Aims and scope

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

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

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

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

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

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

© Copyright 2020, the Korea Genome Organization
It is an open access journal. The articles are distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
EDITOR IN CHIEF

Park, Taesung Seoul National University, Korea

ADVISORY EDITORIAL BOARD

Batzer, Mark A. Louisiana State University, U.S.A.
Church, George M. Harvard University, U.S.A.
Lee, Byungkook National Institute of Health, U.S.A.
Matsuda, Fumihiko Kyoto University, Japan
Sakaki, Yoshiyuki RIKEN Genomic Science Center, Japan
Seo, Jeong-Sun Seoul National University, Korea

ASSOCIATE EDITORS

Cho, Soo Young National Cancer Center, Korea
Choi, Murim Seoul National University, Korea
Han, Kyudong Dankook University, Korea
Huh, Sun Hallym University, Korea
Kim, Sangsoo Soongsil University, Korea
Oh, Bermseok Kyung Hee University, Korea
Oh, S. June Inje University, Korea
Park, Hyun Seok Ewha Womans University, Korea
Yoon, Kyong-Ah Konkuk University, Korea
Won, Sungho Seoul National University, Korea
Woo, Hyun Goo Ajou University, Korea

EDITORIAL BOARD

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

ETHICS EDITOR

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

STATISTICS EDITOR

Han, Buhm Seoul National University, Korea

MANUSCRIPT EDITOR

Chang, Soo Hee Infolumi, Korea

LAYOUT EDITOR

Jeong, Eun Mi M2community, Korea

WEBSITE AND JATS XML FILE PRODUCER

Bae, Hyo-Jeong M2community, Korea
Editorial

Editor's introduction to this issue (G&I 18:3, 2020)
Taesung Park

Original articles

A semi-automatic cell type annotation method for single-cell RNA sequencing dataset
Wan Kim, Sung Min Yoon, Sangsoo Kim

Impact of type 2 diabetes variants identified through genome-wide association studies in early-onset type 2 diabetes from South Indian population
Samuel Liju, Manickam Chidambaram, Viswanathan Mohan, Venkatesan Radha

Molecular characterization and functional annotation of a hypothetical protein (SCO0618) of Streptomyces coelicolor A3(2)

The comparative gene expression concern to the seed pigmentation in maize (Zea mays L)
Kyu Jin Sa, Ik-Young Choi, Ju Kyong Lee

Computational analysis of SARS-CoV-2, SARS-CoV, and MERS-CoV genome using MEGA
Vipan Kumar Sohpal

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

Genome archive

Sequencing and annotation of the complete mitochondrial genome of a threatened labeonine fish, Cirrhinus reba
Mohammad Nazrul Islam, Shirin Sultana, Md. Jobaidul Alam

Application Note

Organizing an in-class hackathon to correct PDF-to-text conversion errors of Genomics & Informatics 1.0
Sunho Kim, Royoung Kim, Hee-Jo Nam, Ryeo-Gyeong Kim, Enjin Ko, Han-Su Kim, Jihye Shin, Daeun Cho, Yurhee Jin, Soyeon Bae, Ye Won Jo, San Ah Jeong, Yena Kim, Seoyeon Ahn, Bomi Jang, Jiheyon Seong, Yujin Lee, Si Eun Seo, Yujin Kim, Ha-Jeong Kim, Hyeji Kim, Hye-Lynn Sung, Hyoyoung Lho, Jaywon Koo, Jion Chu, Juwon Lim, Youngju Kim, Kyungyeon Lee, Yuri Lim, Meongeun Kim, Seonjeong Hwang, Shinhye Han, Soheyun Bae, Sua Kim, Suhyeon Yoo, Yeonjeong Seo, Yerim Shin, Yonsoo Kim, You-Jung Ko, Jihee Baek, Hyejin Hyun, Hyemin Choi, Ji-Hye Oh, Da-Young Kim, Hyun-Seok Park
Editor’s introduction to this issue (G&I 18:3, 2020)

