<td>3</td>
<td>MAPK3</td>
<td>1.88148120172653E+37</td>
</tr>
<tr>
<td>4</td>
<td>INS</td>
<td>1.88148118653989E+37</td>
</tr>
<tr>
<td>5</td>
<td>TNF</td>
<td>1.88148116290353E+37</td>
</tr>
<tr>
<td>6</td>
<td>MAPK8</td>
<td>1.88148112775536E+37</td>
</tr>
<tr>
<td>7</td>
<td>MMP9</td>
<td>1.88147822258162E+37</td>
</tr>
<tr>
<td>8</td>
<td>CXCL8</td>
<td>1.88147770682472E+37</td>
</tr>
<tr>
<td>9</td>
<td>EGF</td>
<td>1.88147075212715E+37</td>
</tr>
<tr>
<td>10</td>
<td>PTGS2</td>
<td>1.88128391642781E+37</td>
</tr>
</tbody>
</table>

These top 10 ranked nodes represent the top 10 hub genes of lip and oral cavity carcinoma.

<sup>a</sup>Based on these results, VEGFA was identified as the highest-ranked hub gene with the highest maximum clique centrality (MCC) score.
metalloproteinase-9 (MMP9), interleukin-8 (CXCL8), pro-epidermal growth factor (EGF), and prostaglandin G/H synthase 2 (PTGS2).

The top 10 genes in the string network ranked by the MCC method were then analyzed for gene enrichment analysis using WebGestalt. To identify the functions, a gene enrichment analysis of the hub genes was performed, and an FDR ≤ 0.05 was set as the cutoff value [13].

Fig. 3B shows the results of the cellular component GO term enrichment analysis, which suggested that all 10 hub genes were significantly enriched in the endomembrane system. The genes are also present in the membrane-enclosed lumen, extracellular space, membrane, and vesicle, and actively function in the protein-containing complex, endoplasmic reticulum, endosome, cell projection, and extracellular matrix. Other cellular components involving these hub genes include the nucleus, mitochondrion, Golgi apparatus, cytosol, cytoskeleton, cell envelope, and vacuole. The cell polarity position of the endomembrane pool of Cdc42 and the possible role of this pool in cancer-related alterations are known. The Golgi apparatus was noted to be rapidly oriented towards the posterior end of the plasma membrane, meaning that its integrity is necessary for guided cell motility and polarized secretion [16]. In the form of single-strand malfunctions, tobacco smoking causes significant damage to DNA [17], and increases in protein thiols and lipid peroxidation/oxidation [18]. Mitochondrial DNA is vulnerable to harm from reactive oxygen species such as superoxide radicals, hydrogen peroxide, and hydroxyl radical due to a lack of a
defensive histone backbone [19]. In the vast system of close and distant cell-to-cell communication, extracellular vesicles (EVs) are secreted by most cell types [20]. Under the influence of hypoxia-inducible factors, EVs that developed under a stressful environment showed enhanced proliferation and migration of oral cancer cells [21]. As for the molecular function of the lip and oral cavity carcinoma hub genes (Fig. 3C), most of these genes are involved in protein binding, with moderate functions in ion binding, transferase activity, nucleotide binding, and enzyme regulatory activity. Protein binding is a secreted glycoprotein that enhances cell-cell and cell-extracellular matrix permeability and induces the production of IL-1, IL-6, and other blood monocyte cytokines, contributing to the invasion and metastasis of lip and oral cancer cells [22]. Antioxidant disruption to salivary DNA and proteins can encourage oral squamous cell carcinoma [23]. The mechanism through which oxidative damage is involved in oral cancer is that when the salivary DNA is extracted from exfoliated oral epithelium, the oxidized proteins and DNA contained in the saliva interact with salivary free radicals [24]. Hydrolase is involved in the metabolism of tobacco carcinogens [25]. Hydrolase activity is linked to an increased risk of oral cavity, pharyngeal, and laryngeal cancers, which are smoking-related cancers, and recent study supported its involvement in lung cancer [26]. In terms of biological processes (Fig. 3A), most of the hub genes are involved in cell communication, metabolic processes, multicellular organismal processes, developmental processes, responses to stimuli, and localization and biological regulation. These hub genes are also involved in cell proliferation, cellular component organization, and multi-organism processes. Cancer cells, which have altered glucose and lipid metabolism, show radical improvements in energy metabolism function compared to normal cells. Tumor metabolism studies have reported that oncogenic signaling pathways stimulated metabolic reprogramming to upregulate lipid, carbohydrate, protein, DNA, and RNA biosynthesis, leading to enhanced tumor development. Under aerobic conditions, cancer tissues have elevated glycolysis levels in the cytosol, even with fully functioning mitochondria, as a result of phosphoinositide 3-kinase (PI3K)/AKT signaling; this is known as the Warburg effect [27]. The cell's regular activities and organization are closely regulated by excitatory or inhibitory input [28]. In tumor cells, pathways are altered, enabling them to divide quickly, sequester blood vessels that stimulate growth, remove or enhance signals to create abnormal functional or structural alterations, and penetrate local or remote sites of normal tissue [29]. Fig. 4 shows the biological pathways (Reactome pathways). Most of these genes are involved in activation of the AD-1 family of transcription factors, signal attenuation, and MAPK3 (ERK1) activation. The transcription factor protein-1 (AP-1) superfamily activator is known to modulate gene expression during the development of many cancers and has been recognized as a potential target for modern therapeutic applications. Its components are involved in RAF-independent MAPK 1/3 activation, interleukin-10 signaling, MAPK targets/nuclear events mediated by MAPKs and high-affinity IgE receptor-mediated MAPK activation including interleukin-4 and interleukin-3 signaling. These hub genes were found to be involved in phosphatidylinositol-5-phosphate/phosphatidylinositol-4,5-bisphosphate and im-

Fig. 4. Gene enrichment analysis of 10 recognized significant hub genes (false discovery rate [FDR] ≤ 0.05) based on the biological pathways of the Reactome database.
mediate early response 3-regulated PI3K/AKT signaling, regulation of regulate phosphatidylinositol 3-kinase and Akt/Protein Kinase B (PI3K/AKT) signaling, senescence-associated secretory phenotype, cellular senescence, and signaling by interleukin. The PI3K/Akt pathway is a central controller of viability in response to cell stress (e.g., pH, nutrient, and oxygen levels), and deregulation of the PI3K signaling pathway leads to cancer [30].

VEGFA encodes a platelet-derived growth factor/vascular endothelial growth factor (VEGF) family member that acts as a glycosylated mitogen, increasing endothelial permeability, angiogenesis, vasculogenesis, endothelial cell growth, and cell migration. VEGFA functions as a central stimulator of angiogenesis, which is an important trait of cancer that plays a crucial role in tumor growth. The production of VEGFA is induced by the generation of hypoxic conditions within tumors [31]. A study reported that VEGFA mRNA levels were 53-fold higher in oral carcinoma tissues than in normal tissues. Hence, VEGFA functions as a potent autocrine survival factor for cancer cells. The risk of oral cancer may be correlated with VEGFA locus haplotypes, and the haplotype effect may be more substantial than a single susceptibility polymorphism [32]. Compared to the normal oral mucosa, multiple studies have shown upregulation of VEGFA expression in cancerous tissues. VEGF levels in oral cancer patients were also found to be significantly higher than normal controls, in an analysis that included clinical stage and lymph node metastasis. This suggests that VEGFA levels may be a reliable biomarker and that VEGFA may be a potential target for developing chemotherapy strategies for oral carcinoma patients.

IL-6 is a pleomorphic cytokine involved in various physiological and pathological processes, such as responses to trauma and infection and the progression of inflammation and tumors. IL-6 appears to lead to oral cancer pathogenesis via multiple pathways and biological processes [33]. IL-6 can stimulate the release of matrix metalloproteinase 1 and 9, which are responsible for malignant growth and neoangiogenesis in oral squamous cell carcinoma. Several keratinocyte mechanisms, including cell formation, survival, and differentiation, are also modulated by IL-6. By triggering global hypomethylation and changes in DNA methylation trends in oral cancer cells, IL-6 can contribute to the growth of oral cancer [34]. IL-6 is linked with increased tumor growth and metastasis, and may therefore be involved in this disease’s pathogenesis. Serum IL-6 was detected at higher concentrations than salivary IL-6 in oral cancer patients. Therefore, serum IL-6 was proposed as a diagnostic or prognostic marker for oral cancer and pre-cancer.

Tumor cell development, differentiation, apoptosis, angiogenesis, invasion, and metastasis are associated with MAPK3 and MAPK8. The repression of MAPK signaling caused by irregular gene expression leads to abnormal responses, whereas regulated MAPK inhibits the development of inhibitory proteins in cell cycles. PI3K/Akt gene mutations lead to irregular activation of the MAPK pathway in oral cancer. This observation suggests an intricate relationship between the MAPK and PI3K/Akt pathways [35].

INS encodes insulin. By modifying the insulin receptor substrate-1 (IRS-1) pathway, diabetes can also raise the risk of certain forms of cancer, including oral carcinoma. Changes in pathway thus an intermediate step towards neoplasia, involving cytoskeleton modifications and decreased cell adhesion. Both integrin and focal adhesion kinase involvement induce IRS-1 activation upon INS activation by tyrosine phosphorylation [36]. TNF promotes cell proliferation and apoptosis, and was reported to be present at high levels in patients with oral leukoplakia, oral lichen planus, and oral submucous fibrosis, which have been claimed to be clinical biomarker for oral cancer [37].

MMP-9 is also a possible biomarker of oral cancer. MMP-9 is a family of enzymes that has been found to be linked to tumor progression because they are active in extracellular matrix breakdown. More specifically, MMP-9 is a family of zinc-dependent proteinases associated with type IV collagen, a key source of the basal lamina, and other forms of collagen in various pathological conditions [38].

CXCL8, also known as interleukin-8 (IL-8), is involved in oral cancer invasiveness through activation of MMPs. Metastatic activity was correlated with interactions between IL-8 and MMP. Well-established inflammatory cytokines in oral cancer cells have been shown to affect MMP development [39].

The EGF receptor is a cell-surface tyrosine kinase known to regulate the metastasis and recurrence of oral cancer. Abnormal stimulation of the downstream signaling pathways promotes the epithelial-to-mesenchymal transition, which ultimately results in neoplastic cells with elevated invasive and metastatic capability [40].

In oral cancers, PTGS2 or cyclooxygenase 2 was elevated. Manifesting as broad hypomethylation in the promoter region of the CpG island area, PTGS2 expression occurs at various organ sites in response to stress, cigar smoke, and pharmacological drugs, altering the methylation of the PTGS2 promoter [41].

Based on the 10 predicted hub genes as potential therapeutic targets for lip and oral cancer, we identified several antineoplastic drugs based on the DGIldb database. Specifically, nine small-molecule drugs (ranibizumab, siltuximab, sulindac, pomalidomide, dexrazoxane, endostatin, pamidronic acid, cetuximab, and apricoxib) were identified as potentially having therapeutic effects for lip and oral cavity cancer based on their interaction scores in the DGIldb database. The interaction scores were calculated based on the evidence score and relative drug and gene specificity. Table 2 shows...
small-molecule drugs with potential therapeutic effects for lip and oral cavity cancer based on the highest interaction score for each predicted hub gene. However, it is indeed necessary to support promising therapeutic targets with more studies.

Our analysis has many advantages over previous work [42-44]. First, this study had a broad sample size retrieved from the DisGeNET database, which covers research from expert-curated repositories, genome-wide association study catalogs, animal models, and scientific literature. We further studied the functional and pathway enrichment of important genes. To our best knowledge, we identified some heretofore unreported prognostic biomarkers, such as MAPK3, INS, and PTGS2. However, further clinical validation of these reported biomarkers is needed.

**Conclusion**

A comprehensive perspective was provided by the bioinformatics analysis to understand the mechanism underlying lip and oral cavity carcinoma development. In this study, the following hub genes were identified as being involved in lip and oral cavity carcinoma through network analysis: VEGFA, IL6, MAPK3, INS, TNF, MAPK8, MMP9, CXCL8, EGF, and PTGS2. In total, 472 gene-disease associations and 10 hub genes were identified and recognized as target biomarkers for lip and oral cavity carcinoma. We also identified several antineoplastic drugs with potential applications for lip and oral cavity cancer. A detailed study of the genes’ biological mechanisms and their pathways may provide potential targets for the therapeutic drug monitoring of lip and oral cavity carcinoma. Nevertheless, further studies are required to understand the development of lip and oral cavity carcinoma to unravel its mechanism more completely.

---

**ORCID**

Sarmilah Mathavan: https://orcid.org/0000-0001-6105-502X
Chin Siang Kue: https://orcid.org/0000-0002-9793-5616
Suresh Kuma: https://orcid.org/0000-0001-5682-0938

**Authors’ Contribution**

Conceptualization: SK. Data curation: SM, CSK, SK. Formal analysis: SM, CSK, SK. Methodology: SK. Writing - original draft: SM, CSK, SK. Writing - review & editing: SM, CSK, SK.

**Conflicts of Interest**

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

**Supplementary Materials**

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

**References**


Genetic alterations in Wnt family of genes and their putative association with head and neck squamous cell carcinoma

Jain Aditya¹, A. S. Smiline Girija¹, A. Paramasivam², J. Vijayashree Priyadharsini²*

¹Department of Microbiology, Saveetha Dental College, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Chennai 600077, India
²Biomedical Research Unit and Laboratory Animal Centre-Dental Research Cell, Saveetha Dental College, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Chennai 600077, India

Introduction

The head and neck squamous cell carcinoma (HNSCC) represents a heterogeneous group of cancer affecting the mucosal surfaces of several organs including nasal cavity, paranasal sinuses, oral cavity, tongue, pharynx, and larynx [1]. HNSCC accounts for more than 330,000 deaths worldwide with more than 650,000 cases of HNSCCs reported annually [2]. The development of HNSCC is strongly associated with long-term tobacco use, excessive consumption of strong alcohols or, especially in the case of oropharyngeal...
tumors, the infection with human papillomavirus (HPV), usually HPV type 16 or 18 \[3\]. The incidence of HNSCC is high in males when compared to females, especially in eastern Europe and India with over 20 males affected per 100,000 individuals \[4\]. Despite relatively easy access for clinical inspection, these tumors are frequently detected at a late stage, when therapeutic options are less effective in curing patients, who are then at a greater risk of the development of recurrent tumors or metastasis \[5\]. Thus, the overall survival rates in this group of patients remain relatively low (~50%), especially when patients are diagnosed with advanced stages of the disease \[6\]. There is a need for novel biomarkers which could improve the clinical management of HNSCC, including better prognosis and disease monitoring. Moreover, the development of new therapeutic options is also necessary for the improvement of treatment outcomes.

Wnt signaling is vital for a plethora of cellular function ranging from homeostasis to the development of mature tissue. Embryonic development also requires Wnt-mediated canonical signaling \[7–9\]. Moreover, Wnt signaling is inevitable for regulating cellular proliferation, apoptosis, metastasis, and migration of cells \[10\]. Of note, Wnt operates through either canonical or non-canonical pathways which are differentiated by β-catenin involvement \[11\]. Cell cycle progression, differentiation, fate determination, and migration are generally orchestrated by canonical Wnt signaling. Altered Wnt/β-catenin signaling has been considered a promoting event for various types of cancers and the oncogenic potential of Wnt signaling has been discussed in numerous cancer types, including breast, pancreatic, colon as well as head and neck \[12\]. The present study investigates the genetic alteration within the Wnt family genes employing in silico approach. The study is first of its kind that reports frequency and type of mutation in genes of the Wnt signaling pathway which provides a clue on the putative association of these genes with HNSCC.

**Methods**

**Sample data set**

The data is obtained from the cBioportal database which is a web source for obtaining, analyzing and exploring genomes. It contains the description of patients from different cohorts and provides information on the genetic alterations among various samples and genes. The Cancer Genome Atlas, TCGA (Firehose Legacy) data set consists of a total of 528 cases of head and neck squamous cell carcinoma, of which 504 samples had sequencing and copy number alteration data. The complete profile of mutated, amplified, deleted genes for each sample has been recorded in the database. Table 1 contains the demographic data of the patients analyzed in the study. Oncoprint data was obtained on submitting user-defined queries on 19 genes of the Wnt family, which was further analyzed for gene expression profile.

**Oncoprint data analysis**

Oncoprint analysis is the shortened and concise summation of the genetic alterations in graphical format. It provides data on gene alterations based on user-defined query on a specific gene or a gene family. The details on frequency distribution of variations in each of the genes, the variant allele frequency, gene deletions, amplifications, insertions, frameshift etc., were recorded (Fig. 1) \[13,14\]. These information were used as baseline to track mutations or variations, gene expression, and survival of patients based on the gene alterations using several other computational tools.

**gnomAD analysis**

The genome aggregation database (gnomAD) hosts information on 125,748 exome sequences and 15,708 whole genome sequences from unrelated individuals sequenced and deposited as part of

---

**Table 1. Demographic details of patients analyzed in the present study (as obtained from the cBioportal site)**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>386</td>
</tr>
<tr>
<td>Female</td>
<td>142</td>
</tr>
<tr>
<td>Mutation count</td>
<td>6–3,181</td>
</tr>
<tr>
<td>Diagnosis age (y)</td>
<td>19–90</td>
</tr>
<tr>
<td>Smoking status</td>
<td></td>
</tr>
<tr>
<td>Smokers</td>
<td>515</td>
</tr>
<tr>
<td>Data not available</td>
<td>12</td>
</tr>
<tr>
<td>Unknown</td>
<td>1</td>
</tr>
<tr>
<td>Alcohol history</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>352</td>
</tr>
<tr>
<td>No</td>
<td>165</td>
</tr>
<tr>
<td>Data not available</td>
<td>11</td>
</tr>
<tr>
<td>Neoplasm histologic grade</td>
<td></td>
</tr>
<tr>
<td>Grade 1</td>
<td>63</td>
</tr>
<tr>
<td>Grade 2</td>
<td>311</td>
</tr>
<tr>
<td>Grade 3</td>
<td>125</td>
</tr>
<tr>
<td>Grade 4</td>
<td>7</td>
</tr>
<tr>
<td>Grade GX</td>
<td>18</td>
</tr>
<tr>
<td>Data not available</td>
<td>4</td>
</tr>
<tr>
<td>Race category</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>452</td>
</tr>
<tr>
<td>African</td>
<td>48</td>
</tr>
<tr>
<td>Asian</td>
<td>11</td>
</tr>
<tr>
<td>American Indian or Alaska native</td>
<td>2</td>
</tr>
<tr>
<td>Data not available</td>
<td>15</td>
</tr>
</tbody>
</table>

---

https://doi.org/10.5808/gi.20065
various disease-specific or population genetic studies. The data source obtained from oncoprint was used to identify whether the variations observed in the present study were novel or reported elsewhere in any other population (Table 2). The exhaustive data source also provides information on minor allele frequencies which will provide a clue as whether the variant identified is a mutation or a polymorphism [15].

**Fig. 1.** (A) Oncoprint analysis depicting gene alterations in the Wnt family of genes. Each of the grey bars represent patients with head and neck squamous cell carcinoma. (B) Frequency of gene amplification among different categories of smokers.

<table>
<thead>
<tr>
<th>Genetic alteration</th>
<th>WNT1</th>
<th>WNT2</th>
<th>WNT3B</th>
<th>WNT3</th>
<th>WNT3A</th>
<th>WNT4</th>
<th>WNT5A</th>
<th>WNT5B</th>
<th>WNT6</th>
<th>WNT7A</th>
<th>WNT7B</th>
<th>WNT8A</th>
<th>WNT8B</th>
<th>WNT9A</th>
<th>WNT9B</th>
<th>WNT10A</th>
<th>WNT10B</th>
<th>WNT11</th>
<th>WNT16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missense mutation (unknown significance)</td>
<td>0.6%</td>
<td>1.6%</td>
<td>2%</td>
<td>1%</td>
<td>1.2%</td>
<td>0.6%</td>
<td>1.4%</td>
<td>3%</td>
<td>1.8%</td>
<td>0.8%</td>
<td>0.8%</td>
<td>0.6%</td>
<td>1.4%</td>
<td>2.2%</td>
<td>0.6%</td>
<td>1.8%</td>
<td>0.4%</td>
<td>5%</td>
<td>1.5%</td>
</tr>
<tr>
<td>Truncating mutation (unknown significance)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Amplification | | | | | | | | | | | | | | | | | | |%
| Deep deletion | | | | | | | | | | | | | | | | | | |%
| No alterations | | | | | | | | | | | | | | | | | | |%

<table>
<thead>
<tr>
<th>Patient smoking history category</th>
<th>[Unknown]</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>No data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. The list of genes, proteins encoded, genetic alterations, loci, frequency of alteration, and variant allele frequency in genes of the *Wnt* signaling pathway.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Protein</th>
<th>Alteration</th>
<th>Loci</th>
<th>% of alteration</th>
<th>Variant allele frequency in tumor sample</th>
<th>gnomAD data</th>
</tr>
</thead>
<tbody>
<tr>
<td>WNT1</td>
<td>Wnt family member 1</td>
<td>Deep deletion</td>
<td>12q13.12</td>
<td>0.6</td>
<td>0.16</td>
<td>Novel</td>
</tr>
<tr>
<td>WNT2</td>
<td>Wnt family member 2</td>
<td>Gene amplification</td>
<td>7q31.2</td>
<td>1.6</td>
<td>0.42</td>
<td>Novel</td>
</tr>
<tr>
<td>WNT2B</td>
<td>Wnt family member 2B</td>
<td>Deep deletion</td>
<td>1p13.2</td>
<td>2</td>
<td>0.63</td>
<td>Novel</td>
</tr>
<tr>
<td>WNT3</td>
<td>Wnt family member 3</td>
<td>Gene amplification</td>
<td>17q21.31–q21.32</td>
<td>1</td>
<td>0.68</td>
<td>rs1483494147</td>
</tr>
<tr>
<td>WNT3A</td>
<td>Wnt family member 3A</td>
<td>Gene amplification</td>
<td>1q42.13</td>
<td>1.2</td>
<td>0.33</td>
<td>Novel</td>
</tr>
<tr>
<td>WNT4</td>
<td>Wnt family member 4</td>
<td>Gene amplification</td>
<td>1q36.12</td>
<td>0.6</td>
<td>0.26</td>
<td>Novel</td>
</tr>
<tr>
<td>WNT5A</td>
<td>Wnt family member 5A</td>
<td>Deep deletion</td>
<td>3p14.3</td>
<td>1.4</td>
<td>0.28</td>
<td>rs750646727</td>
</tr>
<tr>
<td>WNT5B</td>
<td>Wnt family member 5B</td>
<td>Gene amplification</td>
<td>12p13.33</td>
<td>3</td>
<td>0.36</td>
<td>Novel</td>
</tr>
<tr>
<td>WNT6</td>
<td>Wnt family member 6</td>
<td>Deep deletion</td>
<td>2q35</td>
<td>1.8</td>
<td>0.29</td>
<td>rs766635655</td>
</tr>
<tr>
<td>WNT7A</td>
<td>Wnt family member 7A</td>
<td>Gene amplification</td>
<td>3p25.1</td>
<td>0.8</td>
<td>0.23</td>
<td>rs759013954</td>
</tr>
<tr>
<td>WNT7B</td>
<td>Wnt family member 7B</td>
<td>Gene amplification</td>
<td>22q13.31</td>
<td>0.8</td>
<td>0.28</td>
<td>rs751362548</td>
</tr>
<tr>
<td>WNT8A</td>
<td>Wnt family member 8A</td>
<td>Deep deletion</td>
<td>5q31.2</td>
<td>0.6</td>
<td>0.25</td>
<td>Novel</td>
</tr>
<tr>
<td>WNT8B</td>
<td>Wnt family member 8B</td>
<td>Deep deletion</td>
<td>10q24.31</td>
<td>1.4</td>
<td>0.36</td>
<td>Novel</td>
</tr>
<tr>
<td>WNT9A</td>
<td>Wnt family member 9A</td>
<td>Gene amplification</td>
<td>1q42.13</td>
<td>2.2</td>
<td>0.17</td>
<td>Novel</td>
</tr>
</tbody>
</table>

Continued
**Protein protein interaction network analysis**

The STRING database is a collection of known and predicted protein-protein interactions. These interactions could either be direct (physical) and indirect (functional) associations. They are derived from computational predictions and text mining of protein interactions in different organisms and information aggregated from several other primary databases [16]. The gene exhibiting the highest frequency of gene alteration was selected from the entire family and investigated for gene expression and derivation of expression based survival curves for different combinations of parameters such as sex, ethnicity, tumor grade etc., Functional enrichment of the protein network and Kyoto Encyclopedia of Genes and Genomes pathway was derived from the protein protein interaction (PPI) network. The strength score is the ratio between the number of proteins annotated with a term interacting in the network and the number of proteins which is expected to be annotated with this term in a random protein network of the same size. The false discovery rate demonstrates the significance of the enrichment process. The false discovery rate is denoted by p-values which are corrected for multiple testing within each category using Benjamini-Hochberg procedure.

**Gene expression and survival analysis**

The expression of the gene presenting with highest frequency of gene alteration in HNSCC was analyzed using the UALCAN (http://ualcan.path.uab.edu/cgi-bin/TCGA-survival) database. Survival curve analysis based on the tumor grade and expression profile was performed to demonstrate the putative role of Wnt family of genes with HNSCC. Gene expression data is expressed as transcripts per million (TPM) which is a normalization method for RNA-sequencing data. The TPM values which were used for the generation of box-whisker plots were also used to determine the significant difference between the groups. The t test was performed using PERL script with the comprehensive Perl archive network (CPAN) module. Combined survival effect analysis of gene expression and other clinical parameters such as race, sex, tumor grade, and cancer subtypes were assessed using log-rank test that generated a p-value which was further used to indicate statistical significance of survival correlation between groups [17].

**Results**

**Demographic data**

The dataset (TCGA, Firehose Legacy) included in the present study had information on 528 HNSCC samples. The male:female ratio was found to be 2.7:1, with age groups ranging from 19 to 90 years. The number of individuals with the history of smoking and alcohol were roughly around 98% and 67%, respectively. There were five different groups of categories for smoking viz., 1-lifelong non-smoker, 2-current smoker, 3-current informed smoker for
> 15 years, 4-current reformed smoker ≤ 15 years, 5-current reformed smoker, duration not specified. The dataset had samples from patients of American (85.6%), African (9.1%), Asian (2.1%), and American Indian (0.4%) descent. The distribution of patients based on the histologic grade of neoplasm is given in Table 1, of which 59% of patients had grade 2 tumor.

Oncoprint data analysis
The oncoprint data analysis revealed alterations in 19 genes, of which **WNT11** (5%) harbored the highest frequency of gene amplification and deep deletion. When the pattern of amplification was assessed in different groups of smokers a greater frequency of gene amplification was observed in current smokers (n = 6) when compared to other categories (Table 2, Fig. 1B).

gnomAD analysis
The gnomAD analysis revealed several novel and reported variants as demonstrated by the oncoprint data. The variations in **WNT3** (R85Q), **WNTSA1** (R93V, G341S), **WNT6** (R46W, T105M), **WNT7A** (R90C), **WNT10A** (A240V), and **WNT11** (L65P, R202H) genes were reported. Other missense mutations were found to be novel (Table 2). Further investigations are warranted to identify the consequences and association of these variations with HNSCC.

Protein network analysis
The protein interaction network reveals the major interactions of **WNT11** with genes such as **DVL1**, **DVL2**, **DVL3**, **FZD1**, **FZD2**, **FZD3**, **FZD4**, **FZD6**, **FZD7**, **FZD8** which play key roles in governing cell polarity, embryonic development, formation of neural synapses, cell proliferation, and many other processes in developing and adult organisms (Fig. 2). Majority of genes interacting with **WNT11** exhibit significant upregulation of the transcripts with **FZD2** and **DVL3** showing a marked difference (p < 10^-12) in the expression pattern. The expression score is demonstrated by a p-value which denotes the significant difference between two groups of samples viz., normal and HNSCC (Table 3). The functional enrichment analysis showed eleven nodes and 55 edges. The proteins were found to interact more among themselves suggestive of a biologically connected group. The overall PPI enrichment p-value was found to be < 1.0 x 10^-16. Pathways derived from Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis returned predictions which were more inclined towards other cancer types such as basal cell carcinoma, hepatocellular carcinoma, etc. (Table 4).

Gene expression and survival analysis
The gene expression profile of **WNT11** between normal and primary tumor samples revealed a significant difference with a p-value of 3.043 x 10^-3. The relative expression profile of **WNT11** gene in different grades of HNSCC also returned significant values between normal vs. grade 2, normal vs. grade 4, grade 1 vs. grade 4, grade 2 vs. grade 4, and grade 3 vs. grade 4 (Fig. 3A). The expression pattern of **WNT11** gene produced significant difference between normal and female HNSCC subjects (p = 2.169 x 10^-5). Significant difference was not observed between groups normal.

Fig. 2. The proteins network interaction of **WNT11** gene.
vs. male and male vs. female (Fig. 3B). Although the present observation does not confirm sex predilection of WNT11 gene expression with HNSCC, a significant difference in the survival probability between male and female subjects was observed with low/medium level expression ($p = 0.021$). Furthermore, female subjects who presented with a low/medium level expression exhibited low survival probability when compared to male subjects. A $p$-value less than 0.05 was considered to be significant (Fig. 4).

The expression profile of WNT11 was checked in other types of squamous cell carcinoma such as lung (LUSC) and esophageal cancers (ESCC). Although both LUSC and ESCC produced a significant difference with respect to normal and primary tumors, WNT11 expression was upregulated in ESCC cases which was in consonance with HNSCC type. In both the cancer types, there was upregulation of WNT11 gene, whereas in case of LUSC WNT11 expression was downregulated. The receptor proteins interacting with WNT11 have also shown significant differences based on the levels of gene expression in different sex groups (data not shown). These results add to the association of WNT11 gene alterations with HNSCC.

### Discussion

Genetic variations such as single nucleotide variants and copy number variants have long been associated with oral cancer and other devastating diseases. Wnt signaling is vital for a plethora of cellular function ranging from homeostasis to the development of mature tissue. Embryonic development also requires Wnt-mediated canonical signaling [7–9]. Moreover, Wnt signaling is inevitable for regulating cellular proliferation, apoptosis, metastasis, and migration of cells [10]. Of note, Wnt operates through either canonical or non-canonical pathways which are differentiated by β-catenin.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Protein</th>
<th>Expression profile</th>
<th>Expression score ($p$-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FZD1</td>
<td>Frizzled-1; receptor for Wnt proteins</td>
<td>Upregulated</td>
<td>$2.237 \times 10^{-9}$</td>
</tr>
<tr>
<td>FZD2</td>
<td>Frizzled-2; receptor for Wnt proteins</td>
<td>Upregulated</td>
<td>$&lt;10^{-12}$</td>
</tr>
<tr>
<td>FZD3</td>
<td>Frizzled-3; receptor for Wnt proteins</td>
<td>Upregulated</td>
<td>$1.372 \times 10^{-1}$</td>
</tr>
<tr>
<td>FZD4</td>
<td>Frizzled-4; receptor for Wnt proteins</td>
<td>Downregulated</td>
<td>$3.639 \times 10^{-2}$</td>
</tr>
<tr>
<td>FZD6</td>
<td>Frizzled-6; receptor for Wnt proteins</td>
<td>Upregulated</td>
<td>$1.624 \times 10^{-12}$</td>
</tr>
<tr>
<td>FZD7</td>
<td>Frizzled-7; receptor for Wnt proteins</td>
<td>Upregulated</td>
<td>$8.622 \times 10^{-1}$</td>
</tr>
<tr>
<td>FZD8</td>
<td>Frizzled-8; receptor for Wnt proteins</td>
<td>Upregulated</td>
<td>$2.518 \times 10^{-1}$</td>
</tr>
<tr>
<td>DVL1</td>
<td>Segment polarity protein dishevelled homolog DVL-1</td>
<td>Upregulated</td>
<td>$4.600 \times 10^{-2}$</td>
</tr>
<tr>
<td>DVL2</td>
<td>Segment polarity protein dishevelled homolog DVL-2</td>
<td>Upregulated</td>
<td>$1.624 \times 10^{-12}$</td>
</tr>
<tr>
<td>DVL3</td>
<td>Segment polarity protein dishevelled homolog DVL-3</td>
<td>Upregulated</td>
<td>$&lt;10^{-12}$</td>
</tr>
</tbody>
</table>

*Differentially expressed genes with a statistically significant gene expression.

<table>
<thead>
<tr>
<th>KEGG pathways</th>
<th>Strength</th>
<th>False discovery rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal cell carcinoma</td>
<td>2.49</td>
<td>$1.76 \times 10^{-16}$</td>
</tr>
<tr>
<td>Melanogenesis</td>
<td>2.3</td>
<td>$8.06 \times 10^{-15}$</td>
</tr>
<tr>
<td>Signaling pathways regulating pluripotency stem cells</td>
<td>2.15</td>
<td>$1.93 \times 10^{-23}$</td>
</tr>
<tr>
<td>Wnt signaling pathway</td>
<td>2.14</td>
<td>$2.11 \times 10^{-23}$</td>
</tr>
<tr>
<td>Gastric cancer</td>
<td>2.12</td>
<td>$2.26 \times 10^{-23}$</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>2.12</td>
<td>$2.26 \times 10^{-23}$</td>
</tr>
<tr>
<td>mTOR signaling pathway</td>
<td>2.12</td>
<td>$2.26 \times 10^{-23}$</td>
</tr>
<tr>
<td>Cushing’s syndrome</td>
<td>2.11</td>
<td>$2.26 \times 10^{-23}$</td>
</tr>
<tr>
<td>Hippo signaling pathway</td>
<td>2.11</td>
<td>$2.26 \times 10^{-23}$</td>
</tr>
<tr>
<td>Hepatocellular carcinoma</td>
<td>2.08</td>
<td>$3.38 \times 10^{-23}$</td>
</tr>
<tr>
<td>Notch signaling pathway</td>
<td>2.05</td>
<td>$3.07 \times 10^{-6}$</td>
</tr>
<tr>
<td>HTLV-1 infection</td>
<td>1.89</td>
<td>$2.96 \times 10^{-71}$</td>
</tr>
<tr>
<td>Proteoglycans of cancer</td>
<td>1.86</td>
<td>$2.27 \times 10^{-14}$</td>
</tr>
<tr>
<td>Human papillomavirus infection</td>
<td>1.79</td>
<td>$3.50 \times 10^{-70}$</td>
</tr>
<tr>
<td>Pathways in cancer</td>
<td>1.58</td>
<td>$6.22 \times 10^{-18}$</td>
</tr>
</tbody>
</table>

KEGG, Kyoto Encyclopedia of Genes and Genomes; mTOR, mammalian target of rapamycin.
Fig. 3. (A) Box-Whisker plot showing relative expression profile of \(WNT11\) gene in different grades of head and neck squamous cell carcinoma (HNSCC). The X axis denotes The Cancer Genome Atlas (TCGA) samples and Y axis denotes the transcripts per million values. The comparison of gene expression patterns between different grades of HNSCC returned significant values between normal vs. grade 2 \((p = 3.8 \times 10^{-2})\), normal vs. grade 4 \((p = 4.3 \times 10^{-3})\), grade 1 vs. grade 4 \((p = 7.11 \times 10^{-3})\), grade 2 vs. grade 4 \((p = 2.54 \times 10^{-11})\), and grade 3 vs. grade 4 \((p = 1.9 \times 10^{-3})\). A p-value less than 0.05 was considered to be significant. (B) Box-Whisker plot showing relative expression profile of \(WNT11\) gene in male and female HNSCC subjects. The X axis denotes the TCGA samples and Y axis denotes the transcripts per million values. The comparison of gene expression patterns between male and female viz., normal vs. male \((p = 9.923 \times 10^{-2})\), normal vs. female \((p = 2.169 \times 10^{-2})\), male vs. female \((p = 2.12 \times 10^{-1})\). A p-value less than 0.05 was considered to be significant.

Fig. 4. Kaplan–Meier plots showing the association of \(WNT11\) gene expression in combination with the sex with head and neck squamous cell carcinoma patient’s survival. The x-axis represents time in days and y-axis shows the survival probability. The blue line indicates low/medium expression in male patients and the red line indicates low/medium level expression of the \(WNT11\) gene in female patients. A significant difference in the survival probability was observed between the two groups \((p = 0.021)\). Female subjects with a low/medium level expression presented with a low survival probability when compared to male subjects. A p-value less than 0.05 was considered to be significant.
in involvement [11]. Both pathways are activated by the binding of Wnt protein to the Frizzled (Fzd) seven transmembrane receptor. The fundamental difference between these two pathways is the involvement of β-catenin [18,19].

Non-canonical Wnt signaling pathways, which are independent of β-catenin, rely on the signal transduction of Wnt through Fzd as well as its co-receptors such as receptor tyrosine kinase-like orphan receptor 2 or receptor-like tyrosine kinase [20]. On the other hand, the canonical Wnt signaling pathway, also known as Wnt/β-catenin signaling pathway, involves the activation of cytoplasmic β-catenin signaling cascades upon Wnt signal transduction at the cell membrane [21]. Altered Wnt/β-catenin signaling has been considered a promoting event for various types of cancers. Canonical Wnt signaling pathway acts as a master regulator for a wide range of biological effects through up- or down-regulation of genes that act as direct effectors, transcription regulators, or other signaling pathway regulators. Thus, Wnt target gene expression can either directly or indirectly activate cell cycle progression, cell proliferation, cell differentiation, cell migration, inhibit apoptosis, and regulate embryonic development.

The involvement of canonical Wnt signaling pathway in the formation of HNSCC has also been examined in several experimental studies. Evidence of the association between the pathway and HNSCC was first discovered through cDNA arrays on patient samples. The study found that Fzd, Fzd homolog 3, and Dvl homolog genes, which are functionally important in the canonical Wnt signaling pathway, were highly expressed in HNSCC by two to five fold when compared to normal tissues [22]. Furthermore, another study has also demonstrated that the gene expression levels of Wnts, particularly Wnt11 and Wnt10b, were markedly higher by 17 and 3-fold, respectively, in HNSCC cells compared to normal oral squamous epithelial cells [23]. Over-expression of these major components in HNSCC cells compared to normal cells clearly signifies the involvement of abnormal activation of Wnt signaling cascades in HNSCC. These reports were in agreement with the observation made in the present study wherein the up-regulation of WNT11 correlated with concomitant increase in the gene expression level of interacting genes. Recently, the role of canonical Wnt signaling pathway in regulating self-renewal of HNSCC cancer stem cells is also being emphasized in several studies. Abnormal activation of the pathway has been correlated with increased proliferation and thus self-renewal of cancer stem cells in HNSCC [24]. These observations were confirmed by our computational analysis wherein the KEGG pathways arising out of functional enrichment analysis revealed involvement of WNT11 in multiple pathways viz., pathways associated with cancer, infections, syndrome, and signaling process regulating pluripotency of stem cells.

Although gene expression profiling is well documented in case of WNT family of genes, the effect of gene alterations such as mutations, deletions, and copy number variations upon the expression of the genes in the family is scarce or limited. The present study throws light on those alterations which might act as putative drivers in establishing tumorigenesis. Numerous novel variants and reported variants were identified in HNSCC patients. The potential role of these variants in the disease process is yet to be explored. Similar studies have already been carried out to unravel the potential markers with putative association with HNSCC. The survival curve analysis provides cues on the prognostic significance of these markers in relation to the disease [25,26].

Computational tools provide a cost-effective alternative for analyzing the genetic alterations especially in a complex disorder such as cancer. Although the crosstalk between multiple signaling pathways play a vital role in the development of cancer, further understanding of one important molecular mechanism, which is the canonical Wnt signaling pathway, is critical as targeting the pathway could be a promising approach in eradicating the treatment failure and relapse in HNSCC. With all the pros addressed, the study design also suffers certain drawbacks such as population bias and reproducibility in other ethnic groups. The identification of polymorphic variants and chromosomal abnormalities would aid us in preparing a panel of markers intended for use as early diagnostic leads. These markers can be further validated using genotyping methods or next-generation sequencing approaches to derive a strong association with the disease phenotype.

**ORCID**

Jain Aditya: https://orcid.org/0000-0003-0369-7248  
A. S. Smiline Girija: https://orcid.org/0000-0001-6106-0064  
A. Paramasivam: https://orcid.org/0000-0001-6117-1178  
J. Vijayashree Priyadharsini: https://orcid.org/0000-0001-7884-5466

**Authors’ Contribution**

Conceptualization: JVP, AP. Data curation: JVP, AP. Formal analysis: JVP, AP. Funding acquisition: JVP, AP. Writing - original draft: JA, ASSG. Writing - review & editing: JVP, AP.

**Conflicts of Interest**

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

The authors are grateful to all the consort and groups involved in the compilation of data from patients for public use. Our sincere thanks also go to all the patients who have indirectly contributed to the scientific community through providing consent for sharing their data for research use.

The authors Dr. Vijayashree Priyadharsini J. and Dr. Paramasivam A. thank the Science and Engineering Research Board (SERB), Government of India for the financial support rendered through the core research grant (CRG) (No. CRG/ 2019/003756) and (EEQ) (No.EEQ/2019/00411).

References


Introduction

Cancer is a multifactorial disease that gets influenced by several factors including genetic change, lifestyle, viral infection, bacterial infection and epigenetic effects. Cancer causes an elevated physical toll along with amplified psychological stress that disrupts homeostasis [1]. In terms of fatality, cancer undoubtedly falls in the category of diseases that accounts for high death cases and stands second following cardiac diseases. Every year about 1 in 6 deaths occur due to cancer globally which is about 10 million deaths per year [2,3]. Cancer’s effect on the older population (aged 70 or above) is perniciously leading to a high fatality rate which was projected to be 14.4% in older males and 9.6% in older females in 2019 [4].
Cancer has seven hallmarks which include: selective growth and proliferative advantage, altered stress response favoring overall survival, apoptosis, invasion and metastasis, metabolic reprogramming, an abetting microenvironment, and immune modulation [5]. When it comes to aiding both normal and abnormal cell proliferation, angiogenesis plays a vital role [6]. Angiogenesis refers to construction of new capillary blood vessels from pre-existing blood vessels to supply sufficient molecular oxygen, nutrients and other essentials to the proliferating cells. Through the process of angiogenesis, cellular waste and debris are also removed hence angiogenesis or vascularization has a significant role in maintaining cell viability, development, and proliferation [7-9]. Tumor cell proliferation is pronouncedly dependent on angiogenesis because when tumors are devoid of nascent blood vessels to supply them with the necessary factors required for proliferation, they remain benign and ultimately die from necrosis and apoptosis [7,10,11]. Angiogenesis also amplifies the cancer state by providing the abnormal cells with a network to carry out metastasis and corresponding secondary infection [12]. However, several factors either upregulate or downregulate angiogenesis hence, the process is susceptible to being either positively or negatively altered by activators and inhibitors [7,13].

Among the activators of angiogenesis, vascular endothelial growth factors (VEGFs) play a fundamental role as signaling proteins that stimulate new blood vessel formation by vasculogenesis and angiogenesis throughout our lifetime [14,15]. Usually these signaling proteins bind to specific VEGF receptors which then elicit a cellular response of vessel formation [16].

The VEGF proteins are made up of five known sub-families namely VEGF-A (the highly conserved founding member), VEGF-B, VEGF-C, VEGF-D (also known as e-fos-induced growth factor) and the viral VEGF-Es encoded by strains D1701, NZ2, and NZ7 of the parapoxvirus Orf (which causes pustular dermatitis) [17]. VEGF-A is the prototypical member of a family of associated growth factors that includes placental growth factor [17]. The different classes of VEGFs carry out different functions in relation to angiogenesis [18]. The VEGF class that gains the most attention in terms of research is the VEGF-A class as it is thought to be the primary class of VEGF that promotes systemic primary blood vessel development [17]. The discrete functions of VEGF-A that have been identified are follows: increasing endothelial cell migration, increasing permeability of blood vessels, and maintenance of uniform neovascularization [17]. VEGF-B takes embryonic vasculogenesis to completion in combination with VEGF-A [19]. VEGF-C was found to uniquely contribute to lymphomagenesis as it binds to the VEGF receptor (VEGFR)-3 receptor and VEGF-D plays a role in pulmonary angiogenesis through binding to the VEGFR-3 receptor as well. There are also two other classes of VEGF namely VEGF-E and VEGF-F [17]. VEGF-E is encoded by viruses that synergistically along with virus particles such as IL-10 helps wound healing as found in mice and for the VEGF-F case, it is usually isolated and found in snake venom [20].

As far as the mechanism goes for VEGF binding, VEGF-A can bind with either of the corresponding receptors VEGFR-1 or VEGFR-2 located on the surface of the endothelial cells [21]. However, VEGF-A most commonly binds to the VEGFR-2 to stimulate vessel growth [22]. The other receptor VEGFR-3 is specific to another class of VEGF (VEGF-C) and it is thought that the pathway upon binding that receptor stimulates the proliferation of lymphatic cells [21]. All of these receptors are tyrosine kinase receptors which causes dimerization and activation by transphosphorylation which ultimately results in vessel formations [23].

Anti-angiogenic drugs and in particular anti-VEGF agents have entered the clinical armamentarium against cancer. However, a number of complications in terms of vascular events have been found succeeding treatment. The vascular endothelial growth factor signaling pathway (VSP) inhibitors include antibodies that work both extracellularly and intracellularly on VEGF and VEGFR, respectively. VSP inhibitors have possibilities of eliciting damage to endothelial lining due to depleted endothelial cell turnover [24]. Inhibitor Mediated vascular anomalies also include arterial and/or venous thrombosis, and renal vascular injury [25]. Bevacizumab retains the highest frequency of bleeding complications, in particular epistaxis, hemoptysis, and gastrointestinal bleeding. Although a higher incidence of severe hemorrhages has not been consistently demonstrated during the treatment with bevacizumab, mild bleeding episodes appear clearly increased in the experimental arm of most trials. Trials with other small-molecule tyrosine kinase inhibitors like sunitinib or sorafenib showed an overall lower rate of bleeding complications, but still significantly higher than the control arm in many cases [26].