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

In this issue, there are eight articles: six Original Articles, one Application Note, and one article in the category of Genome Archives. The first original article, by the group of Kim et al. (Soongsil University, Seoul, Korea) proposed a novel method for the analysis of single-cell RNA sequencing. Specifically, their article described a semi-automatic method that calculates a normalized score for each cell type based on a user-supplied cell-type-specific marker gene list. Second, Liju et al. (Madras Diabetes Research Foundation, Chennai, India) identified associations of genetic variants with early-onset of type 2 diabetes in a South Indian population. Although the sample size was not large, the association of the HHEX variant rs1111875 was successfully demonstrated for the first time in a South Indian population.

The third article, by Ferdous et al. (Mawlana Bhashani Science and Technology University, Tangail, Bangladesh), provided a molecular characterization and functional annotation of a hypothesized protein of Streptomyces coelicolor A3(2), which is a Gram-positive soil bacterium known for the production of several antibiotics used in various biotechnological applications. The application of several bioinformatics tools successfully revealed the characteristics of this hypothesized protein from the genome of S. coelicolor, including its structure, function, and homologous proteins. The fourth article, by the group of Sa et al. (Kangwon National University, Chuncheon, Korea), presented a comparative gene expression analysis regarding seed pigmentation in maize by comparing differently expressed genes from three inbred lines, including a pigment-accumulating seed type (CM22) and non-pigmented seeds (CMS5 and CM19).

Recent active research efforts have produced many articles on the coronavirus disease 2019 (COVID-19) pandemic. In this issue, there are two articles related to COVID-19. The first one, by Sohpal (Beant College of Engineering & Technology, Gurdaspur, India) presents a computational analysis of the genome of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, the virus that causes COVID-19) with the genome of the Middle East respiratory syndrome–related coronavirus. Using molecular evolutionary genetic analysis (MEGA) from the National Center for Biotechnology Information (NCBI) for statistical analysis, the best substitution pattern and transition/transversions (R) were compared. The final Research Article was by Apio et al. (Seoul National University, Seoul, Korea), which presents the 95% confidence intervals for the SARS-CoV-2 antibody retention rate for the Korean population using two recently performed antibody tests in Korea. The most conservative 95% confidence interval estimation showed that as of 00:00 on September 15, 2020, there were at least 32,602 undetected cases of COVID-19 in Korea.

The one article in the Genome Archives categories, by Islam et al. (Sher-e-Bangla Agricultural University, Dhaka, Bangladesh) presented the sequencing and annotation of the complete mitochondrial genome (16,597 bp in size) of a threatened labeonine fish, Cirrhinus reba, collected from the Khulna region of Bangladesh.

The Application Note, by Park and his students (Ewha Womans University, Seoul, Ko-
rea) described the initiation of the first *Genomics & Informatics* Annotation Hackathon (GIAH) event, focusing on improving earlier versions of the full-text corpus of *Genomics & Informatics* by semi-automatically detecting and correcting PDF-to-text conversion errors and optical character recognition errors.

**ORCID**

Taesung Park: https://orcid.org/0000-0002-8294-590X
A semi-automatic cell type annotation method for single-cell RNA sequencing dataset

Wan Kim, Sung Min Yoon, Sangsoo Kim*
Department of Bioinformatics and Life Science, Soongsil University, Seoul 06978, Korea

Single-cell RNA sequencing (scRNA-seq) has been widely applied to provide insights into the cell-by-cell expression difference in a given bulk sample. Accordingly, numerous analysis methods have been developed. As it involves simultaneous analyses of many cell and genes, efficiency of the methods is crucial. The conventional cell type annotation method is laborious and subjective. Here we propose a semi-automatic method that calculates a normalized score for each cell type based on user-supplied cell type–specific marker gene list. The method was applied to a publicly available scRNA-seq data of mouse cardiac non-myocyte cell pool. Annotating the 35 t-stochastic neighbor embedding clusters into 12 cell types was straightforward, and its accuracy was evaluated by constructing co-expression network for each cell type. Gene Ontology analysis was congruent with the annotated cell type and the corollary regulatory network analysis showed upstream transcription factors that have well supported literature evidences. The source code is available as an R script upon request.

Keywords: cell type annotation, co-expression network, regulatory network, single-cell RNA sequencing, transcription factor

Introduction

Next-generation sequencing technology has transformed transcriptomics by allowing simultaneous identification and quantification of the expressed RNA molecules. However, this RNA-sequencing applied to bulk samples provides averaged expression profiles, not single-cell level variations. On the other hand, single-cell RNA-seq (scRNA-seq) isolates single cells from a given bulk sample, and measures the expression profile of a number of RNA species from each cell, offering the potential of cell-type characteristics as well as cell-type profiles of the bulk sample [1-3]. scRNA-seq has been applied to various fields such as neurobiology and cancer biology [1]. For example, immunotherapy has been developed as an effective cancer therapy, and thus underpinning of cancer immunology is crucial in the development of immunotherapy; scRNA-seq is a useful tool in this area [4-6].

scRNA-seq involves simultaneous analyses of many cell and genes, efficiency is crucial. One of the time-consuming steps is cell type annotation. The conventional method is laborious and subjective. To address this problem, some annotation methods have been developed by using additional transcriptomic data as reference coupled with machine-learning technique [7,8]. For instance, SingleR, a cell type annotation tool for scRNA-seq, leverages reference transcriptomic datasets of pure cell types to infer the cell of origin of each of the single cells independently [7]. But these methods require additional data and consume much computing power needed to perform machine-learning. Here we propose a semi-automatic method that calculates a normalized score for each
cell type based on user-supplied cell type-specific marker gene list. The user can easily decide a cut-off score for each cell type based on the plots of score distributions.

**Methods**

**Materials**

We downloaded raw scRNA-seq datasets from the publicly available ArrayExpress DB (E-MTAB-6173) and Gene Expression Omnibus (GSE92332). A 10× Genomics [3] technology was used to generate these two datasets. One dataset (E-MTAB-6173) was generated from male and female mouse cardiac cell pools after depleting endothelial cells to 10% [9]. The other dataset (GSE92332) was measured from small intestinal epithelial cells from female and male mice that were randomly assigned to treatment groups after matching for the sex and age of 7–10 weeks [10].

**Overall analysis workflow**

Our scRNA-seq analysis pipeline is based on a well-established practice of processing 10× Genomics data. The sequencing data were processed with Cellranger to obtain an expression matrix of RNAs for each cell. Subsequent processing was performed with Seurat for various quality control steps involving cell filtering, normalization, and removal of technical variation, followed by preliminary analyses such as dimension reduction, clustering, cell type annotation, and differential expression analysis [11]. Our Cell Type Activity (CTA) annotation method is an alternative to the cell type annotation step in Seurat. For each cell type, the expression matrix of gene-by-cell was used in co-expression network analysis using WGCNA [12]. The modules in the network were analyzed for the enrichment of Gene Ontology terms and identification of upstream transcription factors (TFs) using iRegulon (Fig. 1) [13].

**Alignment and pre-filtering**

STAR was used to map FASTQ reads to mm10-3.0.0 mouse reference genome [14]. Cellranger detects the cases where two cells are captured by a 10× Genomics GEM bead and filters the RNA counts originated from the dead cells. Using default options of Cellranger 3.0.2, the feature-barcode matrix was generated.

**Dimension reduction, clustering, and annotation**

Seurat 3.1.0 was used for principal component analysis, t-stochastic neighbor embedding (t-SNE), clustering, and cell-type annotation of the feature-barcode matrix that had been generated by Cellranger. With the clustered result, each cluster is annotated with...