The mechanisms of bleeding induced by anti-VEGF agents are complex and not yet fully clarified: the main hypothesis is that VEGF could promote endothelial cell survival and integrity in the adult vasculature and its inhibition may decrease the renewal capacity of damaged endothelial cells [27]. Management of bleeding in patients treated with anti-VEGF agents is a challenging task because this complication is at least in part inherent to the efficacy of the drug and because there is also an increased risk of thrombosis, both arterial and venous. So far, only a few preliminary data are available on a strategy to prevent hemorrhage and thrombotic events [28]. However, previous studies have concluded that the deleterious effects of anti-VEGF drugs are not overt during the first stages of administration because of VEGF’s intrinsic roles rele-
Venant to vascular protection [29]. If subsidiary vascular thrombosis and other vascular complications can be minimized, VEGF inhibitors, if not of the conventional kind, can still be favorable in depleting the prognosis of tumor cells through blocking angiogenesis [30].

VEGF molecules have become a choice of interest for cancer therapy among scientists. Using virtual screening (VS) to find inhibitors against VEGFs from libraries of small molecules like alkaloids can be a good approach to inhibit angiogenesis in recent years [31]. VS refers to a computer-based technique used to identify drugs from libraries of small molecules that may be highly likely to interact with a certain enzyme or protein based receptor.

The aim of this study was to select alkaloids having similar binding capabilities as VEGF inhibitors to propose possible therapeutic candidates against tumor angiogenesis which might minimize vascular complications manifested by the current drugs. We curated a library of alkaloids to select ligands having similar binding affinity to that of anti-VEGF drugs. Since alkaloids have minimal side effects and are easier to extract, this study aimed to provide a preliminary list of potential alkaloids that can be used to develop highly effective therapeutics against VEGF molecules that can work against cancer.

**Methods**

**Protein retrieval**

The X-ray crystallographic protein structure of the major regulators of angiogenesis, VEGF-A (302aa, PDB Code: 1VPF), VEGF-B (207aa, PDB Code: 2C7W), VEGF-C (419aa, PDB Code: 2X1X), VEGF-D (354aa, PDB Code: 2XV7) were retrieved from the RCSB Protein Data Bank in PDB format which were going to be used as targets for carrying out the docking experiments. Resolutions of 2.5 Å, 2.48 Å, 3.1 Å, and 2.9 Å were employed for VEGF-A, VEGF-B, VEGF-C, and VEGF-D, respectively.

**Prediction of active site**

In proteins, active sites are clefts formed by specific combinations of amino acids that facilitate the binding of ligands to a target protein often initiating or blocking a chain of reactions. Identification of the residues that make up the active site has a range of applications in molecular docking and de novo drug designing [32]. Computed atlas of surface topography of proteins (CASTp) was used in active site residue analysis [33,34]. CASTp works using Swiss-Prot mapping method as well as Online Mendelian Inheritance in Man (OMIM) mapping method to prognosticates specific amino acid positioning within a protein surface [35,36].

**Ligand retrieval and preparation**

Initially, more than 300 alkaloid compounds were retrieved from different literature sources as control ligands for the purpose of inhibiting VEGFs based on their natural sources, few or no side effects as therapeutic agents and so on. These alkaloids were acquired from PubChem [37] and ZINC databases were used as ligands [38]. The compounds were downloaded in sdf or structural data file format and then converted to pdb format using OPEN Babel converter [39]. In the next step, these ligands were energy minimized and torsion angle of these molecules were changed for flexibility or freedom of movement. Currently, available known drugs were also retrieved and optimized in silico to be used as a ligand molecule for molecular docking analysis.

**Molecular docking**

Structure-based virtual screening was done using molecular docking as it is a viable and effective process for the identification of hits or potential drugs and thus plays a major role in enhancing the lead recognition stage of the pharmaceutical sectors. VS by docking was selected because it is free, easy to use and can take advantage of numerous core processors in addition to having much more orderly search of the probable energy surfaces. VS was performed against the energy minimized models of VEGF-A, VEGF-B, VEGF-C, and VEGF-D using Autodock to carry out automated docking of ligand molecules to their macromolecular receptors. Autodock creates the three binding energy phases: intramolecular energy, internal energy of ligand, and torsional free energy [40]. The final docked energy is determined from the summation of intermolecular energy and internal energy of the ligand. Autodock tools were employed to construct the input pdbqt file for VEGF-A, VEGF-B, VEGF-C, and VEGF-D and also to set up the size and the center of the grid box. All water molecules, cofactors, and ligands were removed from the protein structure and then checked for polar hydrogen atoms in the macromolecules. Afterward, torsion bonds of the ligands were selected. The binding energy of macromolecules coordinate were evaluated by a three dimensional grid box of $80 \times 40 \times 80$ (num.grid points in xyz) and grid center $5.958 \times 2.623 \times 28.642$ (xyz-coordinates), $40 \times 60 \times 44$ (num.grid points in xzy) and grid center $-43.699 \times -24.709 \times -0.6$ (xyz-coordinates), $76 \times 50 \times 70$ (num.grid points in xyz) and grid center $-34.28 \times 2.751 \times 13.25$ (xyz-coordinates) and $30 \times 60 \times 50$ (num.grid points in xzy) and grid center $-30.389 \times -36.541 \times -6.255$ (xyz-coordinates) were created for VEGF-A, VEGF-B, VEGF-C, and VEGF-D respectively (unit of the dimensions, Å). The bound ligand and actual target docking site were represented based on the calculation of the grid map and the final docking complex was visualized in BIOVIA Discovery Studio.
Bioavailability and ADME/Tox test
Absorption, distribution, metabolism, excretion, and toxicity (ADME/Tox) explain in detail the kinetics of drug exposure to the body tissues and pharmacological effects of the compounds. ADME/Tox was assessed with the help of an online server, pre-ADMET [42]. Besides ADME, drug toxicity and its side effects of the compounds, a major concern, was estimated using OSIRIS program [43] and ADME/Tox filter with FAF-Drug-2 [44]. ADME/Tox filter with FAF-Drug-2 also eradicates PAINS (Pan Assay Interference Compounds) which provides further refining steps in the selection process. They provide weak options for drug development but can provide data that in isolation may be evocative of a particular and optimizable fit for potential drugs.

Results and Discussion
Cancer occupies the maximum landscape among the diseases and disorders that are found to be in frequent prevalence, due to its mortality rates as well as multiple other collateral risk factors. Often, cancer is detected at a stage beyond the scopes of cure by therapeutics because of its ability to blend in well with normal cells, which is why conventional treatment measures fail to provide a permanent cure for cancer patients [45,46]. Discovering and developing novel therapeutics against different types of cancer is quite difficult, merely because of the seven hallmarks that cancer imposes [46]. However, like multiple other diseases, different types of cancers have common clinical manifestations across individuals and if these mechanisms and common manifestations can be addressed using drugs, developing effective and consistent treatment methods against cancer will be possible. Among the hallmarks of cancer, angiogenesis is of immense importance and is common in all types of cancers [47]. As angiogenesis is regulated by VEGF-mediated signaling pathways, blocking VEGF action could stop angiogenesis and by extension, halt the growth of cancer cells, which is why VEGF is a suitable target for cancer therapy [48]. Different VEGF families with their receptors and their respective functions are listed in Table 1 and the crystal 3D structures are shown in Fig. 1. In this study, to scrutinize the effectiveness of alkaloids against cancer therapy in comparison with existing drugs that act upon VEGF blocking, we analyzed multiple alkaloids to identify potential inhibitors of multiple VEGFs using computational approaches of protein-ligand docking. Because VS is a widely followed procedure for de novo drug design, it helps in identifying a library of potential inhibitors which can further be analyzed in terms of binding affinity using molecular docking.

Analysis of active site
Possible binding sites for different VEGFs were identified using the CASTp server [34]. The amino acid residues involved in binding pockets are given in Supplementary Table 1. The possible binding residues that were found to be involved in the interaction with lead inhibitors. As calculated by CASTp the binding pocket of VEGF-A, VEGF-B, VEGF-C, and VEGF-D has a volume of 122.264˚ A, 90.134˚ A, 291.758˚ A, and 14.779˚ A and surface area of 161.609, 149.220, 239.334, and 44.37, respectively.

Ligand preparation
Based on ADME properties through VS of 20 compounds were shortlisted to create the ligand library with potential candidates

Table 1. Different types of VEGFs and their functions

<table>
<thead>
<tr>
<th>VEGF family member</th>
<th>Receptor</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>VEGF-A</td>
<td>VEGFR-1</td>
<td>Angiogenesis</td>
</tr>
<tr>
<td></td>
<td>VEGFR-2</td>
<td>Vasodilation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chemotactic</td>
</tr>
<tr>
<td>VEGF-B</td>
<td>VEGFR-1</td>
<td>Embryonic angiogenesis</td>
</tr>
<tr>
<td>VEGF-C</td>
<td>VEGFR-2</td>
<td>Lymphangiogenesis</td>
</tr>
<tr>
<td></td>
<td>VEGFR-3</td>
<td></td>
</tr>
<tr>
<td>VEGF-D</td>
<td>VEGFR-2</td>
<td>Lymphangiogenesis</td>
</tr>
<tr>
<td></td>
<td>VEGFR-3</td>
<td></td>
</tr>
<tr>
<td>VEGF-E</td>
<td>VEGFR-2</td>
<td>Angiogenesis</td>
</tr>
</tbody>
</table>

VEGF, vascular endothelial growth factor.
Fig. 2. The 2D structure of 20 alkaloid compounds.

Table 2. Docking results of different drugs with VEGFs

<table>
<thead>
<tr>
<th>No.</th>
<th>Drug</th>
<th>VEGF-A</th>
<th>VEGF-B</th>
<th>VEGF-C</th>
<th>VEGF-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Iclusig, Ponatinib</td>
<td>−10.8</td>
<td>−9.4</td>
<td>−10.0</td>
<td>−9.1</td>
</tr>
<tr>
<td>2</td>
<td>Votrient, Pazopanib</td>
<td>−10.5</td>
<td>−9.2</td>
<td>−8.8</td>
<td>−7.3</td>
</tr>
<tr>
<td>3</td>
<td>Adriaamycin, Adriamycin</td>
<td>−10.0</td>
<td>−9.8</td>
<td>−9.8</td>
<td>−8.5</td>
</tr>
<tr>
<td>4</td>
<td>Cometriq, Cabozantinib</td>
<td>−9.7</td>
<td>−8.8</td>
<td>−8.6</td>
<td>−7.4</td>
</tr>
<tr>
<td>5</td>
<td>Inlyta, Axitinib</td>
<td>−9.3</td>
<td>−8.4</td>
<td>−8.8</td>
<td>−7.9</td>
</tr>
<tr>
<td>6</td>
<td>Stivarga, Regorafenib</td>
<td>−9.0</td>
<td>−9.6</td>
<td>−9.3</td>
<td>−8.3</td>
</tr>
<tr>
<td>7</td>
<td>Cabometyx, Cabozantinib</td>
<td>−9.0</td>
<td>−8.3</td>
<td>−9.5</td>
<td>−7.7</td>
</tr>
<tr>
<td>8</td>
<td>Lenvima, Lenvatinib</td>
<td>−8.4</td>
<td>−7.6</td>
<td>−7.3</td>
<td>−6.5</td>
</tr>
<tr>
<td>9</td>
<td>Sutent, Sunitinib</td>
<td>−8.3</td>
<td>−7.4</td>
<td>−7.7</td>
<td>−7.6</td>
</tr>
<tr>
<td>10</td>
<td>Nexavar, Sorafenib</td>
<td>−8.3</td>
<td>−8.5</td>
<td>−9.2</td>
<td>−7.4</td>
</tr>
</tbody>
</table>

AutoDock Vina scores are in kcal/mol.
VEGF, vascular endothelial growth factor.
We screened the selected compounds and selected those which exhibited preferable binding energy clusters \[ 49 \]. Protein-substrate binding gives us insights into prediction and ranking of compounds on the basis of their binding and interactions \[ 50 \].

**Molecular docking analysis**

Among the currently available drugs against VEGFs, Ponatinib showed the highest binding free energy (Table 2) which were $-10.8$ kcal/mol, $-9.4$ kcal/mol, $-10.0$ kcal/mol, and $-9.1$ kcal/mol against VEGF-A, VEGF-B, VEGF-C, and VEGF-D, respectively. Hydrogen bonds, electrostatic bonds, and hydrophobic bonds were majorly formed with VEGFs and the interaction sites are shown in Table 3. Because ponatinib, among the drugs that are commonly used for angiogenesis inhibition exhibited a preferable and considerable binding affinity, it was used as the positive control. Now, although ponatinib is a widely used drug, it isn’t devoid of side effects. The most common adverse effects that can occur due to consistent ponatinib usage are thrombocytopenia and pancreatitis. To avoid these additional drawbacks, our aim was to look for alternative therapeutic compounds with minimum to no side effects. From the 20 ligands, we selected potential candidates for VEGF inhibition in Table 4. Among three ligands: moronic acid, cadambagenic acid, and masilinic acid exhibited higher binding energies with subsequent VEGFs which were more than those shown by ponatinib (Table 5). During docking with VEGF-A, Moronic acid formed three conventional hydrogen bonds with C:-Glu30, C:-Thr31, and D:-Thr31 and three hydrophobic bonds with C:-Ile29, D:-Ile29, and D:-Leu32. Most of the bonds were formed in the active site of the protein. With VEGF-B Moronic acid formed a hydrogen bond with A:-Val32 and six hydrophobic bonds with the site A:-Val31, A:-VAL32, B:-ARG29, B:-VAL31, and B:-VAL32. These bonds were formed on the same active site similar to that of ponatinib; however, the binding energy generated from moronic acid-VEGF-B binding was higher than that generated from the binding with ponatinib. Docking with VEGF-C, moronic acid generated only six hydrophobic bonds at E:-Trp126. Finally with VEGF-D five hydrophobic bonds at A:-Ala121, A:-Phe131, and A:-Pro135 were formed. These bind strongly with the active site residues of the VEGFs signaling molecule so it can’t readily bind with its receptor (Fig. 3) and consequently block the signal transduction for angiogenesis. We also assessed their stability and observed that all bonds were of very short distance that indicates the intense bonding strength.

**ADME/Tox test analysis**

ADME/Tox test analysis was carried out to assess the molecular properties, carcinogenicity and oral toxicity of the selected alkaloid candidates for VEGF inhibition (Tables 6 and 7). Their permeability to different cells and the blood brain barrier were also analyzed because all in all, these are the major stakeholders in drug discovery. The results obtained from these assessments validated the use of these alkaloids in effecting cancer treatment.

---

**Table 3. Nonbonding interactions of ponatinib with VEGFs**

<table>
<thead>
<tr>
<th>VEGF</th>
<th>&quot;Bonds Donor (distance, Å) acceptor (bond type)&quot;</th>
<th>Hydrophobic bond</th>
</tr>
</thead>
</table>

Pose predicted by AutoDockVina where, HB, conventional hydrogen bond; CHB, carbon hydrogen bond; E, electrostatic; A, alkyl; Pi-A, pi-alkyl; A-Pi, amide-pi. VEGF, vascular endothelial growth factor.
Table 4. Docking results of different alkaloids with VEGFs

<table>
<thead>
<tr>
<th>No.</th>
<th>Alkaloid</th>
<th>VEGF-A</th>
<th>VEGF-B</th>
<th>VEGF-C</th>
<th>VEGF-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Moronic acid</td>
<td>-12.9</td>
<td>-13.2</td>
<td>-11.9</td>
<td>-12.2</td>
</tr>
<tr>
<td>2</td>
<td>Cadambagenic acid</td>
<td>-12.5</td>
<td>-12.2</td>
<td>-11.5</td>
<td>-11.5</td>
</tr>
<tr>
<td>3</td>
<td>Masilinic acid</td>
<td>-12.4</td>
<td>-12.6</td>
<td>-11.5</td>
<td>-12.0</td>
</tr>
<tr>
<td>4</td>
<td>Nortripterifordin</td>
<td>-10.4</td>
<td>-9.7</td>
<td>-10.1</td>
<td>-10.0</td>
</tr>
<tr>
<td>5</td>
<td>Michellamine</td>
<td>-10.2</td>
<td>-10.1</td>
<td>-9.9</td>
<td>-8.9</td>
</tr>
<tr>
<td>6</td>
<td>Cadamble</td>
<td>-10</td>
<td>-9.6</td>
<td>-9.3</td>
<td>-7.8</td>
</tr>
<tr>
<td>7</td>
<td>Repandusinic acid</td>
<td>-9.8</td>
<td>-10.4</td>
<td>-9.3</td>
<td>-9.2</td>
</tr>
<tr>
<td>8</td>
<td>3a-Dihydrocadamine</td>
<td>-9.6</td>
<td>-9.0</td>
<td>-9.4</td>
<td>7.4</td>
</tr>
<tr>
<td>9</td>
<td>Hinokiiflavone</td>
<td>-9.6</td>
<td>-8.9</td>
<td>-9.1</td>
<td>-8.1</td>
</tr>
<tr>
<td>10</td>
<td>Robustaflavone</td>
<td>-9.4</td>
<td>-8.8</td>
<td>-9.0</td>
<td>-8.0</td>
</tr>
</tbody>
</table>

AutoDock Vina scores are in kcal/mol.
VEGF, vascular endothelial growth factor.

Table 5. Molecular docking nonbonding interactions of moronic acid with VEGFs

<table>
<thead>
<tr>
<th>VEGF</th>
<th>&quot;Bonds&quot; Donor (distance, Å) acceptor (bond type)</th>
<th>Hydrogen bond</th>
<th>Hydrophobic bond</th>
</tr>
</thead>
<tbody>
<tr>
<td>VEGF-A</td>
<td>d:LIG1:O (3.095) C:GLU30:O (HB)</td>
<td>C:ILE29 (4.848) d:LIG1 (A)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D:THR31:CA (3.185) d:LIG1:O (CHB)</td>
<td>d:LIG1 (5.243) D:ILE29 (A)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A:VAL32 (4.752) d:LIG1 (A)</td>
<td>B:ARG29 (5.031) d:LIG1 (A)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B:VAL31 (3.847) d:LIG1 (A)</td>
<td>B:VAL31 (4.271) d:LIG1 (A)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B:VAL32 (4.060) d:LIG1 (A)</td>
<td>B:VAL32 (4.060) d:LIG1 (A)</td>
<td></td>
</tr>
<tr>
<td>VEGF-C</td>
<td>E:TRP126 (4.423) d:LIG1 (Pi-A)</td>
<td>E:TRP126 (3.690) d:LIG1 (Pi-A)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E:TRP126 (3.961) d:LIG1 (Pi-A)</td>
<td>E:TRP126 (3.961) d:LIG1 (Pi-A)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E:TRP126 (4.740) d:LIG1 (Pi-A)</td>
<td>E:TRP126 (4.445) d:LIG1 (Pi-A)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E:TRP126 (3.799) d:LIG1 (Pi-A)</td>
<td>E:TRP126 (3.799) d:LIG1 (Pi-A)</td>
<td></td>
</tr>
<tr>
<td>VEGF-D</td>
<td>A:ALA121 (3.871) d:LIG1 (A)</td>
<td>A:PRO135 (5.167) d:LIG1 (A)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A:PHE131 (3.786) d:LIG1 (Pi-A)</td>
<td>A:PHE131 (4.910) d:LIG1 (Pi-A)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A:PHE131 (3.847) d:LIG1 (Pi-A)</td>
<td>A:PHE131 (3.847) d:LIG1 (Pi-A)</td>
<td></td>
</tr>
</tbody>
</table>

VEGF, vascular epithelial growth factor; HB, conventional hydrogen bond; CHB, carbon hydrogen bond; A, alkyl; Pi-A, pi-alkyl.

Conclusion
In this study, we adapted in silico approaches of drug discovery to identify potential alkaloids that can prove effective in cancer treatment through VEGF receptor blocking hence obstructing angiogenesis. Through VS and molecular docking analysis, we were able to find three potential alkaloids that showed considerable binding affinity to VEGF active sites. Although in vivo interactions with

![Graphical representation of molecular docking of VEGF-A, VEGF-B, VEGF-C, and VEGF-D with Moronic acid (green color indicate Moronic acid and the dashed-line indicate bonds). VEGF, vascular endothelial growth factor.](https://doi.org/10.5808/gi.20068)
Table 6. ADME prediction of final selected 10 alkaloids using pre-ADMET tool

<table>
<thead>
<tr>
<th>No.</th>
<th>Compound</th>
<th>Human intestinal absorption (HIA, %)</th>
<th>Caco-2 cell permeability (nm/s)</th>
<th>MDCK cell permeability (nm/s)</th>
<th>Skin permeability (logKp, cm/h)</th>
<th>Blood brain barrier penetration (C.brain/C.blood)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Moronic acid</td>
<td>97.629</td>
<td>22.270</td>
<td>0.043</td>
<td>−1.96099</td>
<td>4.35022 (high)</td>
</tr>
<tr>
<td>2</td>
<td>Cadambagenic acid</td>
<td>94.671</td>
<td>21.010</td>
<td>0.044</td>
<td>−2.90457</td>
<td>2.72322 (high)</td>
</tr>
<tr>
<td>3</td>
<td>Maslinic acid</td>
<td>84.065</td>
<td>21.302</td>
<td>0.805</td>
<td>−5.13957</td>
<td>0.258554 (middle)</td>
</tr>
<tr>
<td>4</td>
<td>Nortripterifordin</td>
<td>95.204</td>
<td>21.983</td>
<td>112.722</td>
<td>−3.06612</td>
<td>2.51705 (high)</td>
</tr>
<tr>
<td>5</td>
<td>Michellamine</td>
<td>90.663</td>
<td>20.059</td>
<td>0.043</td>
<td>−3.17237</td>
<td>2.32226 (high)</td>
</tr>
<tr>
<td>6</td>
<td>Cadambine</td>
<td>67.555</td>
<td>3.851</td>
<td>0.054</td>
<td>−5.17586</td>
<td>0.0374348 (low)</td>
</tr>
<tr>
<td>7</td>
<td>Repandusinic acid</td>
<td>00.000</td>
<td>15.834</td>
<td>0.043</td>
<td>−2.63834</td>
<td>0.0277558 (low)</td>
</tr>
<tr>
<td>8</td>
<td>3a-Dihyrodambamine</td>
<td>52.205</td>
<td>6.134</td>
<td>0.073</td>
<td>−5.11884</td>
<td>0.035656 (low)</td>
</tr>
<tr>
<td>9</td>
<td>Hinokiflavone</td>
<td>86.954</td>
<td>7.156</td>
<td>0.084</td>
<td>−3.36300</td>
<td>0.280203 (middle)</td>
</tr>
<tr>
<td>10</td>
<td>Robustaflavone</td>
<td>81.196</td>
<td>12.043</td>
<td>0.043</td>
<td>−3.45363</td>
<td>0.122688 (middle)</td>
</tr>
</tbody>
</table>

ADME properties showed that these compounds are good lead molecules.