---

**Fig. 1.** Analysis workflow diagram. This diagram illustrates overall workflow used in our study. FASTQ files were processed for alignment, cell filtering, UMI count and feature (genes) count by using CellRanger. CellRanger is a popular pipeline that processes Chromium single-cell 3′ RNA-sequencing data. Next, Seurat was used to generate clusters of cells. Seurat is an R package offering functions for secondary analysis such as cell QC, dimension reduction, clustering and differential expressed gene analysis. After clustering, we used our Cell Type Activity estimation technique for cell type annotation. To identify modules from each cell type, WGCNA was used. WGCNA is an R package that calculates correlation weighted network by using gene expression data and creates clusters of genes. Enrichment analysis study was conducted on each modules and iRegulon was used to identify potential transcription factors co-regulating the gene sets.
a cell type by identifying the dominant expression of a list of cell type–specific markers. The conventional method requires manual examination of a violin plot for each marker of a cell type, followed by confirmation of its high expression in the feature plot. This is a laborious step, often producing imprecise results especially for the cases having a high number of clusters or many similar cell types. In order to reduce human errors, we automated the cell type annotation step using the algorithm that has been well developed for gene-set analysis [15]. Our method is called Cell Type Activity (CTA) method.

The inputs to CTA comprise a feature-barcode matrix, the cluster membership of each cell, and a list of markers for each cell type of the user’s choice. Suppose the following: there are \( K \) clusters from the Seurat result and these clusters would be annotated with \( C \) different cell types. If each cell type has \( N \) marker genes, there would be a total of \( C \times N \) markers. The following steps are repeated \( C \) times, one for each cell type. For a given cell type, the first step tabulates the cluster median expression of each marker, generating an \( N \times K \) matrix. Each row of the matrix \( (P) \) is then normalized so that it sums to unity \((\sum_{i} p_{ik} = 1, \ i = 1, \cdots, N)\). In the second step, we calculate the weight to emphasize informativeness of a marker gene in classifying the clusters. To calculate this, we use the concept of Gini impurity that measures how homogeneous the groupings are. Because the original Gini impurity reaches its minimum (zero) when only one class is classified, we modified it to give the highest score if a marker gene is expressed specifically in one cluster. The weight for the \( i^{th} \) gene is defined as below:

\[
W_i = 1 + \sum_{k} p_{ik}^2,
\]

where \( p_{ik} \) represents one of the elements of the aforementioned \( P \) matrix. In the third step, a CTA score for the \( k^{th} \) cluster, \( S_k \), is calculated as follows:

\[
S_k = \sum_{i} \frac{E_{ik} \cdot W_i}{\sqrt{N}},
\]

where \( E_{ik} \) is the average expression value of the \( i^{th} \) marker gene in the \( k^{th} \) cluster and \( W_i \) is the weight calculated above (Eq. 1). Lastly, we convert the CTA scores to probabilities by normalizing them to a sum of unity. The user decides a cut-off score based on a cumulative normal distribution curve of the normalized CTA scores. The whole process is repeated for each cell type.

**Co-expression and co-regulatory networks**

WGCNA 1.68 was used to construct cell type–specific co-expression networks. The Seurat output was parsed to be used as an input to WGCNA. The resulting modules were further analyzed for the enrichment of Gene Ontology terms.

For the co-expressed modules, the potential upstream TFs were inferred using iRegulon 1.3 available as an application of Cytoscape. For the TF binding motif search, the 20 kb upstream region of transcription start site of each gene was used.

**Results**

The FASTQ files of mouse cardiac cell pools (ArrayExpress E-MTAB-6173) were processed with Cellranger, resulting in the feature-barcode matrix of 11,701 cells. Seurat was used to analyze the single-cell level heterogeneity. The cells having mitochondrial RNAs more than 25% of the total expressed RNAs were removed. In order to remove cells having unrealistic RNA varieties, only the cells with unique feature counts within the range between 200 and 5,000 were kept, resulting in a total of 11,587 cells and 17,432 genes. For the t-SNE clustering, 24 dimensional components as inferred from the principal component analysis and the resolution value of 2.0 were used, resulting in a total of 35 clusters. In order to annotate an appropriate cell type to each cluster we employed the CTA method (see Methods) using 12 cell types having 10 unique marker genes per cell type (Supplementary Table 1) [9]. The CTA score distribution for each cell type was manually examined to adjust the cut-off (see Supplementary Fig. 1A–F for exemplary plots). The t-SNE clustering with the cell type annotation is shown in Fig. 2, while the number of cells and genes identified for each cell type are listed in Table 1. The CTA matrix (12 rows by 35 columns) is given in Supplementary Table 2.