Table 7. Toxicity of final selected 10 alkaloids using OSIRIS Property Explorer

<table>
<thead>
<tr>
<th>No.</th>
<th>Alkaloid</th>
<th>Toxicity effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M, mutagenic</td>
</tr>
<tr>
<td>1</td>
<td>Moronic acid</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Cadambagenic acid</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Maslinic acid</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Nortripterifordin</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>Michellamine</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>Cadambine</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Repandusinic acid</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>3a-Dihyrodambamine</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>Hinokiflavone</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>Robustaflavone</td>
<td>No</td>
</tr>
</tbody>
</table>

**ORCID**

Shah Md. Shahik: https://orcid.org/0000-0002-1433-5085
Asma Salauddin: https://orcid.org/0000-0002-5658-8638
Md. Shakhawat Hossain: https://orcid.org/0000-0002-0187-8881
Sajjad Hossain Noyon: https://orcid.org/0000-0001-8613-6571
Abu Tayab Moin: https://orcid.org/0000-0002-2954-3215
Shagufta Mizan: https://orcid.org/0000-0002-4958-2537

Md. Thosif Raza: https://orcid.org/0000-0002-2048-6293

**Authors’ Contribution**

Conceptualization: SMS. Data curation: SMS. Formal analysis: SMS, AS, MSH. Methodology: SMS, AS, MSH, SHN, ATM. Writing - original draft: SMS, MSH, SHN, SM, MTR. Writing - review & editing: SMS, SM, MTR, AS, MSH, SHN, ATM.
Conflicts of Interest

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

Acknowledgments

We cordially thank Md. Saiful Islam, PhD Candidate/Researcher at Albert-Ludwigs-Universität Freiburg and Graduate Research Assistant at Universitätsklinikum Freiburg, for his critical and valuable suggestions on the analysis and the manuscript preparation.

Supplementary Materials

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

References

27. Carden CP, Larkin JM, Rosenthal MA. What is the risk of intracranial bleeding during anti-VEGF therapy? Neuro Oncol 2008;


41. Gao YD, Huang JF. An extension strategy of Discovery Studio 2.0 for non-bonded interaction energy automatic calculation at the residue level. Dongwuxue Yanjiu 2011;32:262-266.


In silico discovery and evaluation of phytochemicals binding mechanism against human catechol-O-methyltransferase as a putative bioenhancer of L-DOPA therapy in Parkinson disease

Surya Narayan Rath¹,², Lingaraja Jena³, Rajabrata Bhuyan⁴, Nimai Charan Mahanandia⁵, Manorama Patri²*

¹Department of Bioinformatics, Odisha University of Agriculture and Technology, Bhubaneswar 751003, India
²Neurobiology Laboratory, Department of Zoology, School of Life Sciences, Ravenshaw University, Cuttack 753003, India
³Centers for Advanced Research & Excellence, Department of Laboratory Oncology, All India Institute of Medical Science, New Delhi 110023, India
⁴Department of Bio-Science and Biotechnology, Banasthali Vidyapeeth (Deemed) University, Banasthali 304022, India
⁵Animal Biotechnology Centre, National Dairy Research Institute, Karnal 132001, India

Levodopa (L-DOPA) therapy is normally practised to treat motor pattern associated with Parkinson disease (PD). Additionally, several inhibitory drugs such as Entacapone and Opicapone are also cosupplemented to protect peripheral inactivation of exogenous L-DOPA (~80%) that occurs due to metabolic activity of the enzyme catechol-O-methyltransferase (COMT). Although, both Entacapone and Opicapone have U.S. Food and Drug Administration approval but regular use of these drugs is associated with high risk of side effects. Thus, authors have focused on in silico discovery of phytochemicals and evaluation of their effectiveness against human soluble COMT using virtual screening, molecular docking, drug-like property prediction, generation of pharmacophoric property, and molecular dynamics simulation. Overall, study proposed, nine phytochemicals (withaphysalin D, withaphysalin N, withaferin A, withacnistin, withaphysalin C, withaphysalin O, withanolide B, withasomnine, and withaphysalin F) of plant *Withania somnifera* have strong binding efficiency against human COMT in comparison to both of the drugs i.e., Opicapone and Entacapone, thus may be used as putative bioenhancer in L-DOPA therapy. The present study needs further experimental validation to be used as an adjuvant in PD treatment.

**Keywords:** inhibitors, L-DOPA, Parkinson disease, phytochemicals, *Withania somnifera*

---

**Introduction**

Parkinson disease (PD) pathology is mainly associated with progressive loss or impairment of dopaminergic neurons, occurs as a consequence of chronic inflammation, oxidative stress, deposition of protein aggregates within neurons, depletion of neurotransmitters, abnormal ubiquitination, mitochondrial dysfunction, excitotoxicity of neurons,
and disarrangement or damage of the blood-brain barrier (BBB) [1,2]. Although, influences of genetic and environmental factors in PD [3] is well studied but, complete knowledge on disease pathophysiology is still blurred. Motor related symptoms such as tremor, rigidity, and difficulty in coordination of physical movements [4] are common in PD and developed due to depletion of dopamine within an area of midbrain known as substantia nigra pars compacta. Therefore, motor disturbances in PD are treated through administration of exogenous levodopa or L-DOPA (3,4-dihydroxy L-phenylalanine) which provides only symptomatic relief [4]. In addition, co-supplementation of monoamine oxidase B, aldehyde dehydrogenase, and catechol-O-methyltransferase (COMT) inhibitors is also practised to prevent unwanted inactivation of L-DOPA within the brain [5-7].

Human COMT (EC 2.1.1.6, hCOMT) is a magnesium-dependent intracellular enzyme expressed in glial cells and neurons, and associated with diverse spectrum of neurological disorders as well as cancer [8]. hCOMT enzyme metabolizes catecholamines (norepinephrine, epinephrine, and dopamine) by introducing a methyl group from S-adenosyl methionine (SAM) to their catecholamine group [8,9]. COMT exists in two major forms such as membrane-bound (MB) COMT and soluble (S) COMT. The cellular distribution and orientation of MB-COMT on the cellular membrane is controversial [8]. However, S-COMT plays more significant role in peripheral L-DOPA deactivation than central nervous system (CNS) [10]. L-DOPA is not only a precursor of catecholamines but also an important substrate of COMT. Therefore in L-DOPA therapy, COMT inhibitors such as entacapone (Drug Bank ID: DB00494), tolcapone (Drug Bank ID: DB00323), and opicapone (Drug Bank ID: DB11632) have been used as an adjuvant to prolong the availability of L-DOPA [4] within the brain. Entacapone is a U.S. Food and Drug Administration (FDA)-approved drug that mainly acts peripherally whereas tolcapone acts both peripherally and centrally [11,12].

In PD treatment, lifelong medication is normally recommended by physician to improve the quality of patient’s life [12]. However, continuous uses of synthetic medicines have been reported with adverse effects on hepatic and cardiac health [13,14]. Due to association of serious hepatotoxicity, the drug tolcapone (Drug Bank ID: DB00323) is already withdrawn after investigation. Although, both entacapone and opicapone are FDA-approved drugs, but investigations are still going on to get COMT inhibitors with less side effects. Opicapone is a highly selective, reversible peripheral COMT inhibitor [15,16] but, associated with severe side effects such as dyskinesia, dizziness, dry mouth, and constipation [17]. In this context, several phytochemicals from different neuroprotective plants with antioxidant, anti-inflammatory, antiangiogenic, immune suppressive, anti-apoptosis, protein kinase inhibitor, anti-cholinesterase, anti–cyclooxygenase-1 (COX-1) properties have been identified and reported [2,18]. Therefore, it is essential to identify potent drug-like phytochemicals to be used as alternative medicines for the treatment of PD [2,3,18]. The present study has focused on in silico discovery and assessment of suitable herbal compounds as putative COMT inhibitors which may be experimented for further validation. This study would throw lights on discovery of natural medicine to treat PD patients with no or less risk of side effects.

Methods

Extraction and preparation of drug target structure

The X-ray crystallographic structure of hCOMT (PDB ID: 3BWM) attached with its substrate SAM and a substrate analog, 3,5-dinitrocatechol (DNC) was extracted from PDB (Protein Data Bank) (http://www.pdb.org). Initially, all crystallographic water molecules and DNC were removed from the original structure in order to dock herbal compounds into its substrate binding sites. Further, energy minimization of the target structure was performed after adding hydrogen atoms to obtain a properly optimized position of side chain atoms and hydrogen atoms using Discovery Studio 3.5 suite.

Molecular dynamic simulation of COMT

Molecular dynamic (MD) simulation was performed to study the structural stability of human S-COMT enzyme attached with and without substrates such as SAM and DNC using GROMOS96 54a7 force field of GROMACS 5.0.4 package [19]. The protonation state of the enzyme was achieved at default pH (7.0). Simple point charge water model was embedded in cubic boxes with minimum edge distance of 10 Å from the protein surface to solvate the systems. Further, electrical neutral state was attained by adding chloride ions and replacement of water [20]. Subsequently, steepest descent energy minimization was carried out until reaching to a force tolerance of 1,000 kJ/mol. Afterwards, systems were equilibrated at 300 K for 100 ps (NVT) by restraining all heavy atoms of protein backbone chain, followed by 100 ps of pressure equilibration (NPT). During NPT equilibration, all of the restraints were withdrawn. Velocity rescale thermostat [21] was used with a time constant (τT) of 0.1 ps for temperature coupling. At the same time, isotropic Parrinello-Rahman barostat (1981) was set to 1.0 bar in all directions with a time constant (τP) of 2.0 ps at the time of pressure coupling. Particle mesh Ewald method [22] was employed to take care of long-range Coulomb interactions. Similarly, the linear constraint solver (LINCS) algorithm [23] was used to...
restrict all bond lengths for a time step of 2 fs. During MD simulation, Van der Waals forces and Coulomb interactions cut-off distances were maintained at 1.0 nm of each. Each MD simulations were performed independently for a time period of 50 ns for all of the systems (COMT with and without SAM and DNC) [24].

**Inspection of ligand binding site**
The optimized structure of hCOMT was subsequently inspected to identify apposite active and functional site, where substrate normally binds to initiate its proper biochemical function. The amino acids strongly interacting with the substrate analog DNC and the ion Mg$^{2+}$ were considered as active site for ligand interaction.

**Retrieval and preparation of ligand structures**
Based on literature evidence, we found total 80 numbers of phytochemicals with anti-PD properties from different medicinal plants [2,18,24]. Three-dimensional structures of these compounds were extracted from PubChem (https://pubchem.ncbi.nlm.nih.gov/) database in SDF format and were converted to PDB format using Open Babel [25] to carry out further in silico studies. Structural geometry optimization and protonation state of these ligands were achieved using Discovery Studio 3.5 suite.

**Drug-like property prediction**
Molinspiration (http://www.molinspiration.com/) web server was used to predict the drug-like property of selected phytochemicals. It accepts ligand structure in SMILES (Simplified molecular-input line-entry system) format and predicts its bioactivity and pharmacokinetics properties following Lipinski’s rule of five [26].

**Screening of ligands**
Selected natural compounds were screened computationally against complex structure of hCOMT and SAM in order to identify efficient ligand using PyRx0.8 tool (https://pyrx.sourceforge.io/). PyRx 0.8 is an open source tool [27], used to screen libraries of compounds against potential drug target [24,28]. During virtual screening (VS) a grid of 30, 30, 30 Å in x, y, z direction was centred on drug-binding pocket of hCOMT crystal structure using AutoDock Vina [29] and PyRx 0.8 [27].

**Molecular docking**
Molecular docking was performed to validate the efficiency of selected natural compounds obtained from VS and drug-like property prediction. During docking, two FDA-approved anti-Parkinson COMT inhibitor drugs such as opicapone (DB11632), and entacapone (DB00494) were also included to compare their binding affinity with selected natural ligands. Molecular docking was performed using AutoDock 4.2 (http://autodock.scripps.edu/) and Auto-Dock Tools 4 tool [30]. Each ligand was docked independently with the enzyme COMT. During docking and further studies, Mg$^{2+}$ ion was kept intact in its position. The receptor and ligands were prepared using ADT tool [30]. Kollman charges and polar hydrogen atoms were added to the enzyme structure. Gasteiger partial charge was applied and nonpolar hydrogen atoms were merged within ligand structures. Both receptor and ligands were converted to pdbqt format before docking. A virtual grid box was set around the drug-binding cavity of the target structure with size of 30, 30, 30 Å in x, y, z direction along with spacing of 0.375 Å. Semi-flexible docking was performed by keeping the protein as rigid and allowing ligands to move within the binding cavity. Lamarckian genetic algorithm was employed to perform molecular docking. During the docking process, a maximum of 20 conformers was considered for each docking with 25,000,000 energy evaluation steps. Subsequently, all binding poses of each docking were studied and most energetically as well as geometrically favorable conformation for each independent run was selected for further study. Finally, 2D and 3D view of atomic interaction between best-docked complexes were achieved using Discovery Studio 3.5 and PyMol molecular graphics (http://www.pymol.org) tool, respectively.

**MD simulation of COMT in the presence of SAM and natural ligands**
To confirm the stability and efficacy of natural ligands fitted into the active pocket of COMT and in the presence of SAM, MD simulation of protein-ligand complex [24] was performed for 10 suitable phytochemicals. PRODRG [31] web server was used to prepare each ligand topology. Rest of the protocol was same as described above. Ten independent MD run were performed for 50 ns time period. Trajectories of all 10 simulations were saved in 10 fs interval. Microsoft Excel was used to plot graphs from the produced results.

**Prediction of pharmacophoric features**
Knowledge on different pharmacophoric properties of a lead molecule has a vital role in computer aided drug design (CADD). Presence of few chemical features such as aromatic ring (AR), hydrogen bond donor (HBD), hydrogen bond acceptor (HBA), hydrophobic property (HY) were predicted for all 10 suitable phytochemicals using ZINC Pharmer (http://zincpharmer.csb.pitt.edu/) web server.

https://doi.org/10.5808/gi.20061
**Results**

**COMT identified as a significant drug target in PD**
Possible conversion of exogenous L-DOPA to 3-O-methyldopa in both peripheral and cerebral system occurs due to the metabolic activity of COMT enzyme, is a major concern in PD treatment. In this connection, discovery and synthesis of chemical inhibitors has been aided by several solved structures of both rat and hCOMT enzyme available at PDB. Different crystal structures of human and rat COMT enzyme were observed with diverse substrate specificity and conformation. In addition, significant decrease in protein level as well enzymatic activity of COMT has been reported in association of a valine-methionine polymorphism at position 108 of hCOMT [32]. Here, we have retrieved the soluble form of hCOMT structure (PDB ID: 3BWM, chain A, length: 214, 1.98Å resolution) connected with substrates SAM and DNC. After inspection, it was identified, amino acid residues such as ASP141, LYS144, ASP169, ASN170, GLU199, and Mg$^{2+}$ were strongly interacting with the substrate analog DNC, thus considered as active drug-binding site for further study (Fig. 1).

**Validation of COMT stability by MD simulation in the presence and absence of substrates**
Macromolecules are not static in nature, so their movement causes structural fluctuation with varying energies which may affect their relevant functional phenomena. MD simulation is the one and only computational method to study the functional behavior of biological molecules such as protein or enzyme in different thermodynamical condition with respect to time scales [24,33]. Here, we performed MD simulation to discover the time dependant structural fluctuation and functional stability of the enzyme human S-COMT (PDB ID: 3BWM) in the presence and absence of substrate SAM and the substrate analog DNC. Both of the MD run was performed independently for 50 ns using GROMOS96 54a7 force field of GROMACS 5.0.4 package [18,24,33]. Overall structural consistency and stability of the backbone folding pattern was observed from root mean square deviation (RMSD) plot (Fig. 2A) in both of the systems after around ~10 ns. However, the system with substrates (SAM and DNC) was achieved the stability more quickly with an average deviation of 0.13 nm from the starting structure. Similarly, the overall root mean square fluctuation (RMSF) of COMT with SAM and DNC showed less flexibility as compared to the system in the absence of substrates (Fig. 2B). The overall packing of the systems was justified from the radius of gyration (RG) plot (Fig. 2C). As per RG plot, the packing of atoms in proteins in both of the systems (with and without substrates) were almost same (difference with only ~0.03 nm) throughout the simulation period of 50 ns (Fig. 2C). The overall MD simulation of COMT in the presence of SAM and DNC proved to be more stable than the enzyme COMT alone (Fig. 2A-2C).

**Selection of suitable phytochemicals**
Plant's crude extracts and plant-oriented natural compounds from several medicinal plants have been studied to explore their neuroprotective effect using different *in silico* and *in vivo* models [2,3,18,24,34-36]. Therefore, *in silico* identification of phytochemicals may be useful to discover suitable natural inhibitors against the PD drug target COMT. In the present study, total 80 numbers of phytochemicals (Supplemental Table 1) with medicinal properties were selected from the literature [2,18].

**Pharmacokinetic properties of proposed drug-like phytochemicals**
Determination of pharmacokinetic profile such as absorption, distribution, metabolism, excretion, and toxicology is crucial to verify the suitability of any small compound to be used as a lead molecule [37]. According to Lipinski’s rule of five a lead molecule should have ≤ 10 HBA, ≤ 5 HBD, ≤ 500 molecular weight, ≤ 5 octanol/water partition coefficient (miLogP), ≤ 90 Å square topological polar surface area (TPSA). It is considered, reduction in bioactivity of a lead molecule may occur due to violation of any of these two properties [26]. Here, the prediction proposed only 63 phytochemicals with good pharmacokinetic profile (Supplemental Table 2). Again, as of Lipinski’s rule of five [26], the permeability through cell membrane for a lead compound is evaluated through its computed TPSA value. So, compound with TPSA value greater than 140 Å squared tend to be poor at permeating cell membranes [38]. But, in case of CNS-related drugs, the TPSA value less than 90 Å squared is mostly acceptable which indicates their ability to penetrate through the cell membrane as well as BBB [39]. Therefore, after ADMET analysis, it is recommended, out of 63 com-

---

**Fig. 1.** Crystal structure of human catechol-O-methyltransferase (COMT) (PDB ID: 3BWM, chain A) enzyme with substrate S-adenosyl methionine and 3,5-dinitrocatechol (DNC) (left). Amino acids found interacting with DNC and Mg$^{2+}$ ion within the active pocket of COMT are deciphered in the right side.
pounds (Supplemental Table 2), only 39 (Supplemental Table 3) bioactive compounds have the ability to cross the BBB, thus may be useful to be used as CNS drugs. However, total 17 compounds were strongly violated one or two Lipinski’s rule (Supplemental Table 2), therefore discarded from further study.

**VS recommended efficient natural ligands**

VS of ligands have been utilized successfully as an effective in silico technique for filtering out potential ligands against appropriate drug target [24,27,28,37]. It is an economical and time-saving approach and possibly helps experimental procedure to increase the success rate in drug discovery. Here, site-directed VS was performed for 63 previously studied phytochemicals with good phar-

---

**Fig. 2.** Molecular dynamics simulation (MD) plots of catechol-O-methyltransferase (COMT) enzyme attached with and without S-adenosylmethionine (SAM) and a substrate analog, 3,5-dinitrocatechol (DNC): root mean square deviation (RMSD) plot of backbone (A), root mean square fluctuation (RMSF) plot for residue wise fluctuation (B), and radius of gyration (RG) plot for overall compactness of the system (C). MD simulation plots of COMT enzyme in complex with 10 different phytochemicals of plant *Withania somnifera*: RMSD plot of backbone (D), RMSF plot for residue wise fluctuation (E), and RG plot for overall compactness of the system in presence of phytochemicals (F).
macokinetic properties (Supplemental Table 2). As, COMT enzyme has the ability to degrade L-DOPA both peripherally and centrally [8-10,32] therefore, all of these 63 natural ligands (Supplemental Table 2) were screened to discover compounds with potential binding affinity against the drug target. According to VS result, four phytochemicals such as withaphysalin M, withaphysalin N, withaphysalin F, and withaphysalin O of plant Withania somnifera were showed better binding energy than rest others (Supplemental Table 4). However, on the basis of suitable binding affinity and pharmacokinetic profile 15 natural compounds (Table 1) of W. somnifera plant were subjected for further validation using molecular docking study. 

Molecular docking confirmed the binding efficiency of W. somnifera phytochemicals against COMT

Total 15 suitable phytochemicals (Table 1) of plant W. somnifera were docked into the drug-binding pocket of human S-COMT. The binding energy of protein-ligand complex resulted from molecular docking was compared with VS score of each compound and reported (Table 2). It was confirmed all of these 15 phytochemicals have potential binding efficiency against hCOMT (Table 2). In addition, 10 natural compounds such as withaphysalin M (−7.42 kcal/mol), withaphysalin N (−7.24 kcal/mol), withaphysalin F (−6.48 kcal/mol), withaphysalin O (−6.78 kcal/mol), withaphysalin C (−6.85 kcal/mol), withaphysalin D (−7.84 kcal/mol), withanolide B (−7.63 kcal/mol), withaferin A (−7.53 kcal/mol), withacnistin (−7.13 kcal/mol), and withasomnine (−6.09 kcal/mol) were showed consistency in binding energy scores (Table 2) with VS scores which advocated for their reliability in binding against COMT. Further, two FDA-approved drugs such as opicapone (DB11632), and entacapone (DB00494) were docked within the drug-binding site of COMT. Interestingly, eight phytochemicals such as withaphysalin D (−7.84 kcal/mol; KI: 1.8 μM), withanolide B (−7.63 kcal/mol; KI: 2.54 μM), withaferin A (−7.53 kcal/mol; KI: 3.03 μM), withaphysalin M (−7.42 kcal/mol; KI: 3.67 μM) were showed better binding energy than rest others (Supplemental Table 2). However, on the basis of suitable binding affinity and pharmacokinetic profile 15 natural compounds (Table 1) of W. somnifera plant were subjected for further validation using molecular docking study.