The performance of our CTA annotation method was evaluated with the other scRNA-seq dataset of mouse small intestinal epithelium [10]. The same workflow was applied for cell QC, resulting in a total of 5,188 cells and 14,259 genes. For clustering and visualization, 12 principal components and the resolution value of 0.8 were used, resulting in a total of 15 clusters. We performed the CTA method by using the 11 cell-type lists (Supplementary Table 3). Cumulative distribution of the CTA score was visualized for decision of the cut-off (Supplementary Fig. 2A–K). The final annotation graph visualized by t-SNE is shown in Supplementary Fig. 3, and the CTA matrix is given in Supplementary Table 4. The annotation results for the major clusters were qualitatively congruent with the original work [10].

In order to assess the quality of our CTA annotation method, we performed co-expression and co-regulatory network analyses with the clustering and annotation results of the mouse cardiac cell pools [9]. While the member cells in each cell-type cluster display
was constructed with WGCNA. The network was modularized using the soft threshold value (Supplementary Table 5) that produced a scale-free network for each cell-type cluster. Each co-expression module was further analyzed for the enrichment of Gene Ontology terms (Table 2).

As shown in Table 3, an exemplary module from the dendritic cell (DC) cluster was significantly enriched with immune-related terms that are related to the features of dendritic cells such as immune cell activation, recognition of pathogen (the full listing in Table 3).

Fig. 2. t-distributed stochastic neighbor embedding (t-SNE) result of cardiac non-myocytes. This plot displays final cell type annotation based on Cell Type Activity (CTA) scores. Feature-barcode matrix from single-cell RNA sequencing data was used for dimension reduction by using principal component analysis. After dimension reduction, t-SNE was used to visualize the clusters. The cell type for each cluster was inferred from the CTA score.

Table 1. Annotation summary table

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Cell</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibroblast 1</td>
<td>6,386</td>
<td>16,840</td>
</tr>
<tr>
<td>Macrophage</td>
<td>1,427</td>
<td>15,773</td>
</tr>
<tr>
<td>Endothelial cell</td>
<td>920</td>
<td>14,594</td>
</tr>
<tr>
<td>Smooth muscle cell</td>
<td>712</td>
<td>14,616</td>
</tr>
<tr>
<td>Fibroblast 2</td>
<td>661</td>
<td>15,324</td>
</tr>
<tr>
<td>Pericyte</td>
<td>425</td>
<td>12,968</td>
</tr>
<tr>
<td>B cell</td>
<td>331</td>
<td>12,646</td>
</tr>
<tr>
<td>Dendritic cell</td>
<td>208</td>
<td>13,007</td>
</tr>
<tr>
<td>Schwann cell</td>
<td>179</td>
<td>12,416</td>
</tr>
<tr>
<td>Granulocyte</td>
<td>134</td>
<td>9,392</td>
</tr>
<tr>
<td>T cell</td>
<td>121</td>
<td>11,282</td>
</tr>
<tr>
<td>Natural killer cell</td>
<td>56</td>
<td>10,187</td>
</tr>
</tbody>
</table>

This chart displays summary of cell types after annotation based on Cell Type Activity score. The amount of cells of each cell type is written at cell column. The number of genes that were expressed in the cell type is showed at gene column.

Table 2. Co-expression module summary table

<table>
<thead>
<tr>
<th>Cell type</th>
<th>All modules</th>
<th>GO assigned modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dendritic cell</td>
<td>98</td>
<td>13</td>
</tr>
<tr>
<td>T cell</td>
<td>86</td>
<td>5</td>
</tr>
<tr>
<td>Schwann cell</td>
<td>81</td>
<td>7</td>
</tr>
<tr>
<td>Pericyte</td>
<td>81</td>
<td>2</td>
</tr>
<tr>
<td>Granulocyte</td>
<td>76</td>
<td>6</td>
</tr>
<tr>
<td>Endothelial cell</td>
<td>67</td>
<td>13</td>
</tr>
<tr>
<td>Smooth muscle cell</td>
<td>57</td>
<td>11</td>
</tr>
<tr>
<td>B cell</td>
<td>50</td>
<td>7</td>
</tr>
<tr>
<td>Natural killer cell</td>
<td>46</td>
<td>7</td>
</tr>
<tr>
<td>Macrophage</td>
<td>37</td>
<td>14</td>
</tr>
<tr>
<td>Fibroblast 2</td>
<td>23</td>
<td>5</td>
</tr>
<tr>
<td>Fibroblast 1</td>
<td>12</td>
<td>8</td>
</tr>
</tbody>
</table>

Each column represents the number of modules generated by co-expression network and the number of modules that were enriched Gene Ontology (GO) term, respectively.
This module comprises of 242 genes and the iRegulon motif analysis of their 20 kb upstream of transcription start site inferred 11 potential TFs (Table 4). Among the inferred TFs, STAT1 and CEBPB have more than 100 target genes each. Their regulatory relationship is depicted in Fig. 3, showing that CEBPB regulates STAT1, which is also self-regulated. Having established this, their roles in DC biology was surveyed from the literature. It is known that up-regulation of STAT6 pathway plays an important role in the differentiation of immature DCs, and its down-regulation is related to the maturation of DCs. It is reported that STAT1 pathway works opposite to STAT6 pathway, maturing mDCs [16]. IRF8, an epigenetic and fate-determining TF of plasmacytoid dendritic cell (pDC), modulates chromatin modification of thousands of pDC enhancers. CEBPB forms a negative feedback loop with IRF8, determining the epigenetic fate of monocyte-derived DCs [17].

**Discussion**

Here we propose an efficient semi-automatic processing pipeline of scRNA-seq data, called CTA method. Its quality was assessed by constructing co-expression and co-regulatory networks using the cell type annotation results. Our results were qualitatively congruent with the literature information. In scRNA-seq, many really expressed RNA species are missed. If this drop-out event can be complemented by imputation, much richer information can be retrieved. However, the current implementation of the imputation such as BISCUIT is very slow and not attempted in this work [18].

The strategy demonstrated in this work may find useful applications in inferring regulators of various cell types. For example, for the cell types whose critical differentiation regulators are elusive, co-regulatory network construction of progenitor and differentiated cells may elucidate key modules for the differentiation.

### Table 3. Dendritic cell exemplary module GO term result

<table>
<thead>
<tr>
<th>Enrichment P</th>
<th>Term ID</th>
<th>Term name</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.94E-24</td>
<td>GO:0035456</td>
<td>Response to interferon-beta</td>
</tr>
<tr>
<td>2.15E-20</td>
<td>GO:0006952</td>
<td>Defense response</td>
</tr>
<tr>
<td>3.99E-20</td>
<td>GO:0045087</td>
<td>Innate immune response</td>
</tr>
<tr>
<td>1.61E-19</td>
<td>GO:0035458</td>
<td>Cellular response to interferon-beta</td>
</tr>
<tr>
<td>4.47E-19</td>
<td>GO:0043207</td>
<td>Response to external biotic stimulus</td>
</tr>
<tr>
<td>4.47E-19</td>
<td>GO:0051707</td>
<td>Response to other organism</td>
</tr>
<tr>
<td>1.96E-15</td>
<td>GO:0098542</td>
<td>Defense response to other organism</td>
</tr>
</tbody>
</table>

Enrichment analysis revealed the module that is highly associated with immune Gene Ontology (GO) term. This table represents the list of GO terms from tan module.