Table 1. Binding energy scores of 15 drug-like phytochemicals of plant Withania somnifera resulted from virtual screening against human COMT enzyme

<table>
<thead>
<tr>
<th>No.</th>
<th>Ligand</th>
<th>Binding energy score</th>
<th>Drug likeness (Lipinski’s rule of five)</th>
<th>BBB permeant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Withaphysalin M</td>
<td>−10.3</td>
<td>Suitable</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Withaphysalin N</td>
<td>−10.3</td>
<td>Suitable</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Withaphysalin F</td>
<td>−9.9</td>
<td>Suitable</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Withaphysalin O</td>
<td>−9.5</td>
<td>Suitable</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>Withaphysalin C</td>
<td>−5.8</td>
<td>Suitable</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>Withaphysalin D</td>
<td>−5.3</td>
<td>Suitable</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>Withanolide B</td>
<td>−6.0</td>
<td>Suitable</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>Withaferin A</td>
<td>−6.1</td>
<td>Suitable</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>Withacnistin</td>
<td>−5.9</td>
<td>Suitable</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>Withasomnine</td>
<td>−5.3</td>
<td>Suitable</td>
<td>Yes</td>
</tr>
<tr>
<td>11</td>
<td>Anaferine</td>
<td>−4.6</td>
<td>Suitable</td>
<td>Yes</td>
</tr>
<tr>
<td>12</td>
<td>Calystegine B2</td>
<td>−5.2</td>
<td>Suitable</td>
<td>No</td>
</tr>
<tr>
<td>13</td>
<td>Cuscohygrine</td>
<td>−4.2</td>
<td>Suitable</td>
<td>Yes</td>
</tr>
<tr>
<td>14</td>
<td>Pelletierine</td>
<td>−3.9</td>
<td>Suitable</td>
<td>Yes</td>
</tr>
<tr>
<td>15</td>
<td>Tropine</td>
<td>−4.0</td>
<td>Suitable</td>
<td>Yes</td>
</tr>
</tbody>
</table>

COMT, catechol-O-methyltransferase; BBB, blood-brain barrier.
μM), withaphysalin N (–7.24 kcal/mol; KI: 16.96 μM), withaphysalin C (–6.51 kcal/mol; KI: 11.41 μM), withaphysalin O (–6.34 kcal/mol; KI: 22.55 μM). However, rest five phytochemicals namely cuscohygrine (–6.51 kcal/mol; KI: 16.96 μM), withaphysalin F (–6.48 kcal/mol; KI: 17.83 μM), anaferine (–6.33 kcal/mol; KI: 23.06 μM), pelletierine (–6.31 kcal/mol; KI: 23.5 μM), and withasomnine (–6.09 kcal/mol; KI: 34.35 μM) were appeared as close binding competitors of both of the drugs (Table 3). Again, the presence of ample numbers of amino acid residues in hydrogen bond formation, Van der Waals interaction, and Pi-Alkyl interaction within active site of hCOMT enzyme also established significant interaction of the phytochemicals with COMT (Fig. 3). Again, participation of strong polar interactions (distance ≤ 3Å) between hCOMT and phytochemicals (Table 4, Fig. 4) of plant W. somnifera were perceived in favor of the above observation. To its support, few amino acids such as Tyr 68 (withaphysalin M, withaferin A, withacnistin, withasomnine, withaphysalin F), Lys144 (withaphysalin N, withaphysalin M, withaphysalin C, withaphysalin O, withaphysalin F), Asp145 (withaphysalin D, withanolide B, withaphysalin C, withaphysalin O), and Arg146 (withaphysalin D, withanolide B) were identified as commonly participated in polar interaction within the binding cavity (distance ≤ 3.5Å) of hCOMT (Table 4, Fig. 4) enzyme. However, the overall study confirmed about their strong atomic interaction with hCOMT consequently, subjected for MD simulation.

**MD simulation established structural stability of COMT-phytochemical complex in presence of SAM**

On the basis of recommendation of all previous observations, 10 phytochemicals (withaphysalin M, withaphysalin N, withaphysalin F, withaphysalin O, withaphysalin C, withaphysalin D, withanolide B, withaferin A, withacnistin, and withasomnine) of plant W. somnifera were appeared to have possible impact to block the active site of human S-COMT, insisted authors to perform MD simulation of the enzyme (PDB ID: 3BWM) in the presence of these phytochemicals along with its natural substrate SAM to observe its structural and functional behavior in complex form. Ten independent MD run were performed for protein-ligand complex up to 50 ns time scale. From the RMSD plot of backbone atomic structure, it was identified, hCOMT attached with different phytochemicals were quite consistent after around ~30 ns, suggesting the better stability of the enzyme except in one case, i.e., the COMT and withaphysalin M complex (Fig. 2D). In this case, significant fluctuation in RMSD plot was observed after around ~34 ns which continued till the end of 50 ns MD simulation, indicated about the instability of COMT and withaphysalin M complex (Fig. 2D). Similar type of observation was perceived from the RMSF plot (Fig. 2E). Minor fluctuation in amino acid residual positions was noticed for all COMT-phytochemical complexes except withaphysalin M which pointed out the binding stability of all nine natural compounds (Fig. 2E) of W. somnifera plant. Higher fluctuations in amino acids of COMT were observed near the regions having no specific secondary structure. Similarly, the overall packing of COMT enzyme was found quite stable and compact throughout the simulation period of 50 ns in all nine protein-ligand complexes except withaphysalin M as plotted in RG plot (Fig. 2F). Furthermore, to strengthen this hypothesis, RMSD plot of different phytochemicals and SAM from all of the MD systems were plotted (Fig. 5A and 5B). Quite satisfactory observation was noticed in case of all natural compounds (Fig. 5A) and SAM (Fig. 5B) in their respective enzyme-ligand-SAM complex within time scale of 50 ns MD simulation. Overall, MD simulation results discovered, out of 10 only nine phytochemicals of plant W. somnifera have potential binding stability against the soluble hCOMT enzyme.

**Table 3. Docking scores of 15 phytochemicals and two synthetic inhibitors (opicapone and entacapone) against human COMT resulted from molecular docking**

<table>
<thead>
<tr>
<th>No.</th>
<th>Ligand (phytochemical/drug)</th>
<th>Binding energy score (kcal/mol)</th>
<th>Inhibition constant (μM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Withaphysalin D</td>
<td>–7.84</td>
<td>1.8</td>
</tr>
<tr>
<td>2</td>
<td>Withanolide B</td>
<td>–7.63</td>
<td>2.54</td>
</tr>
<tr>
<td>3</td>
<td>Withaferin A</td>
<td>–7.53</td>
<td>3.03</td>
</tr>
<tr>
<td>4</td>
<td>Withaphysalin M</td>
<td>–7.42</td>
<td>3.67</td>
</tr>
<tr>
<td>5</td>
<td>Withaphysalin N</td>
<td>–7.24</td>
<td>4.93</td>
</tr>
<tr>
<td>6</td>
<td>Withacnistin</td>
<td>–7.13</td>
<td>5.92</td>
</tr>
<tr>
<td>7</td>
<td>Withaphysalin C</td>
<td>–6.85</td>
<td>9.56</td>
</tr>
<tr>
<td>8</td>
<td>Withaphysalin O</td>
<td>–6.78</td>
<td>10.76</td>
</tr>
<tr>
<td>9</td>
<td>Opicapone&lt;sup&gt;a&lt;/sup&gt;</td>
<td>–6.74</td>
<td>11.41</td>
</tr>
<tr>
<td>10</td>
<td>Cuscohygrine</td>
<td>–6.51</td>
<td>16.96</td>
</tr>
<tr>
<td>11</td>
<td>Withaphysalin F</td>
<td>–6.48</td>
<td>17.83</td>
</tr>
<tr>
<td>12</td>
<td>Entacapone&lt;sup&gt;a&lt;/sup&gt;</td>
<td>–6.34</td>
<td>22.55</td>
</tr>
<tr>
<td>13</td>
<td>Anaferine</td>
<td>–6.33</td>
<td>23.06</td>
</tr>
<tr>
<td>14</td>
<td>Pelletierine</td>
<td>–6.31</td>
<td>23.5</td>
</tr>
<tr>
<td>15</td>
<td>Withasomnine</td>
<td>–6.09</td>
<td>34.35</td>
</tr>
<tr>
<td>16</td>
<td>Calystegine B2</td>
<td>–4.98</td>
<td>224.43</td>
</tr>
<tr>
<td>17</td>
<td>Tropine</td>
<td>–4.94</td>
<td>240.97</td>
</tr>
</tbody>
</table>

COMT, catechol-O-methyltransferase.
<sup>a</sup>Drug compounds.

https://doi.org/10.5808/gi.20061
Fig. 3. The 2D view of close amino acid residues participated in h-bond, Van der Waals interaction, and Pi-Alkyl interaction with phytochemicals within the active pocket of human catechol-O-methyltransferase (COMT) are represented: withaphysalin D (A), withaphysalin N (B), withaphysalin M (C), withaferinA (D), withacnistin (E), withaphysalin C (F), withaphysalin O (G), withaphysalin F (H), withasomnine (I), and withanolide B (J).
**Table 4.** Strong atomic interaction predicted between 10 phytochemicals of *Withania somnifera* plant and human COMT enzyme (distance ≤3.5 Å)

<table>
<thead>
<tr>
<th>No.</th>
<th>Phytochemical</th>
<th>Predicted amino acid residues within active site of COMT (distance ≤ 3 Å)</th>
<th>Predicted h-bond residues</th>
<th>Bond</th>
<th>Distance between atoms (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Withaphysalin D</td>
<td>Met 40, Asp 141, His 142, Trp 143, Lys 144, Asp 145, Arg 146, Tyr 147, Asn 170, Pro 174</td>
<td>Asp 145</td>
<td>HN---O</td>
<td>2.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Arg 146</td>
<td>HN---O</td>
<td>3.33</td>
</tr>
<tr>
<td>2</td>
<td>Withanolide B</td>
<td>His 142, Trp 143, Lys 144, Asp 145, Arg 146, Tyr 147, Pro 174</td>
<td>Asp 145</td>
<td>O---HO</td>
<td>2.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Arg 146</td>
<td>HN---O</td>
<td>3.00</td>
</tr>
<tr>
<td>3</td>
<td>Withaphysalin N</td>
<td>Met 40, His 142, Trp 143, Lys 144, Asp 145, Arg 146, Tyr 147, Asn 170, Pro 174</td>
<td>Lys 144</td>
<td>HN---O</td>
<td>2.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>HN---O</td>
<td>2.87</td>
</tr>
<tr>
<td>4</td>
<td>Withaphysalin M</td>
<td>Met 40, Tyr 68, Asp 141, His 142, Trp 143, Lys 144, Asp 145, Tyr 147, Asn 170, Pro 174, Ala 176</td>
<td>Tyr 68</td>
<td>O---HO</td>
<td>2.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Asp 141</td>
<td>O---HO</td>
<td>2.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lys 144</td>
<td>N---O</td>
<td>2.93</td>
</tr>
<tr>
<td>5</td>
<td>Withaferin A</td>
<td>Met 40, Tyr 68, Asp 141, His 142, Trp 143, Tyr 147, Asp 169, Asn 170</td>
<td>Tyr 68</td>
<td>O---O</td>
<td>2.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>His 142</td>
<td>O---HO</td>
<td>2.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tyr 147</td>
<td>O---HO</td>
<td>2.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Asn 170</td>
<td>N---OH</td>
<td>3.38</td>
</tr>
<tr>
<td>6</td>
<td>Withacnistin</td>
<td>Trp 38, Met 40, Tyr 68, His 142, Trp 143, Lys 144, Tyr 147, Asp 169, Asn 170, Pro 174, Leu 198, Glu 199, Arg 201, Asp 205</td>
<td>Tyr 68</td>
<td>OH---O</td>
<td>3.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pro 174</td>
<td>N---O</td>
<td>3.30</td>
</tr>
<tr>
<td>7</td>
<td>Withaphysalin C</td>
<td>Trp 38, Met 40, Tyr 68, His 142, Trp 143, Lys 144, Asp 145, Arg 146, Pro 174, Leu 198, Arg 201</td>
<td>Lys 144</td>
<td>HN---O</td>
<td>2.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Asp 145</td>
<td>HN---O</td>
<td>2.11</td>
</tr>
<tr>
<td>8</td>
<td>Withaphysalin O</td>
<td>Trp 38, Met 40, His 142, Trp 143, Lys 144, Asp 145, Arg 146, Tyr 147, Asn 170, Pro 174, Leu 198, Arg 201</td>
<td>Lys 144</td>
<td>NH---O</td>
<td>1.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Asp 145</td>
<td>O---HO</td>
<td>2.19</td>
</tr>
<tr>
<td>9</td>
<td>Withasomnine</td>
<td>Met 40, Tyr 68, Asp 141, His 142, Trp 143, Lys 144, Tyr 147, Asp 169, Asn 170, Cys 173, Pro 174, Gly 175, Ala 176</td>
<td>Tyr 68</td>
<td>O---NH</td>
<td>2.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Withaphysalin F</td>
<td>Met 40, Tyr 68, Asp 141, His 142, Trp 143, Lys 144, Asp 145, Tyr 147, Asn 170, Pro 174, Ala 176</td>
<td>Tyr 68</td>
<td>O---OH</td>
<td>2.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Asp 141</td>
<td>O---OH</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>His 142</td>
<td>O---HO</td>
<td>3.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lys 144</td>
<td>O---N</td>
<td>3.00</td>
</tr>
</tbody>
</table>

COMT, catechol-O-methyltransferase.

**Fig. 4.** Atomic interaction including h-bond between 10 phytochemicals (withaphysalin O, withasomnine, withaphysalin F, withaphysalin M, withaferin A, withacnistin, withaphysalin C, withaphysalin D, withanolide B, and withaphysalin N) and human catechol-O-methyltransferase enzyme are depicted.
**Presence of pharmacophoric features advocated for phytochemicals efficacy**

Pharmacophore based ligand discovery has a critical role in CADD [37,40]. Therefore, identification of important chemical features such as the presence of AR, hydrophobic feature (HY), HBD, and HBA are necessary to confirm the effect of interaction between a lead molecule and the drug target. Presence of above pharmacophoric properties was discovered in functional groups of all 10 suitable phytochemicals (Table 5, Fig. 6) and thus, strongly recommended for their effectiveness and sensible binding interaction against the PD drug target COMT.

**Discussion**

Disturbance of motor activity has been perceived as a preliminary symptom [4] in PD and is usually treated through administration of L-DOPA [12]. COMT plays a significant role in the metabolism of L-DOPA, thus inactivates exogenous L-DOPA both in peripheral and CNS. Therefore, few COMT inhibitor drugs such as Opicapone and Entacapone are cosupplemented with L-DOPA to maintain the dopamine level within CNS [4,10,12]. On the contrary, the long-term uses of these medicines are associated with the risk of patient’s cardiac and hepatic health [17]. In this context, possible use of plant-oriented natural inhibitors has received growing interest of scientist globally. Since ages, many pant derived natural compounds or phytochemicals have been known to be effective against neurological disorders due to the presence of their antioxidant, anti-inflammatory, antiangiogenic, immune suppressive, anti-apoptosis, protein kinase inhibitor, anticholinesterase, anti-COX-1 properties [2,18,24]. The rationale of this approach is established through several in silico, in vitro, in vivo, and preclinical studies [2,3,18,24,34-36,41]. Additionally, the use of phytochemicals offers advantages over synthetic drugs such as no or minimal side effects.

The immense importance of different neuroprotective natural compounds encouraged authors to investigate binding stability and suitability as putative inhibitors against PD drug target COMT. Based on literature evidence, structures of 80 natural compounds [2,18,24] were retrieved from the public repository and subjected to verify their drug-like property. Pharmacokinetics of a potential inhibitor depends on its good drug-like properties [37] which were verified by following the Lipinski’s rule of five [26]. Upon verification, suitable drug-like property was confirmed in case of 63 phytochemicals (Supplemental Table 2) therefore structures of those phytochemicals were virtually screened against crystal structure of hCOMT within its known drug-binding site. VS technique has been revealed as a promising in silico procedure to identify potential lead compounds against any drug target [24,27,28,33,37]. VS result was proposed four phytochemicals such as withaphysalin M, withaphysalin N, withaphysalin F, and withaphysalin O (Supplemental Table 4) of plant W. somnifera...
with strong binding affinity against hCOMT. However, on the basis of suitable binding affinity and drug-like properties total 15 phytochemicals (Table 1) were selected for further study. Further, molecular docking was performed to confirm binding affinity and binding pattern of these 15 natural compounds. Comparative analysis of VS and docking results was revealed 10 natural compounds (withaphysalin M, withaphysalin N, withaphysalin F, withaphysalin O, withaphysalin C, withaphysalin D, withanolide B, withaferin A, withacnistin, and withasomnine) as suitable due to their consistency in binding scores (Table 2). Further to compare the binding affinity of natural compounds with synthetic COMT inhibitors two FDA-approved drugs namely opicapone (DB11632), and entacapone (DB00494) were also docked within the drug-binding site of COMT. Interestingly, better binding affinity and inhibition constant was found in case of eight phytochemicals (withaphysalin D, withaphysalin N, withaphysalin F, withaphysalin O, withaphysalin C, withaphysalin D, withanolide B, withaferin A, withacnistin, and withasomnine) than both of the drugs (Table 3) which confirmed their efficacy. To its support, interaction analysis was suggested for significant binding pattern between selected 10 natural compounds (Table 2) of plant W. somnifera due to the presence of good numbers of strong hydrogen bond (distance ≤ 3Å), Van der Waals interaction, and Pi-Alkyl interaction within active site of hCOMT enzyme (Table 4, Figs. 3 and 4).

In order to assess the stability and conformational changes in COMT upon binding of these 10 suitable phytochemicals of plant W. somnifera MD simulation was performed for 50ns. The values of RMSD, RMSF, and RG plot suggested the binding of all of these nine phytochemicals of plant W. somnifera except withaphysalin M stabilized the COMT structure in presence of SAM (Fig. 2D-2F) without any conformational shift. However, several random fluctuations were seen initially, but no conformational switching was observed during entire simulation period (Fig. 2D-2F). Notably, RMSD and RMSF values of SAM and all of these 10 phytochemicals were found quite satisfactory in their respective enzyme-ligand-SAM complex (Fig. 5) within 50 ns MD simulation. In addition, the pharmacophoric features of all of these phytochemicals found suitable to be used as lead compounds against PD drug target COMT. (Table 5, Fig. 6). The overall analysis hypothesized, all of these nine phytochemicals (withaphysalin N, withaphysalin F, withaphysalin O, withaphysalin C, withaphysalin D, withanolide B, withaferin A, withacnistin, and withasomnine) of plant W. somnifera have potential binding efficiency and may be used as putative inhibitors against PD drug target COMT.

In conclusion, the present in silico study discovered, total of nine phytochemicals (withaphysalin D, withaphysalin N, withaferin A, withacnistin, withaphysalin C, withaphysalin O, withanolide B, withasomnine, withaphysalin F) of plant W. somnifera (ashwagandha) with good pharmacokinetic profile, pharmacophoric features and stable binding potentiality against hCOMT enzyme. Thus, it
is hypothesized that these phytochemicals may be used as putative bioenhancer in L-DOPA treatment. The present study would throw lights on discovery of natural inhibitors against COMT as an alternative treatment of PD and may be further extended for experimental validation in the future.

**ORCID**

Surya Narayan Rath: https://orcid.org/0000-0002-5458-8351
Lingaraja Jena: https://orcid.org/0000-0001-6669-3610
Rajabrata Bhuyan: https://orcid.org/0000-0001-7777-8230
Nimai Charan Mahanandia: https://orcid.org/0000-0002-9001-8284
Manorama Patri: https://orcid.org/0000-0002-6051-4180

**Authors’ Contribution**

Conceptualization: SNR, MP. Data curation: SNR, NLM. Formal analysis: SNR, LJ, RB, NLM. Methodology: SNR, LJ, RB, NLM. Writing - original draft: SNR. Writing - review & editing: SNR, LJ, RB, MP.

**Conflicts of Interest**

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

**Acknowledgments**

We are thankful to all members of Neurobiology laboratory, Department of Zoology, Ravenshaw University, Cuttack, and Department of Bioinformatics, Odisha University of Agriculture and Technology, Bhubaneswar for technical support and valuable discussions.

**Supplementary Materials**

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

**References**

18. Kumar G, Patnaik R. Exploring neuroprotective potential of Withania somnifera phytochemicals by inhibition of GluN2B-containing


Introduction

The nasuta subgroup of immigrants species of Drosophila representing over nominal 12 species or subspecies exhibits varying degrees of the post and pre-zygotic isolations, making it a potent cluster to study speciation genetics. Comparative investigations among the members of nasuta radiation have provided insights into cellular complexity at genomic and transcriptomic levels. The evolutionary history of the rapidly radiating nasuta subgroup is well characterized, which has allowed for taxonomic inferences [1]. Genome-wide studies have revealed early stages of sex chromosome evolution and the origin of B chromosome in D. albomicans [2]. Transcriptome studies in D. nasuta and D. albomicans have formed the basis for deducing the role of alternative splicing towards lin-
eage-specific evolution in *Drosophila* [3]. *Drosophila nasuta nasuta* (D. n. nasuta) and *Drosophila nasuta albomicans* (D. n. albomicans) belong to the orbital sheen complex, one of the three sheen complexes characterized in the *nasuta* subgroup of *Drosophila* [4]. Despite the documented initial post-zygotic incompatibilities in the F1 hybrids, the combinatorial crossing of various strains of these karyotypically diverged species and their successive maintenance has resulted in an assemblage of stabilized hybrids termed cytoraces [5,6]. In nature, many instances of hybridization in *Drosophila* are reported [7]. However, the number of laboratory-induced interspecific hybrids between closely related *Drosophila* species is relatively high [8].

Cytoraces with their parental species constitute the *nasuta-albomicans* complex (NAC) of *Drosophila*, which is the longest ongoing evolutionary experiment in a laboratory setting in the genus *Drosophila* and the independently evolving populations of cytoraces are currently passing ~850 generations [9]. Cytorace-3 (C3), a member of NAC stemmed from the interracial hybridization event between *D. n. nasuta* males (Coorg strain), and *D. n. albomicans* (Thailand strain) females carry a stable chromosome complement of 2n = 8 resembling *D. n. nasuta* with ten *D. n. nasuta* chromosomes and six *D. n. albomicans* chromosomes (Fig. 1). The rapid divergence in C3 is extensively recorded through comparative studies for male and female reproductive fitness, wing size, genitalia size, body weight, body size, longevity [10], mating latency, and copulation duration in heterogamic crosses [11]. This empirical evidence makes C3 an attractive model to study the molecular basis of early events in racial divergence.

Advances in RNA sequencing technologies and bioinformatics approaches for data analysis have enabled the precise measurement of gene expression levels. De novo transcriptome assembling is a widely used method for conducting spatio-temporal and condition specific gene expression profiling in non-model organisms in lieu of reference genomes. This dissemination of technology has allowed researchers to investigate evolutionary questions once limited to model organisms [12]. Recently, transcriptomics’ increased applicability has permitted quantifying hybridization-induced effects at a transcription level in interspecific hybrids and pure species [13-16]. Interspecific hybrids display a wide range of genetic changes in the hybrid genomes, ranging from introgression, chromosomal rearrangements, and differential gene expression to epigenetic modifications [13-15,17-19]. Such genetic changes invariably provide an insight into fixed accumulated species-specific changes limiting transgenerational assessments in hybrids due to failed reproduction, restricting the formation of true-bred hybrid populations. Therefore, the extent of genetic divergence between species can influence the outcome of hybridization and is well argued in the context of *Drosophila* [20].

In this study, we examined the expression profiles of mature ovaries of virgin females and testes from naïve males of C3 employing Illumina sequencing technology and de novo transcriptome assembling strategies. The study aims to comprehensively quantify gene expression divergence in C3 and deduce the inheritance of gene expression levels through comparative analysis with their parental transcriptomes. To our knowledge, this is the first report focused on a systematic exploration of expression profiles in the gonads of an independently evolving established homoplody hybrid lineage of *Drosophila*. The present findings amalgamate the existing works in understanding the onset of hybridization driven speciation and provide a valuable resource for subsequent studies.

### Methods

**RNA sequencing and de novo transcriptome assembly**

Fly stocks of C3 were maintained under laboratory conditions of 22 ± 1°C, 12:12 LD cycle in standard wheat cream agar medium. Ovaries were dissected from 3–5 days old virgin females and testes from 3–5 days old naïve males. For each tissue type, 60 flies were dissected in ice-cold phosphate-buffered saline. To preclude mating induced transcriptional responses which are sex specific, the study focused on profiling only the gonadal transcriptomes in the virgin flies. RNA extraction was performed using the ZR-DuetTM RNA MiniPrepPlus kit (Zymo Research, Irvine, CA, USA), followed with cDNA libraries preparation using NEBNext Ultra RNA library prep kit for Illumina (New England Biolabs, Ipswich, MA, USA) and validated using Agilent DNA HS (Agilent Technologies, Santa Clara, CA, USA) and Qubit DNA BR assay kits (Life Technologies, Carlsbad, CA, USA). The sequencing of QC passed libraries was carried on the Illumina HiSeq 2500 (Illumina, San Diego, CA, USA) platform for 2 × 101 bp read length. Reads of parental lineages that belong to the same bio-project were downloaded from NCBI accessions SRR8398945 and SRR10875322 (D.
n. albomicans ovary and testis), SRR8398946 and SRR10875323 (D. n. nasuta ovary and testis). A multistep raw data processing approach was undertaken to produce high quality read datasets (Fig. 2). The datasets were checked for sequence quality using FASTQC v0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Subsequently, QC passed datasets were then corrected for low K-mer values using Rcorrector [21] followed by rRNA contamination removal using the SILVA rRNA database [22]. Following this, high quality read datasets from each library were de novo assembled to construct a tissue-specific transcriptome assembly for C3 and a co-assembled reference assembly of each tissue by pooling high quality reads from C3, D. n. nasuta and D. n. albomicans using Trinity v2.6.6 with default parameters [23]. The co-assembling strategy has been implemented in studies involving closely related species which have not diverged extensively. Considering divergence history of the parental lineages and C3 this strategy is deemed suitable in the present study and enhances the ability of the assembler to construct lowly expressed transcripts. Redundancy imparted in assembling stages was reduced by clustering transcripts with > 95% similarity using CD-HIT-EST v4.6 [24].

**Structural and functional annotations of assembled transcriptomes**

TransDecoder v5.5.0 [23] was run on clustered assemblies to extract open reading frames (ORFs) from transcripts and predict potential coding sequences using default parameters. Assignment of functional definitions to assembled transcripts was performed by querying against UniprotKB/Swiss-Prot database (release June 2019), EggNOG [25], Pfam v31.0 [26], TMHMM v2.0 [27], and SignalP v4.1 [28] databases and results compiled using Trinotate v3.1.1-pl526_5 [29] pipeline.

**Differential gene expression analysis of Illumina sequence data**

Transcript abundance was estimated by aligning high quality reads used for assembly construction against the final co-assembled reference assembly using Bowtie2 v2.3.4.3 [30]. Following this, raw gene-level counts were estimated using RNAseq by Expectation-Maximization (RSEM) package [31]. A Gene with a potential ORF and with a minimum one transcript per million (TPM) value was considered as expressed and was used in downstream expression analysis. Pairwise comparisons were performed for differential gene expression analysis between C3 and D. n. nasuta and
and 72°C) was carried using the cDNA with the selected primer genes, an additional PCR (40 cycles of 95°C for 10 s, 60°C for 20 s, 95°C for 10 s) was performed in a 20 µL reaction volume according to the manufacturer’s protocol. 

As a confirmatory step to check for the presence of expressed genes, an additional PCR (40 cycles of 95°C for 10 s, 60°C for 20 s, 95°C for 10 s) was performed in a 20 µL reaction volume according to the manufacturer’s protocol. 

Real-time quantitative polymerase chain reaction validation of DEGs

The reliability of the differential expression results quantified by whole transcriptome sequencing (WTS) was validated through real-time quantitative polymerase chain reaction (RT-qPCR). Twelve DEGs (six each from ovary and testis) which were commonly differentially expressed in C3 ovaries and testes relative to the parental transcriptomes were selected for validation. Primers were designed using Primer3 v0.4.0 software [36] and were synthesized at Xcelris Labs Ltd. (Ahmedabad, India). Gene list with primer sequence information along with respective Tm values is provided in Supplementary Table 1. Ovaries were dissected from 3–5 days old virgin females and testis from 3–5 days old naive males. Total RNA was extracted separately from ovary and testis tissues (30 pairs each) of the three samples using the GeneJET RNA purification kit (Thermo Fisher Scientific, Waltham, MA, USA) and quantified using Nanodrop 1000 spectrophotometer (Thermo Fisher Scientific). cDNA synthesis was carried by RevertAid first-strand synthesis kit (Thermo Fisher Scientific) in a 20 µL reaction volume according to the manufacturer’s protocol. As a confirmatory step to check for the presence of expressed genes, an additional PCR (40 cycles of 95°C for 10 s, 60°C for 20 s, and 72°C) was carried using the cDNA with the selected primer pairs. Quantification was carried in biological triplicates and technical duplicates in StepOnePlus Real-Time PCR System (Applied BioSystems, Foster City, CA, USA) using SYBR Green chemistry (Maxima SYBR Green/ROX qPCR Master Mix 2×) (Thermo Fisher Scientific). The specificity of primers in amplifying target genes was reconfirmed by melt curve analysis. Gapdh2 was used as an endogenous control following which relative gene expression quantification was calculated by the 2\(^{-ΔΔct}\) method.

Accession numbers

The accession numbers for the C3 ovary and C3 testis sequences reported in this paper are deposited in Sequence Read Archive: SRR8398943, SRR10875320 under BioProject ID PRJNAS12942 and PRJNA600771.

Results

Transcriptome sequencing and assembly construction

Tissue-specific sequencing of the C3 using Illumina Hiseq2500 paired-end yielded a total of 16,419,268 and 10,466,674 reads corresponding to ovarian and testes datasets. Sequence reads with PHRED score < 20 were discarded. Subsequent removal of rRNA contamination, adaptor sequences, and K-mer correction generated 15,946,326 and 10,100,250 high-quality filtered reads for ovary and testis, further used for transcriptome assembly construction. Tissue-specific assemblies were constructed using C3 reads adopting the de novo approach with Trinity software. Tissue-specific assemblies were used to assess the coding potential of transcripts and for assigning functional definitions. We assembled 16,642 transcripts from ovarian, and 20,823 transcripts from the testis reads of C3. Redundancy removal through clustering resulted in the retention of ~92% and ~94% transcripts from C3 ovary and testis assemblies, respectively. Alignment rate of 96.92% of ovarian reads and 95.93% of testis reads deduced from back-mapping the reads to the constructed assemblies are indicative of good quality assemblies suitable for transcriptome characterization. Assembly statistics are provided in Table 1, and the functional annotation of the assembled transcripts against various databases is summarized in Table 2. Comparative enrichment analysis of gene ontology (GO) terms for the testis and ovary transcriptome of C3 was categorized into “biological process,” “molecular functions,” and “cellular component” (Fig. 3).

A combined reference assembly of 15,273 transcripts (ovarian co-assembly) was generated by pooling the high-quality filtered reads from ovaries of parental lineages and C3. Similarly, a testis-specific co-assembly (testis co-assembly) was also generated with 19,465 transcripts. Transcripts that carried a potential ORF
Among the pairwise comparisons, genes differentially expressed in C3 relative to both its parents were listed, and their KEGG enrichment analysis was performed. Two hundred twenty-five up-regulated genes in C3 ovary constituted 48 KEGG pathways, of which, ten were significantly enriched. Five hundred forty-six down-regulated genes in C3 ovary constituted 51 KEGG pathways, of which, seven were significantly enriched. 169 up-regulated genes in C3 testis constituted 59 KEGG pathways, of which, ten were significantly enriched. Two hundred forty-four down-regulated genes in C3 testis constituted 143 KEGG pathways, of which, 56 were significantly enriched. A detailed list of these enriched KEGG pathways is provided in Supplementary Table S2-S5. Enrichment analysis of GO terms on these common DEGs was performed, and the top ten GO terms from the three categories, “biological process,” “molecular functions,” and “cellular component,” are shown in Supplementary Figs. 1-4.

**Gene expression inheritance analysis indicates the predominance of conserved levels of expression in the hybrid lineage**

Among the 7,114 genes expressed in ovary and 8,673 genes expressed in testis of C3, conserved inheritance pattern was predominant for both testis (85%) and ovary (73%). A total of 3% of genes were under the additive category in the testis and 5% in the ovary, while the remaining genes were misexpressed in C3. Four hundred thirteen of the misexpressed genes in the testis and 772 in the ovary were either over- or under-dominant in C3. Number of genes following the D. n. nasuta dominance was twice (270 genes) when compared to D. n. albomicans dominance (133 genes) in the ovary. This trend was the opposite for testis transcriptome, where only 105 genes showed D. n. nasuta dominance and 207 showed D. n. albomicans dominance. Few genes that are over-/under-dominant in C3 also exhibit D. n. nasuta/ D. n. albomicans dominance. The results are summarized in Fig. 5.
Validation of DEGs through RT-qPCR

Six DEGs each for ovary and testis were validated using RT-qPCR to check the reliability of the differential expression results quantified by WTS. Of the six, five genes closely tallied DEG analysis results conducted on ovarian Illumina sequencing data (Fig. 6A and 6B). One gene antdh was not differentially expressed in C3 relative to both the parents. Among the testis transcriptomes, three genes closely matched the DEG analysis results conducted on Illumina sequencing data (Fig. 6C and 6D). For the genes E(spl)m3-HLH and CG3339 expression were detected only in the C3 sample. This trend was similar to WTS data recorded for these genes. Only gene pug did not coincide with the WTS data recorded for these genes. Only gene pug did not coincide with the WTS data and was found upregulated in C3 than the parental testis transcriptomes. This demonstrated the reliability of the WTS technique to quantify gene expression. Overall, RT-qPCR validations were in good agreement with WTS data.

Discussion

This study explores gene expression divergence patterns in an evolving population of a hybrid lineage, C3. This stable population of ~850 generations represents nearly 30 years of laboratory evolution. The results of our comparative transcriptome analyses with parental lineages revealed differential expression of genes in the gonadal tissues of C3. Besides, the expression divergence recorded in this model surpasses the expression divergence recorded between the parents in the ovarian (8%) and testis transcriptomes (4%) (Supplementary Fig. 5). The genome of C3 being an admixture of the two parental genotypes emanated from chromosome recombination emphasizes the potential impact of hybridization on the genome, invariably leading to C3 acquiring unique and more divergent expression patterns in a brief duration.

Our analysis showed six inheritance patterns a gene could follow as a consequence of hybridization in a genome. A significant proportion of the conserved inheritance pattern of expressed genes observed in ovaries (73%) and testes (85%) of C3 is similar to earlier reports made in the fertile hybrids obtained with D. pseudoobscura species [13] and D. mojavensis/D. arizonae [38] with a recent
divergence history. With an estimated divergence time of 0.3-0.5 MY between D. n. nasuta and D. n. albomicans, the parental lineages of C3 equally represent a recently diverged sibling species pair with D. n. albomicans diverging from D. n. nasuta or nasuta like an ancestor [39,40]. Therefore, the recorded proportion of conservation in gene expression could be attributed to the recent divergence of the parental races, further suggesting the process of hybridization has not affected these genes. Co-adaptation between cis and trans-regulatory networks is documented in sea urchins to mammals [41,42].

This co-adaptation might be maintaining the optimal gene expression levels in the conserved category. An asymmetrical pattern of expression divergence was observed in the non-conserved category. More genes in C3 were differentially expressed relative to the maternal parent D. n. albomicans in ovarian and testis transcriptomes. Among the non-conserved categories, a proportion followed D. n. nasuta dominance while a minor portion indicated D. n. albomicans dominance in the ovarian transcriptome, indicating that the alleles of these expressed genes might be from D. n. nasuta and is favored over D. n. albomicans in the ovaries and the testis, the trend was opposite with more genes showing D. n. albomicans dominance. A plausible explanation for this bidirectional preference could be due to the unequal amounts of parental genetic content found in C3. When C3 was initially derived, it contained precisely 50% genome from each parent, and in subsequent generations, parental chromosome selection coupled with recombination may have led to unequal contributions. This is evident in the present stabilized karyotype of C3, which resembles D. n. nasuta and comprising 40% of D. n. albomicans chromosomes. Differential expression can potentially influence phenotypic and trait changes [43], with the upregulation, mainly conferring benefits to an organism while the down-regulation precluding benefits. Assessment of functionality and pathway categorization of the common DEGs in ovaries and testis of C3 revealed a broader distribution of the genes across multiple pathways, majorly comprising metabolism-related processes. The number of genes distributed across

Fig. 4. Pairwise differential gene expression analysis in C3 transcriptomes. Log2-fold change (M) is plotted on the X axis and the absolute value of the difference in expression (D) on the Y axis: C3 and Drosophila nasuta nasuta ovaries (A), C3 and Drosophila nasuta albomicans ovaries (B), C3 and D. n. nasuta testes (C), C3 and D. n. albomicans testes (D). Black points represent all statistically insignificant non-differentially expressed genes and red points represent significant differentially expressed genes. C3, Cytorace-3.
metabolism-related pathways is primarily downregulated. The results hint that these genes might influence the short life span recorded for C3 [44]. Although this is speculative, the role of known candidate genes related to metabolism and aging needs a detailed investigation. Additionally, it provides an opportunity of identifying potential novel longevity-related genes in this model.

Based on existing research in C3, we have an understanding of divergence in this model. C3 differed from parental lineages and other cytoraces for key life-history related traits like lifetime female fecundity, lifetime fertility, and ovariole numbers, which represent

---

**Fig. 5.** Inheritance patterns of gene expression levels deduced in C3 transcriptomes. (A) Representation of six inheritance patterns of gene expression levels. Conserved: gene which is not differentially expressed in C3 in comparison with both the parents and also not a DEG between the parental transcriptomes, Over-dominant: gene for which expression is significantly higher in the C3 than both parents, Under-dominant: gene for which expression is significantly lower in the C3 than both parent, Additive: gene which is differentially expressed in parents and show intermediate expression in C3, and nasuta/albomicans dominant: gene which is differentially expressed in C3 and is exhibiting similar patterns of expression with a corresponding parental lineage. D. n. n corresponds to *Drosophila nasuta nasuta* and D. n. a corresponds to *Drosophila nasuta albomicans*. (B) Expression inheritance deduced in the C3 ovary. (C) Expression inheritance deduced in the C3 testis. C3, Cytorace-3.
indicative measures of population fitness [45]. These population differences could be attributed to the observed expression patterns as these traits are polygenic and mostly influenced by genes having a function in the ovary and testis. In Drosophila, nearly 76% of all the genes are expressed in the testis, and 47% of the genes are expressed in the ovaries providing the transcriptional diversity required to control the mechanisms of spermatogenesis and oogenesis, respectively [3]. With the amount of gene expression conservation that is quantified, it can be argued that the conserved genes are contributing to the normal functioning of the ovaries and testis in the C3. The quantified expression divergence could be contributing to the observed phenotypic variations, with the gonads seemingly tolerating the expression divergence.

Novel expression profiles can evolve in the hybrids when two independently evolving lineages are crossed together. The consequences of this co-evolution of the expression networks in the genetic background of a hybrid can be deleterious. Studies have shown various abnormalities affecting the reproductive system leading to complete sterility [46,47]. Regulatory incompatibility models have shown hybridization-induced profound effects at a transcription level with consequences like misexpression of genes in the F1 hybrids of D. melanogaster and D. sechellia [14]. Under-expression of genes with gonadal atrophy was recorded in female hybrids of D. melanogaster and D. simulans accompanied with significant over-expression of male-biased genes [15]. Significant misexpression for genes of spermatogenesis and reproductive proteases was seen in sterile hybrids of Drosophila [13,16]. Misexpression of regulatory factors and metabolic regulatory genes was observed in the hybrids of Drosophila [19]. However, alternatively, the co-evolution of expression networks in the hybrids might constitute novel and heritable gene expression patterns imparting beneficial effects. Evidence of altered gene expression levels in hybrids contributing to transgressive hybrid phenotypes is well documented in plants [48,49]. Over time, independently evolving lineages accumulate genetic variations in regulatory and coding sequences caused by genetic and environmental factors like a re-

Fig. 6. RT-qPCR validation of the transcriptome data. Log2-fold changes (Log2FC) between C3 and Drosophila nasuta nasuta ovaries (A), C3 and Drosophila nasuta albomicans ovaries (B), C3 and D. n. nasuta testis (C), and C3 and D. n. albomicans testis (D) are shown along Y axis for genes represented on X axis. White bars indicate Log2FC evaluated by Illumina sequencing data (WTS-NOISeq) and black bars indicate Log2FC derived from RT-qPCR. RT-qPCR results were in good agreement with Illumina sequencing analysis except for the gene antdh in the ovary and pug in testis which failed to match the WTS pattern of fold change. RT-qPCR, real-time quantitative polymerase chain reaction; C3, Cytorace-3; WTS, whole transcriptome sequencing.
spontaneous mechanism to the evolutionary challenges imparted on an adapting organism. Even in hybrid populations, in which two parental genomes clash, such response mechanisms are expected to occur. Therefore, the accumulated genetic variations in a persisting hybrid population could have stemmed from such response mechanisms. Before evolving into an independent genetic entity, the hybrid products of *D. n. nasuta* and *D. n. albomicans* exhibited F1 heterosis and F2 breakdown for a few critical parameters of fitness [50]. Fertility tests in F2 and backcross progeny recorded a more considerable number of sterile males than females [50]. Fertile surviving recombinants reproduced, and with the gradual diminishment of karyotypic mosaicism, the stabilized forms became established populations [6,7].

Cytoraces are thought to evolve at a more accelerated rate due to hybridization and inbreeding, which may lead to faster fixation of variations than the natural populations. This accelerated accumulation of differences have been demonstrated through several molecular studies in cytoraces for increased genetic variability in inter-simple sequence repeat [51], nucleotide variations observed in Sod1 and Rpd3 genes, higher levels of RPD3 and SIR2 proteins [44], *always early* (*aly*) a meiotic arrest gene has shown eight times greater rate of substitutions amongst cytoraces than their parents and amongst species of subgenera [9]. Empirical evidence of assortative mating in the C3 males [52] is the first report from the *nasuta-albomicans* complex of *Drosophila*, making it a potential model for studying hybridization-induced behavioral trait evolution. The assortative mating contributes to premating isolation and is more likely to evolve if the parental species are intermediately diverged [20]. Though complete reproductive isolation from their parents has been unachieved in the cytoraces, especially in C3, they constitute a distinct lineage due to the mixed ancestry. Examples of such hybrid taxon with incomplete reproductive barriers are reported in Oxford ragwort [53], Appalachian swallowtail butterflies [54], Cottus fishes [55], and Italian sparrows [56]. The available literature and the quantified expression divergence in C3 reflect the extent to which hybridization has altered the genetic architecture permitting a sustainable population.

In our attempt to profile the gonadal gene expressions in the laboratory evolved *Drosophila* hybrid lineage through comparative transcriptome approaches, we have recorded an accumulated expression divergence which surpasses that between parental lineages illustrating the strong impact of hybridization driving gene expression changes in a brief duration. This transcriptome dataset represents the first in an independently evolving homoploid hybrid lineage exhibiting incipient speciation in *Drosophila*. This documentation in the gonads prompts an extension from the whole organism's perspective and specific somatic tissues in this model. Besides, an evolutionarily unfinished product like C3 and in general cytoraces in their current incipient stages provides more information about the process of ongoing hybridization-induced changes than a finished evolutionary product. Our study contributes primarily toward the enrichment of genomic resources in the context of ecological speciation and provides an unprecedented global view of the transcriptomes for fundamental support in future research.

**ORCID**

Koushik Ponnanna: https://orcid.org/0000-0003-3821-3493
Stanfy M. D’Souza: https://orcid.org/0000-0001-5618-6930
Nallur B. Ramachandra: https://orcid.org/0000-0002-0919-7781

**Authors’ Contribution**

Conceptualization: NBR. Data curation: KP, SMD. Formal analysis: KP. Funding acquisition: NBR. Methodology: KP, SMD. Writing - original draft: KP. Writing - review & editing: KP, SMD, NBR.

**Conflicts of Interest**

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

**Acknowledgments**

We thank the Department of Studies in Genetics and Genomics, the University of Mysore for providing the infrastructure. We thank Dr. Amruthavalli C. and Dr. M.S. Ranjini (DBT project co-investigators) for the support. We thank Rajanikanth C. for generous help with the illustrations. This work was supported by Department of Biotechnology (DBT), Ministry of Science and Technology, Government of India (grant number BT/PR9871/BID/7/472/2013) to NBR; DST-KSTePS Fellowship (LIF-02) from Karnataka Science and Technology Promotion Society to K.P and DST-IN-SPIRE Fellowship (IF150898) from Department of Science and Technology, Government of India to S.M.D.

**Supplementary Materials**

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

31. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioin-
formatics 2011;12:323.


https://doi.org/10.5808/gl.20051
Computational evaluation of interactions between olfactory receptor OR2W1 and its ligands

S. June Oh

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

Mammalian olfactory receptors are a family of G protein-coupled receptors (GPCRs) that occupy a large part of the genome. In human genes, olfactory receptors account for more than 40% of all GPCRs. Several types of GPCR structures have been identified, but there is no single olfactory receptor whose structure has been determined experimentally to date. The aim of this study was to model the interactions between an olfactory receptor and its ligands at the molecular level to provide hints on the binding modes between the OR2W1 olfactory receptor and its agonists and inverse agonists. The results demonstrated the modes of ligand binding in a three-dimensional model of OR2W1 and showed a statistically significant difference in binding affinity to the olfactory receptor between agonists and inverse agonists.

Keywords: docking score, homology modeling, molecular docking, olfactory receptor

Introduction

Olfactory receptors (ORs), which belong to the G protein-coupled receptor (GPCR) family, are cell membrane receptors with a unique structure consisting of seven transmembrane regions. Many three-dimensional (3D) structures of these seven-transmembrane receptors (7TMRs) have been revealed using techniques such as X-ray crystallography and cryogenic electron microscopy. However, the overall 3D structure of ORs, which play a very important role in the external sensory and signal transmission system of animals, has not yet been reported. Accordingly, the binding mode of ORs and their ligands remains to be elucidated.

According to recent reports, ORs, which are generally considered to function as chemosensors in the olfactory organs, appear to be expressed throughout the mammalian body [1,2]. Due to the importance of the various roles and functions of ORs in the body, the discovery of ligands that bind to ORs and the study of the binding mechanism have emerged as an important topic.

Recently, various studies have been conducted on the relationship between ORs and their ligands [3-5]. According to the tissue-specific pattern of human olfactory receptor 2W1 (OR2W1) mRNA expression from BioGPS (http://biogps.org/#goto=genereport&id=26692), the five tissues with the highest OR2W1 expression are the following: superior cervical ganglion (9.9), liver (7.15), Burkitt’s lymphoma (7.05), cardiac myocytes (6.85), and heart (6.7). In 2012, Adipietro et al. [4] reported ligand selectivity and differences in the receptor potency (EC_{50}) of several primate ORs including OR2W1, and we adopted the results of their report in the current experiment.

The present study was designed to develop a homology model of OR2W1 and to in-
vestigate the relationship between the receptor and its corresponding agonists and inverse agonists by simulating the docking modes.

**Methods**

**Homology template for OR2W1**

Molecular docking is a computational procedure that attempts to predict noncovalent binding of a macromolecule (e.g., a membrane receptor) and a small molecule (e.g., a ligand) [6]. To study the interaction between OR2W1 and its ligands, we started with the receptor responses to 42 chemically diverse odorants reported previously by Adipietro et al. [4]. The amino acid sequence for OR2W1 (UniProt: Q9Y3N9) was obtained from UniProt KnowledgeBase (https://www.uniprot.org/) [7]. To pick out the amino acid sequences of 7TMRs with a higher BLAST score (Bits) than 45 or a lower E-value than 1e-07 [8,9], we analyzed the amino acid sequence of OR2W1 against the locally-built BLAST database of 7TMR amino acid sequences registered in the RCSB Protein Data Bank (PDB) [10,11]. Based on their similarity results for OR2W1, four PDB entries of three 7TMRs were selected as the experimental templates for the homology modeling of OR2W1: human serotonin receptor (HTR2A), bovine rhodopsin (Rho), and two turkey β1-adrenoceptor (ADR1B1) PDB models.

In order to apply the activated structure of the receptor as a template, we adopted the PDB model of each receptor binding to an agonist or the corresponding G protein: HTR2A (PDB code: 6WHA; BLASTP score [E-value]: 50.4 bits [1e-08]), Rho (PDB code: 5TE3; BLASTP score [E-value]: 48.5 bits [4e-08]), ADR1B1 (PDB codes: 61BL, 6H7J; BLASTP score [E-value]: 48.5 bits [4e-08]). For these four PDB models, models were built from multiple templates using MODELLER (R. 9.25) [12]. Multiple sequence alignment (MSA) of the four PDB models was executed using Clustal Omega [13]. We applied the MSA result to MODELLER and the application automatically combined these four templates to build the model for OR2W1 using information from multiple templates to build the 3D structural model of OR2W1. After confirming the 3D model of OR2W1 with a Ramachandran plot, it was used for the subsequent docking experiment.

In addition to the confirmation of the 3D model of OR2W1, tools such as HMMTOP [14] and Phobius [15] were used to determine the transmembrane region of the OR. Hydrophobicity analysis was also conducted to verify the transmembrane region using the hydrophobicity scale from Kyte and Doolittle [16].

**Molecular docking and scoring**

Automated molecular docking is widely used for the prediction of receptor-ligand complexes in interaction analysis and molecular design. There are several freely available programs for molecular docking analysis, including AutoDock4 [17], AutoDock Vina [6], idock [18], and smina. Smina was created as a fork of AutoDock Vina to provide enhanced support for minimization and scoring [19]. Based on a report that the above scoring methods fared worse, on average, than simply using the output from smina alone [20], subsequent docking and scoring experiments were performed using the smina program.

We utilized AutoDockTools4 (ADT4) [17], which accompanies AutoDock4, to adjust the experimental conditions of the 3D docking space inside the membrane receptor. The binding site grid box was visually defined for each receptor by employing the grid setting feature of ADT4. To compare whether the experiment using smina was performed properly, we calculated the agonist binding affinity and conformation of human adenosine receptor A2A (AdoRA2A) as well as OR2W1. We downloaded the 3D structural model of AdoRA2A (PDB code: 5G53) from RCSB PDB and removed its ligand to obtain a structural model of the receptor in an activated conformation. The AdoRA2A model was then subjected to molecular docking with N-ethyl-5’-carboxamidoadenosine (NECA) using smina.

Each run with smina was executed using the default parameters with the exception of the 3D coordinates of the search space, so that the program output nine docking poses for each run. As smina accepts a ligand in the SDF format, the 3D SDF files of 22 small molecules were downloaded from PubChem of NCBI and used for the docking experiment with the corresponding receptor.

**Statistical evaluation of binding affinities**

We hypothesized that there would be a difference in binding affinity between OR2W1 agonists and inverse agonists, and tried to verify the significance of the difference. Based on the OR2W1 ligand binding data from the previously reported experimental results [4], we statistically compared the binding affinity scores of agonists and inverse agonists of OR2W1 using a protein modeling and docking experiment. We implemented the statistical evaluation using the R statistical package.

**Results**

**Homology model of OR2W1**

Among the ligands tested for binding to OR2W1, (+)-carvone produced the greatest response to the receptor [4]. The homology model of human OR2W1 generated by MODELLER is shown in Fig. 1. This structure was considered to be stable according to the Ramachandran plot of the preferred model structure (Supplementary Fig. 1). Additionally, the amino acid sequences of the seven
transmembrane regions of the OR2W1 model were compared to those produced by sequence prediction tools such as HMMTOP [14]. The transmembrane domains from the OR2W1 model and the two predicted results of sequence prediction were aligned properly, supporting the validity of the 3D model of OR2W1. Even in the worst case of transmembrane region alignment, only four amino acid residues were short at the N-terminus of the second transmembrane helix compared to those of HMMTOP. Moreover, the amino acid residues of the extracellular side of the receptor transmembrane region almost matched the predicted results of the sequence prediction tool.

Docking, scoring, and analysis of receptor-ligand binding

As a criterion for comparison, we carried out molecular docking of AdoRA2A and NECA using smina according to the method described above. The best binding affinity score obtained from the docking experiment of AdoRA2A and NECA using smina was –8.7 kcal/mol. This docked pose of NECA with AdoRA2A is shown in Fig. 2 (scaled ball and stick).

As shown in Fig. 2, the conformation of the docking model based on the results of smina showed almost the same structure as that previously obtained by X-ray crystallography [21]. The amino acid residue of Asn253 (6.S5; Ballesteros-Weinstein nomenclature [22], in yellow) in AdoRA2A, which interacts with the amino group of NECA, was found to be critical for activation of the receptor in previous studies of the 3D structure of AdoRA2A [23-25].

The homology model of OR2W1 generated with MODELLER and its best docking conformation with (+)-carvone is presented in Fig. 3. The docking conformation of OR2W1 and (+)-carvone showed a minimum binding energy of –7.2 kcal/mol (Table 1).

As shown in Fig. 3, (+)-carvone is surrounded by the third, sixth, and seventh transmembrane domains of the receptor, and is located very close to the Tyr252 amino acid residue in the sixth transmembrane region. This Tyr252 residue is not only at a position similar to the Asn253 residue of AdoRA2A, but also appears to form pi-pi stacking with the agonist (+)-carvone.

Smina predicts the binding of a receptor to a ligand and outputs the corresponding binding affinity along with its positional con-
formation. Although the challenge of selecting the correct docked pose remains [19], the best model selected from each smina prediction not only had a zero distance of the lower and upper bound root-mean-square deviation from the best mode, but also the lowest binding affinity energy between receptor and ligand among the nine suggested modes. The example docking model of AdoRA2A-NECA shown in Fig. 2 was also adopted as described above. The binding affinity of the best model obtained from each docking experiment is shown in Table 1. The mode that scored best according to the smina scoring function was chosen as the representative mode and its affinity score was subjected to further statistical analysis.

The average binding affinities of the 12 agonists and 10 inverse agonists were –6.325 and –4.9, respectively. These two binding affinity groups were checked for normality using the Shapiro-Wilk test, and the resulting p-values of the agonist and inverse agonist groups were 0.5229 and 0.08436, respectively. Therefore, the binding affinity data from each group were considered to follow a normal distribution. The F-test was performed to determine whether the variances of the two groups were homogeneous, and as a result, the p-value was 0.1475. Accordingly, the two-sample t-test assuming equal variances was conducted to evaluate the difference between the average values of the two groups, and the difference was found to be significant (p = 0.01019, two-sided).

In addition, as a result of performing the Wilcoxon rank sum test under the assumption that the data did not follow a normal distribution, the p-value was found to be 0.01333. This suggests that the median values of the two distributions are not equal. From the above results, it seems reasonable that the modeled binding affinity value of the agonists to OR2W1 is lower than that of the OR2W1 inverse agonists.

**Discussion**

Many biological processes are regulated by signaling systems that cross cell membranes. GPCRs, including ORs, are proteins that play an important role in the physiology of higher organisms. ORs play an essential role in responding to changes in the environment by transmitting external signals to the body. In the case of GPCRs, the chemical change and fate of the ligand before and after binding remains unknown, and the 3D structure of the OR has not been revealed. Therefore, studies of specific chemical changes and structural activation mechanisms occurring in the binding process of ORs and ligands are very limited.

In this study, a computational model was generated for the OR OR2W1, the specific structure of which is unknown, and the binding condition with the ligand was simulated. In addition, by revealing a statistically significant difference in binding affinity between agonists and inverse agonists of the ligand, a helpful hint for screening tests to find novel ligands of the OR was provided.

**ORCID**

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

**Conflicts of Interest**

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

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

References

Machine learning based anti-cancer drug response prediction and search for predictor genes using cancer cell line gene expression

Kexin Qiu¹, JoongHo Lee¹, HanByeol Kim¹, Seokhyun Yoon¹², Keunsoo Kang³

¹Department of Computer Science, Dankook University, Yongin 16890, Korea
²Department of Electronics and Electrical Engineering, Dankook University, Yongin 16890, Korea
³Department of Microbiology, Dankook University, Cheonan 31116, Korea

Although many models have been proposed to accurately predict the response of drugs in cell lines recent years, understanding the genome related to drug response is also the key for completing oncology precision medicine. In this paper, based on the cancer cell line gene expression and the drug response data, we established a reliable and accurate drug response prediction model and found predictor genes for some drugs of interest. To this end, we first performed pre-selection of genes based on the Pearson correlation coefficient and then used ElasticNet regression model for drug response prediction and fine gene selection. To find more reliable set of predictor genes, we performed regression twice for each drug, one with IC₅₀ and the other with area under the curve (AUC) (or activity area). For the 12 drugs we tested, the predictive performance in terms of Pearson correlation coefficient exceeded 0.6 and the highest one was 17-AAG for which Pearson correlation coefficient was 0.811 for IC₅₀ and 0.81 for AUC. We identify common predictor genes for IC₅₀ and AUC, with which the performance was similar to those with genes separately found for IC₅₀ and AUC, but with much smaller number of predictor genes. By using only common predictor genes, the highest performance was AZD6244 (0.8016 for IC₅₀, 0.7945 for AUC) with 321 predictor genes.

Keywords: cell line gene expression data, drug response prediction, machine learning, predictor genes

Introduction

Cancer is one of main causes of death worldwide. Anti-cancer drug therapy is an important part of cancer treatment and an effective use of them can prolong patient’s survival. According to many clinical data, patients with the same cancer have quite different response to the same treatment or the same drugs due to genomic specificity. Recently, targeted anti-cancer therapy [1,2] considering gene-specific effects has been proposed as a new cancer therapy. In order to develop specific targeted therapy for cancer patients in clinical treatment, many clinical trials are required. However, there are many obstacles such as sample limitations, complicated operations, high environmental requirements, and high cost, which far from meeting the demand.

With the rapid development of artificial intelligence, many machine learning based
drug response prediction models were proposed utilizing genomic information and anti-cancer drug response data. In 2011, Riddick et al. [3] used the random forest algorithm to establish a regression model of drug response, and successfully predicted the drug response of 19 breast cancer and seven glioma cell lines, which was advanced to other methods such as based on differential gene expression. In 2014, Gheeleeher et al. [4] used Ridge regression based on baseline gene expression levels and in vitro drug sensitivity in cell lines to establish a regression model and used it to predict clinical drug response. On the other hand, some studies have shown that the structural similarity between drugs may have similar response to cancer cell lines that have similar gene expression profile [5-7]. Specifically, Shivakumar and Krauthammer [8] reported that the similarity between drugs is useful to predict the drug response. Based on this research background, we designed an improved drug response prediction model based on cancer genomics data and explored the predictor genes possibly related to the drug response.

**Methods**

**Data**

The data used in this work is from Genomics of Drug Sensitivity in Cancer (GDSC) [9] which was developed by the Sanger Research Institute in the United Kingdom. We considered 12 drugs and gene expression data for 1,000 human cancer cell lines. The drug response indicators used were the half maximal inhibitory concentration (IC$_{50}$) and the area under the curve (AUC) [10]. The former is the concentration at which the compound reaches 50% reduction in cell viability and the latter is the area under the fitted dose response curve. Biologically, the smaller the IC$_{50}$ and AUC, the greater the response of the cancer cells to the drug.

**Method**

Based on the gene expression data of the cancer cell lines and the two types of response indicators, we used a machine learning algorithm to construct a drug response prediction model. We first pre-selected genes based on the p-value of Pearson correlation coefficients [11] and then used ElasticNet to predict drug response and to further select the predictor genes among the pre-selected ones. Specifically, we performed ElasticNet regression separately on the two response values, from which common predictor genes were identified. These common genes were used again to predict drug response hoping that the prediction performance is better than, or at least similar to, those obtained separately for the two response indicators. To confirm biological significance of predictor genes, we provide heatmap and gene ontology analysis results. Fig. 1 shows the entire experimental workflow.

**Preprocessing**

Before processing the data, we took logarithm on IC$_{50}$ and normalized the cell line gene expression data using the robust multichip average [12].

**Feature selection based on Pearson correlation coefficient**

For some drugs, there are thousands of genes in the gene expression data, but not many genes have strong correlation with the drug responses. Therefore, it is very important to pre-select the relevant genes first. Although ElasticNet has capability for gene selection, it is subject to data dependency and/or batch effect and, sometimes, it ignores genes that are really important to predict drug responses. In this paper, to overcome such problem, we used two-step gene selection, where we first used the Pearson correlation coefficient to pre-select genes and then applied ElasticNet to fine select the predictor genes. In particular, we used p-value of Pearson correlation coefficient between the drug response and the expression of each gene, with which genes with p = 0.05 or less were selected in the first feature selection.

**ElasticNet-based feature selection and drug response prediction**

ElasticNet [13] is a linear regression model trained with both ℓ₁ and ℓ₂ regularization. It is useful when there are so many features that are correlated with one another. In our data, the number of features (genes) is much larger than the number of samples and the prediction might be subject to overfit. Hence, to appropriately select genes and to suppress generalization error, we used ElasticNet to predict the drug response. The ElasticNet was selected based on the preliminary experiments where we compared ElasticNet with two well-known models, SVR [14] and Xgboost [15]. The former can be configured to a non-linear regressor by using various kernel functions and we used radial basis function kernel and the latter is an improved version of decision tree based gradient boosting algorithm. The two algorithms were shown to perform good for many applications, while, according to our preliminary experiments, they seem to have higher overfit than ElasticNet as the numbers of predictor genes that are common for the two response indicators were smaller than that for the ElasticNet. Fig. 2 summarizes the comparison for the 12 drugs in terms of Pearson correlation coefficients between the predicted IC$_{50}$ and the measured ones.
**Fig. 1.** Experimental workflow. GDSC, Genomics of Drug Sensitivity in Cancer; AUC, area under the curve.

**Fig. 2.** A comparison of three regression methods in terms of Pearson’s correlation coefficients (PCC) between the predicted IC₅₀ and the measured ones. GDSC, Genomics of Drug Sensitivity in Cancer.
Results

Prediction of IC\textsubscript{50} and AUC

In the first experiment, we predict the two drug response indicators, IC\textsubscript{50} and AUC separately. In ElasticNet, there are two key hyper parameters, a.k.a. the penalty weight $\alpha$ and the relative weight of $\ell_1$ penalty $\lambda$, where $\alpha$ is an arbitrary positive real while $0 \leq \lambda \leq 1$. $\lambda = 0$ corresponds to the Ridge regression, where we have only $\ell_2$ penalty while $\lambda = 1$ corresponds to LASSO regression where we have $\ell_1$ penalty only. These two hyper parameters must be optimized to achieve the best performance. To this end, we performed grid search for a set of combinations $(\alpha, \lambda)$. Through this, the best performance for drug response prediction were obtained for the 12 drugs as summarized in Table 1.

For all the 12 drugs, the correlation coefficient between the estimated IC\textsubscript{50} and the true ones were higher than 0.65, where three of them were reached 0.8, e.g., 0.823 for AZD6244, 0.819 for Nutlin-3a and 0.811 for 17-AAG. Similar performances were also observed for AUC, where 17-AAG and nilotinib showed correlation coefficient exceeding 0.8. The results seem to be statistically significant as the p-value for the correlation coefficient of each drug was less than 0.01.

Of note, not only the number of genes to obtain the optimal predictive performance were quite different for each drug but also the gene sets for the two response indicators of the same drugs were only partly overlapped. The latter suggest us that there might exist dependency on the response indicators and it would be interesting to check the prediction performance using the common predictors. Hopefully, they will be more reliable predictors of the practical drug responses.

Drug response prediction based on common predictor genes

In the previous experiments, we found the predictor genes separately for the two response indicators and it is also interesting to evaluate the performance when using only common genes. It could be a more stable group of predictor genes for drug response. To find the predictor genes that are commonly effective in the two response indicators, the relative weights of the $\ell_1$ and $\ell_2$ of the ElasticNet were fixed to 0.5. Then we adjusted $\alpha$ to make the number of selected genes for the two response indicators are similar to each other and then took the intersection of them to obtain the common predictors to be used for the drug response prediction. The results are summarized in Supplementary Fig. 1 and the scatter plots of the predicted versus the true responses were shown in Supplementary Fig. 2 for IC\textsubscript{50} and 3 for AUC.

As the number of common genes increases, the predictive performance for each response indicators changed similarly, but it is confirmed that the performances were saturated or slightly decreased after reach the peak. According to the trend of the prediction accuracy curve, we found the points at which the performance was the best for both the drug response indicators simultaneously. The results are summarized in Table 2, where the Pearson correlation coefficients for IC\textsubscript{50} and AUC of six drugs were higher than 0.7 only with 200 predictor genes. Comparing with the results in Table 1 for separate predictors for each response indicator, it can conclude that even with only those common predictor genes, one can have similar predictive performance suggesting that these genes are more reliable predictors on the two response indicators and are closely related to the underlying biological mechanism that governs the drug response. For comparison, we provided the performance of the drug responses in the literature for the same GDSC dataset in the last column of Table 2.

Table 1. Comparisons of the PCC between the estimated response and the true value for the 12 drugs in GDSC

<table>
<thead>
<tr>
<th>Drug name</th>
<th>Predict IC\textsubscript{50}</th>
<th>Predict AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of features</td>
<td>PCC</td>
</tr>
<tr>
<td>17-AAG</td>
<td>566</td>
<td>0.811</td>
</tr>
<tr>
<td>AZD-0530</td>
<td>262</td>
<td>0.612</td>
</tr>
<tr>
<td>AZD6244</td>
<td>570</td>
<td>0.823</td>
</tr>
<tr>
<td>Erlotinib</td>
<td>253</td>
<td>0.603</td>
</tr>
<tr>
<td>Lapatinib</td>
<td>261</td>
<td>0.698</td>
</tr>
<tr>
<td>Nilotinib</td>
<td>475</td>
<td>0.782</td>
</tr>
<tr>
<td>Nutlin-3a</td>
<td>475</td>
<td>0.819</td>
</tr>
<tr>
<td>PD-0325901</td>
<td>570</td>
<td>0.775</td>
</tr>
<tr>
<td>PD-0332991</td>
<td>527</td>
<td>0.743</td>
</tr>
<tr>
<td>PHA-665752</td>
<td>224</td>
<td>0.635</td>
</tr>
<tr>
<td>PLX4720</td>
<td>499</td>
<td>0.715</td>
</tr>
<tr>
<td>Sorafenib</td>
<td>297</td>
<td>0.619</td>
</tr>
</tbody>
</table>

PCC, Pearson’s correlation coefficient; GDSC, Genomics of Drug Sensitivity in Cancer; AUC, area under the curve.
GSDC data set also provides binary indicator of drug response, with which the cell lines are labelled as either “sensitive (S)” or “resistant (R)” to a specific drug. And it would be interesting if the two groups show non-negligible difference in the expression of the predictor genes or not. Fig. 3 shows the heatmap for the predictor genes for four drugs, where we can identify the differences in their expressions between the two group and can qualitatively judge the effectiveness of the predictor genes we found. The heatmap analysis [16] shows that the predictor genes can also distinguish the drug sensitivity of cell lines to a certain extent, even though it is not our focus in this work. Rather, it would be more interesting to check what biological processes these genes are involved in response to a certain drug treatment.

**Notes on biological implication of the predictor genes**

To show the biological implication of the drug response, we used Metascape [17] to perform gene enrichment analysis. The predictor genes for the 12 drugs were listed in Supplementary Table 1 and the results of gene enrichment analysis for the 12 drugs are shown in Supplementary Table 2. Through the enrichment analysis of predictor genes, we found various pathways that were mostly related to cancer, such as cell proliferation and developmental process. For example, the negative regulation of cell population proliferation (GO:0008285) is a process that stops, prevents, or reduces the rate or extent of cell proliferation [18]. If predictor genes of drug found by machine learning are in this pathway, this drug may be effective for cancer.

Of note, AZD6244 is an inhibitor of the MAPK cascade [19]. The predictor genes we found were confirmed to be related to the regulation of the MAPK cascade through the enrichment analysis. Nutlin-3a is known to be an inhibitor of the MDM2-p53 (TP53) interaction [20]. The first significant pathway of the predictor genes appeared to be the p53 downstream pathway. It can be seen that some genes that are important to predict drug response are related to the mechanism of drug action. For example, NQO1 found to be one of the predictor genes of 17-AAG, the overexpression was known to increase the sensitivity to the drug 17-AAG [21]. Among the predictor genes of Nutlin-3a, the regulation of HIPK2 determines the response of tumor cells to the p53 activating drug Nutlin-3a [22]. platelet-derived growth factor receptor A, one of the predictor genes of PD-0332991, is known to play an important role in cell signaling pathways that affect cell growth and differentiation and are associated with an array of clinically significant neoplasms [23]. For other drugs, it may be a new mechanism of action for drugs which is not yet known.

**Discussion**

Although the model proposed in this study shows good predictive performance for GDSC, there are still some limitations. First, the characteristic of cancer cell line may be quite different from the in vivo cancers and it should be verified whether this will be effective in clinical trial. Second, we perform drugs response prediction mainly based on gene expression data. While, the response of drugs is not only related to gene expression levels, but also to structural variations such as gene mutations. Therefore, more study is required to utilize such information and integrate them into the model to improve the predictive power.

Cancer is one of the leading causes of death worldwide. If one can find a new treatment by accurately predicting drug response, the probability of recovery will also be increased. Although there are still huddles to overcome in drug response prediction, advanc-
es in machine learning techniques will make it possible to introduce new ideas for drug response prediction that can provide accurate drug treatments and make it practical for clinicians and non-experts.

**ORCID**

Kexin Qiu: https://orcid.org/0000-0002-5606-1860  
JoongHo Lee: https://orcid.org/0000-0003-4729-9938

HanByeol Kim: https://orcid.org/0000-0001-5149-220X  
Seokhyun Yoon: https://orcid.org/0000-0002-0464-2233  
Keunsoo Kang: https://orcid.org/0000-0003-0611-9320

**Authors’ Contribution**

Conceptualization: KQ. Data curation: KK, SY. Formal analysis: KQ, HK. Funding acquisition: SY. Methodology: KQ, JL. Writing

---

**Fig. 3.** Heatmap for the predictor genes of the four selected drugs: for lapatinib (A), for nilotinib (B), for PD-0332991 (C), and for sorafenib (D). The type abbreviation S stands for “sensitive” and R for “resistant.”
Conflicts of Interest

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

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1F1A1066320).

Supplementary Materials

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

References

Introduction

The novel coronavirus disease 2019 (COVID-19) presents an important and urgent threat to global health. Since the outbreak in early December 2019 in the Hubei province of China, the number of patients confirmed to have the disease has exceeded 118 million as the disease spread globally, and the number of people infected is probably much higher [1]. More than 2.6 million people have died from COVID-19 (up to 11 March 2021) [2]. Despite public health responses aimed at containing the disease and delaying the spread [3,4], several countries have been confronted with a critical care crisis, and more countries could follow [5].

To mitigate and suppress the burden of COVID-19 on the healthcare system, while also protecting the general public, especially the highly susceptible group of people, re-
bust models that predict the prognosis of COVID-19 were urgently needed to support decisions about shielding, hospital admission, treatment, and population level interventions [6]. In this situation, prediction tools can help project different scenarios such as (1) number of possible confirmed (new) cases, (2) number of possible hospitalized cases, (3) number of possible death cases and so forth. As a consequence, prediction tools are useful for several different purposes [7].

Other features, such as social distancing, stay-at-home orders, use of facemasks or self-quarantine, travel restriction, and contact tracing could help predict what comes next. For better understanding, prediction models are important for better estimation about the disease and its possible threats such as the number of cases based on the level of severity can help determine the need of numbers of ventilators and other sophisticated medical equipment. Furthermore, countries need to shape their health system responses in accordance with the need [8]. Therefore, access to accurate outbreak prediction models is essential to obtain insights into the likely spread and consequences of infectious diseases. Governments and other legislative bodies rely on insights from prediction models to suggest new policies and to assess the effectiveness of the enforced policies [9].

For COVID-19, predictive modeling, in the literature, uses broadly susceptible exposed infected recovered (SEIR/SIR), agent-based, curve-fitting models. Besides, machine-learning models that are built on statistical tools have widely been used too [7]. Here, we employ statistical models; segmented Poisson, negative binomial (NB), and local likelihood regression (LLR), mathematical model SEIR, deep-learning based model long short-term memory (LSTM), and tree based gradient boosting machine (GBM) for prediction of future COVID-19 pandemic situation of Korea. The COVID-19 daily confirmed cases of Korea was divided into two regions: capital area (Seoul metropolitan area) and non-capital area (non Seoul metropolitan area). Domestic which is the sum of Capital and non-capital areas was also analyzed (Fig. 1). The daily confirmed cases of these regions were then split into train (January 20, 2020–December 31, 2020 and January 20, 2020–January 31, 2021) and test (January 1, 2021–February 28, 2021 and February 1, 2021–February 28, 2021) datasets. The prediction performance of the above models were tested using relative mean square error (RMSE). RMSE takes the total squared error and normalizes it by dividing by the total squared error of a simple predictor. Thus, the smaller the RMSE value, the better the prediction performance of the model.

Therefore, with increasing daily confirmed cases since the beginning of 2021, in Korea and elsewhere, such models could help in the response to pandemic by informing decisions about planning, resource allocation, and decision of the social distancing.

## Methods

### COVID-19 confirmed cases data

The daily series of confirmed cases of COVID-19 for South Korea from January 20, 2020 to February 28, 2021 was obtained from...
Kaggle (from January 20 to June 30, 2020) [10] and Korea public data portal of the Ministry of Health and Welfare (from July 1, 2020 to February 28, 2021) [11]. The combined data was divided into two regions: capital or Seoul Metropolitan area (capital; Seoul, Incheon, and Gyeonggi-do) and non-capital or non-Seoul Metropolitan area (non-capital; other cities beside Seoul, Incheon, and Gyeonggi-do). The analysis was conducted on the Domestic area (capital and non-capital), Seoul Metropolitan area, and non-Seoul Metropolitan area data, respectively. The data was split into two subsets. First subset is composed of training (January 20, 2020–December 31, 2020) and test data (the last 59 days, January 1, 2021–February 28, 2021). And second subset is composed of training (January 20, 2021–January 31, 2021) and test data (the last 28 days, February 1, 2021–February 28, 2021) for downstream analysis with the test data used for prediction analysis.

Prediction models
As one model may not give the best prediction of the COVID-19 situation of Korea, we present prediction results estimated by different models that can apply the above data. Many models are available and have been implemented for forecasting the pandemic situation of many countries and look at a few. In this section, we introduce segmented Poisson model, LLR model, deep-learning based LSTM model, NB model, SEIR model, and GBM model used for predicting the COVID-19 situation of Korea.

Segmented Poisson model
Here, we regarded the confirmed cases as a function of time $t$ based on a segmented Poisson model. Let $Y_t$ be the confirmed cases at day $t$ which is the number of days since the first case occurred. Poisson model is defined as;

$$Y_t \sim \text{poisson}(\mu_t),$$

where $\mu_t$ is the expectation of $Y_t$ with segments.

Breakpoints were considered in the daily confirmed cases during the analysis by splitting the daily confirmed cases into segments (Supplementary Table 1). These breakpoints were decided using some of the aforementioned significant events linked to the spread of COVID-19 in South Korea. Since there are three breakpoints, four segments are defined as follows:

$$
\log(\mu_t) = \\
\begin{cases} \\
\beta_0 + \beta_1 t + \beta_2 \log(t+1), & (t = 0, 1, \ldots, c_1 - 1) \\
\beta_0 + \beta_1 (t-c) + \beta_2 \log(t-c+1), & (t = c_1, \ldots, c_2 - 1) \\
\beta_0 + \beta_1 (t-c) + \beta_2 \log(t-c+1), & (t = c_2, \ldots, c_3 - 1) \\
\beta_0 + \beta_1 (t-c) + \beta_2 \log(t-c+1), & (t = c_3, \ldots, \infty) \\
\end{cases}
$$

where $c_i (i=1,2,3)$ are breakpoints.

**NB model**

NB model is defined as [12];

$$\lambda_{F_{i+1}} \sim \text{NB}(\lambda_i, \phi),$$

where $\lambda_i$ is the conditional expectation of $Y_t$ given $F_{i+1}$ as the history of the joint process $\{Y_t : t \in \mathbb{N}\}$. Conditional mean and variance of $Y_t$ are defined as;

$$E(Y_t|F_{i+1}) = \lambda_i$$

$$\text{VAR}(Y_t|F_{i+1}) = \lambda_i + \lambda_i^2/\phi,$$

where $\phi$ is the dispersion parameter. And overdispersion parameter $\sigma^2$ is defined as $\sigma^2 = 1/\phi$. NB distribution is defined as;

$$P(Y_t = y | F_{i+1}) = \frac{\Gamma(\phi + y) \Gamma(y + 1)}{\Gamma(\phi) \Gamma(y + \lambda_i)} \left( \frac{\phi}{\phi + \lambda_i} \right)^\phi \left( \frac{\lambda_i}{\phi + \lambda_i} \right)^y$$

where $y = 0, 1, \ldots, n$. For estimating $\lambda_i$, $l = \{1,7,21\}$ were used as lagged confirmed cases and $\log(\lambda_i) = \beta_0 + \sum_{c_i=1}^{3} \beta_i \log(Y_{i+1} + 1)$, were used for the model. For NB model, package ‘tscount’ were used to analyze the confirmed cases of South Korea as time series count data [13].

**LLR model**

Our LLR model is based on Poisson model which previously mentioned. For this model, local quadratic approximation is fitted within a smoothing window of bandwidth $h$, which is the number of the nearest past observations to be used in the local fit. We use tricube kernel of weight $W(u) = (1–|u|)^3$ for each point. The local quadratic log-likelihood is defined as;

$$L(a) = \sum_{i=1}^{n} w_i(t) \{ Y_i \phi_i + a_1(t_i) + a_2(t_i)^2 \},$$

where $w_i(t) = W(\frac{t_i - a}{h})$ and $l$ are the log-likelihood function based on Poisson distribution assumption. The local likelihood estimate is made by maximizing over the parameter $a = (a_{\phi}, a_1, a_2)^T$.

We utilize a rolling origin cross validation to select optimal bandwidth of the smoothing window [14]. Validation sets are divided at the local peaks of counts. Validation MSE is cumulated by each validation set’s counts being predicted using past validations sets. The bandwidth with the smallest validation MSE is selected as optimal bandwidth. And then using optimal bandwidth we finally fitted LLR model. For LLR model [15], package ‘locfit’ was used [16].
**Long short-term memory**

Here, LSTM network is considered [17]. Let the input data $X_t$ be as a set of vector consisting of $Y_{t-h}$ to $Y_{t-1}$ according to day $t$, where $h$ is the bandwidth having values {7, 14, 21, 28, 35, 42, 49, 56}. Among these, the optimal $h$ is selected using validation set, which is last 7 days of the training period. The data is normalized using minmax normalizer to transform data to be in the range of 0 to 1.

The LSTM architecture is described in Fig. 2 [18]. Each blocks in the model use current input value $X_t$ with $C_{t-1}$ and $h_{t-1}$ to be trained. The $C_{t-1}$ and $h_{t-1}$ are the state and output of the last block, respectively. We assume four blocks with 64 units, each with a 0.2 dropout layer. The optimization is held using the adam optimizer to minimize the MSE during the training process.

Once the optimal model is built, it is applied to the test data for prediction. The prediction is performed sequentially by using the current output as the part of input of the next prediction. The analysis was performed using Python version 3.7.6, and ‘keras’ library.

**SEIR model with least squares**

The infectious disease dynamic can be formulated with a mathematical model. We consider the SEIR model to fit the dataset of COVID-19 daily confirmed cases and predict the incidence of COVID-19 epidemic in Korea. In SEIR model, population is divided in four groups: susceptible (S), exposed (E), symptomatic and infectious (I) and recovered (R) individuals. This model includes the spread of infection during the latent period. The latency of COVID-19 infection is biologically realistic. The SEIR model is defined by the following the system of ordinary differential equations [19-22]:

$$\frac{dS}{dt} = -\beta S \frac{I}{N}$$

$$\frac{dE}{dt} = -\beta S \frac{I}{N} - \kappa E$$

$$\frac{dI}{dt} = -\kappa E - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

where $\beta$ is the transmission rate, $\gamma$ is the recovery rate, and $1/\kappa$ is the average incubation period. The initial condition of this model $S(0), E(0), I(0), R(0)$ must satisfy the condition $S(0) + E(0) + I(0) + R(0) = N$, where $N$ is the total population size. In data fitting, the unknown parameters in model were estimated by a least squares algorithm. The numerical simulation and analysis were performed in MATLAB 2020a.

**Gradient boosting machine**

GBM is a tree based machine learning algorithm that can be used for regression and classification problems. GBM consist of weak regression learner and decision trees. The decision tree uses the input value to determine which regression learner is best to make predictions.

Based on adaptive boosting algorithm, GBM can build a strong regression learner by iteratively combining a set of weak regression leaners. GBM use gradient descent for minimizing loss function of a strong regression learner. Like other boosting algorithms, GBM adds models into the tree using greedy style [23]:

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x),$$

where $F_m$ is the updated model, $F_{m-1}$ is previous model and $\rho_m h_m$ is

---

Fig. 2. Long short-term memory (LSTM) model architecture. (A) Overall architecture of LSTM. (B) The LSTM block architecture.
the newly added model. \( h_n \) is the trained base learner which minimizes the loss function \( L \) and \( \rho \) is the multiplier which is found by solving one dimensional optimization problem.

\[
\rho_n = \arg\min \sum_{i=1}^{n} L \left( y_i, F_m^{-1}(x_i) + \rho h_m(x_i) \right).
\]

To build GBM, ‘LightGBM’ library was used [24].

**Model assessment**

To evaluate the above models, RMSEs for the train and test datasets for each of the fitted models were calculated as follows:

\[
RMSE = \sum \frac{(\hat{y}_i - y_i)^2}{n},
\]

where \( n \) is the number of data points, \( y_i \) is the observed values, \( \hat{y}_i \) is the predicted values from a fitted model and \( y^3 \) is the mean of observed values. To compare models predicting different regions, having different scale of confirmed cases, RMSE measure was chosen.

**Results**

The COVID-19 daily confirmed cases of the country were divided into two regions (non-capital and capital) with the total being domestic and analysed using the above models. The data was split into two subsets and used in the training and prediction analysis of the models.

As for model evaluation, in Table 1, for comparison of models in the whole country and the two regions, we observe that the train RMSE is always lower than the test RMSE, with the domestic region producing the highest RMSE values for all models. Also, the segmented Poisson model gives higher RMSE values when compared with other methods. With the first data subset: in the whole country (domestic), SEIR model and GBM had the lowest train RMSE values while NB and LLR had the lowest test RMSE values. In the Capital region, GBM and LLR have the lowest train RMSE while LLR and SEIR have the lowest test RMSE values, respectively. The non-capital region showed that SEIR and GBM have the lowest train RMSE while GBM and LLR have the lowest test RMSE values, respectively.

With the second data subset: in the country, LLR and SEIR had the lowest RMSE while NB and GBM had the lowest train RMSE values, respectively. Capital region showed that LLR and NB had the lowest train RMSE while NB and GBM had the lowest test RMSE values. In the non-capital region, SEIR and LLR had the lowest train RMSE while NB and GBM had the lowest test RMSE values.

Therefore, taking into lower train and test RMSE values for all Table 1. RMSE for the regions and models following the the two data subsets

| Region         | Model                  | RMSE of data split 1 | RMSE of data split 2 | |
|----------------|------------------------|-----------------------|-----------------------|
| Domestic       | Segmented Poisson      | 0.088                 | 1194.103              | 0.251                 | 16.415               |
|                | Negative binomial      | 0.057                 | 0.409                 | 0.063                 | 2.088                |
|                | Local regression       | 0.037                 | 0.793                 | 0.039                 | 14.856               |
|                | LSTM                   | 0.051                 | 23.117                | 0.083                 | 6.327                |
|                | SEIR                   | 0.033                 | 0.956                 | 0.035                 | 2.658                |
|                | GBM                    | 0.022                 | 1.507                 | 0.082                 | 0.591                |
| Capital        | Segmented Poisson      | 0.075                 | 668.199               | 0.235                 | 5.312                |
|                | Negative binomial      | 0.061                 | 1.311                 | 0.064                 | 3.078                |
|                | Local regression       | 0.042                 | 1.135                 | 0.046                 | 3.846                |
|                | LSTM                   | 0.054                 | 14.8                  | 0.074                 | 3.934                |
|                | SEIR                   | 0.073                 | 0.410                 | 0.072                 | 3.109                |
|                | GBM                    | 0.021                 | 1.960                 | 0.095                 | 0.892                |
| Non-capital    | Segmented Poisson      | 0.118                 | 1131.838              | 0.195                 | 34.935               |
|                | Negative binomial      | 0.097                 | 0.912                 | 0.103                 | 1.157                |
|                | Local regression       | 0.074                 | 0.522                 | 0.076                 | 33.232               |
|                | LSTM                   | 0.087                 | 15.207                | 0.119                 | 4.774                |
|                | SEIR                   | 0.036                 | 0.610                 | 0.036                 | 1.964                |
|                | GBM                    | 0.015                 | 0.607                 | 0.083                 | 0.855                |

RMSE, relative mean squared error; LSTM, long short-term memory; SEIR, susceptible exposed infected recoverd; GBM, gradient boosting machine.

https://doi.org/10.5808/gi.21028
region and both data subsets, we can conclude that LLR model, GBM, SEIR model and then NB model were the best prediction models for forecasting of the COVID-19 situation of Korea. Segmented Poisson model tended to have the highest test RMSE values in all scenarios.

A look that the prediction plots of the these models shows that the daily COVID-19 confirmed cases will decline in the country (domestic), Seoul metropolitan (capital) and non-Seoul metropolitan (non-capital) areas using LLR, and NB models using the first data subset, while it will increase and stay constant using the segmented Poisson and LSTM models, respectively (Supplementary Figs. 1–3). With the second data subset, daily COVID-19 confirmed cases will decline in the three regions as predicted by NB, segmented Poisson and LSTM models, while it will increase in the country and non metropolitan areas but will decline in the metropolitan areas, using LLR model (Supplementary Figs. 4–6). The SEIR and GBM models shows a decrease in daily confirmed cases in the country and the two regions for all data subsets (Supplementary Figs. 7 and 8).

Discussion

The objective of our analysis was to predict the future COVID-19 situation of South Korea using daily confirmed cases. We employed six different models in this analysis and all models gave some different prediction results for different data subsets and regions. The evaluation of the models using RMSE showed that local likelihood regression, GBM, SEIR and NB models had the lowest RMSE values, making them the best models, though LSTM gave better RMSE values compared to segmented Poisson model. LLR, GBM, SEIR, NB and LSTM models mainly predicted a decline in COVID-19 daily confirmed cases in the country and the two regions of Korea. We can reasonably take that these results portray the future situation of the country.

With the first dataset, NB, SEIR, and LLR, respectively showed the best test performance in domestic, capital, and non-capital areas, while with second dataset, GBM showed the best test performance for all regions. In case of NB model, we found that the coefficient of confirmed cases of the day before had the largest value. This means that confirmed cases of the day before can affect the prediction of most future confirmed cases. In case of GBM, we could obtain feature importance plots of the model (Supplementary Fig. 9). We discovered that the confirmed cases of the last day was the most important feature for all regions. Thus, the models using the confirmed cases of the past days seemed to perform better than other models without using such data. The parameters of the mathematical model SEIR can be easily interpreted as a rate or transition parameter. However, the local regression does have too many parameters to assign meaningful interpretation.

Note that the number of daily confirmed cases varied across regions of Korea, so we first fitted these models for each of the regions and compared them with the number of observed cases. However, the comparison of prediction models based on regional data was not convincing due to the small sample sizes. Instead of fitting the models for each region, we considered combining the two regions of capital and non-capital areas, which provided enough sample sizes. We also tried predicting the number of deaths due to COVID-19. However, the data was not large enough (with only 1,669 deaths as of March 14, 2021) [25] to provide reliable fitted results from the models.

In our study, there is one challenge that predictions made reflect interventions in place at the time the model was developed. So, one can argue the influence of government intervention policies in the above observed results. However, our comparison result is still valid because all models reflected the same intervention effects. Actually, the Korean government has maintained a high level of social distancing with a ban in gatherings of more than five persons [26], in their efforts to lower the triple-digit number of daily confirmed cases that has been observed since the start of this year. According to Heo et al’s study [27] on the COVID-19 situation of Korea, the Korean government social distancing policy was predicted to lead to a decrease in the daily confirmed cases observed in the country but with only segmented Poisson model which according RMSE value, did not perform well as compared to the other models. In future, we hope to control for the influence of government interventions using the other models, to give a whole picture of the future COVID-19 situation of the country.

A good understanding of the epidemic dynamic would greatly enhance the control and prevention of COVID-19 as well as other infectious diseases. Therefore, taking precaution when using prediction to support a decision, for example, return to work or lowering of the social distancing level, is highly encouraged too.

ORCID

Taewan Goo: https://orcid.org/0000-0001-9427-2290
Catherine Apio: https://orcid.org/0000-0003-0240-794X
Gyujin Heo: https://orcid.org/0000-0003-3499-2848
Doeun Lee: https://orcid.org/0000-0002-2791-8431
Jong Hyeok Lee: https://orcid.org/0000-0002-3573-1853
Jisun Lim: https://orcid.org/0000-0002-1058-3292
Kyulhee Han: https://orcid.org/0000-0002-7243-5711
Taesung Park: https://orcid.org/0000-0002-8294-590X
Authors’ Contribution

Conceptualization: TP. Data curation: KH, GH. Formal analysis: TG, GH, DL, JHL, JL. Funding acquisition: TP. Methodology: TP. Writing - original draft: TG, CA. Writing - review & editing: TP, TG, CA.

Conflicts of Interest

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

Supplementary Materials

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

References


Instructions for authors

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

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

Research and publication ethics

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

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

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

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

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

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

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

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

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

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

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

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

Preparation of manuscripts

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

Manuscript type

Original articles
Original research articles are full scientific reports of original research. The manuscript should be organized as follows: Title Page, Abstract & Keywords, Introduction, Methods, Results, Discussion, Acknowledgments, References, Tables, and Figure Legends. The Results and Discussion can be combined.
**Application notes**

Application notes are short communications about novel software, new algorithm implementations, databases, and network services (web servers and interfaces). The manuscripts include the following: Title Page, Abstract & Keywords, Availability, Introduction, Main Text, References, and Supplementary Information.

**Clinical genomics**

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

**Genome archives**

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

**Letters to the editor**

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

**Review articles**

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

**Opinions / Commentaries**

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

**Minireviews**

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

**Research communications**

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

**Manuscript Format**

**Title**

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

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

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

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

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

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

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

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

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

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

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

Examples of references are given below:

Journal article

Books

Book sections

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

Conference paper

Dissertation/Thesis

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

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

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

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

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

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

Supplementary materials
Supplementary materials can be provided to support and enhance scientific information. Supplementary files offer additional possibilities for publishing supporting applications, sequence alignment, background datasets, microarray hybridization experiments, high-resolution images, movies, sound clips, and more. Supplementary files will be published alongside the online version of the article on the Genomics & Informatics web site. This material will not be edited or formatted; thus, the authors are responsible for the accuracy and presentation of all such material. Accepted file formats for supplementary materials:
- Quick Time files (.mov)
Selection and Description of Participants

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

Submission of Manuscript

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

Peer review and revision of manuscripts

Peer review

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

Manuscript revision

When a manuscript is returned to the corresponding author for revision, the reviewed manuscript must be re-submitted within one month, unless the authors request an extension. A galley proof
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, authors should contact the editorial office.

**Clinical data sharing policy**
This journal follows the data sharing policy described in "Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors” (https://doi.org/10.3346/jkms.2017.32.7.1051). As of July 1, 2018, manuscripts submitted to ICMJE journals that report the results of clinical trials must contain a data sharing statement as described below. Clinical trials that begin enrolling participants on or after January 1, 2019 must include a data sharing plan in the trial’s registration. The ICMJE’s policy regarding trial registration is explained at www.icmje.org/recommendations/browse/publishing-and-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 (&quot;learned intermediary&quot;) identified for this purpose.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>For what types of analyses?</td>
<td>Any purpose</td>
<td>To achieve aims in the approved proposal.</td>
<td>For individual participant data meta-analysis.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>By what mechanism will data be made available?</td>
<td>Data are available indefinitely at (link to be included).</td>
<td>Proposals should be directed to xxx@yyy. To gain access, data requestors will need to sign a data access agreement.</td>
<td>Proposals may be submitted up to 36 months following article publication. After 36 months, the data will be available in our University's data warehouse but without investigator support other than deposited metadata.</td>
<td>Not applicable</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Information regarding submitting proposals and accessing data may be found at (link to be provided).</td>
<td></td>
</tr>
</tbody>
</table>

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

ICMJE, International Committee of Medical Journal Editors.

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

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

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

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

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

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

**Contact address**

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

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

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

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

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

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

<table>
<thead>
<tr>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>
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, facsimle and E-mail address of corresponding author, (6) running title (no more than 50 characters).

☐ 3. Abstract in unstructured format within 250 words.

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

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

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

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

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

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

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