### Table 4. Dendritic cell inferred TF summary table

<table>
<thead>
<tr>
<th>Inferred TF</th>
<th>Target gene count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stat1</td>
<td>156</td>
</tr>
<tr>
<td>Cebpb</td>
<td>109</td>
</tr>
<tr>
<td>Ar</td>
<td>19</td>
</tr>
<tr>
<td>Sox9</td>
<td>32</td>
</tr>
<tr>
<td>Atf2</td>
<td>21</td>
</tr>
<tr>
<td>Yy1</td>
<td>14</td>
</tr>
<tr>
<td>Cebpe</td>
<td>8</td>
</tr>
<tr>
<td>Mybl2</td>
<td>10</td>
</tr>
<tr>
<td>Snai2</td>
<td>9</td>
</tr>
<tr>
<td>Tbx15</td>
<td>27</td>
</tr>
<tr>
<td>Bdp1</td>
<td>12</td>
</tr>
</tbody>
</table>

The gene set from the exemplary module of dendritic cell was used for regulatory network analysis, which enables to infer co regulation transcription factor (TF). This chart shows revealed TF names and related gene from the module.

**Fig. 3.** Dendritic cell exemplary module regulatory network. Co-expression network was conducted to find modules of dendritic cell. Gene list in the module that highly enriched immune functions was analyzed to identify co-regulating transcription factors. Network visualization illustrates interactions between transcription factors (TFs; green) and genes (pink) from the module. The analysis revealed that two TF (CEBPB and STAT1) are significantly related to the genes. There is also relationship between two TFs and self-regulatory loop for the downstream STAT1.

**ORCID**

Wan Kim: https://orcid.org/0000-0001-9934-0360
Sung Min Yoon: https://orcid.org/0000-0001-6383-8546
Sangsoo Kim: https://orcid.org/0000-0001-9836-9823

https://doi.org/10.5808/GI.2020.18.3.e26
Authors’ Contribution

Conceptualization: SMY, WK, SK. Data curation: SMY, WK. Formal analysis: SMY, WK. Funding acquisition: SK. Methodology: SMY, WK. Writing – original draft: SMY, WK, SK. Writing – review & editing: WK, SMY, SK.

Conflicts of Interest

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

Acknowledgments

This work is based on the Master’s degree thesis of SMY at Soongsil University, Seoul, Korea. We acknowledge the financial support from the Soongsil University Research Fund. The computational resources were kindly provided by Korea Institute of Science and Technology Information (GSDC & KREONET).

Supplementary Materials

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

References

Impact of type 2 diabetes variants identified through genome-wide association studies in early-onset type 2 diabetes from South Indian population

Samuel Liju¹, Manickam Chidambaram¹,⁴, Viswanathan Mohan¹,², Venkatesan Radha¹*

¹Department of Molecular Genetics, Madras Diabetes Research Foundation, Chennai 600086, India
²Dr. Mohan’s Diabetes Specialties Centre, ICMR Centre for Diabetes Advanced Research and WHO Collaborating Centre for Non-communicable Diseases Prevention and Control, Chennai 600086, India

The prevalence of early-onset type 2 diabetes (EOT2D) is increasing in Asian countries. Genome-wide association studies performed in European and various other populations have identified associations of numerous variants with type 2 diabetes in adults. However, the genetic component of EOT2D which is still unexplored could have similarities with late-onset type 2 diabetes. Here in the present study we aim to identify the association of variants with EOT2D in South Indian population. Twenty-five variants from 18 gene loci were genotyped in 1,188 EOT2D and 1,183 normal glucose tolerant subjects using the MassARRAY technology. We confirm the association of the HHEX variant rs1111875 with EOT2D in this South Indian population and also the association of CDKN2A/2B (rs7020996) and TCF7L2 (rs4506565) with EOT2D. Logistic regression analyses of the TCF7L2 variant rs4506565(A/T), showed that the heterozygous and homozygous carriers for allele ‘T’ have odds ratios of 1.47 (95% confidence interval [CI], 1.17 to 1.83; p = 0.001) and 1.65 (95% CI, 1.18 to 2.28; p = 0.006) respectively, relative to AA homozygote. For the HHEX variant rs1111875 (T/C), heterozygous and homozygous carriers for allele ‘C’ have odds ratios of 1.13 (95% CI, 0.91 to 1.42; p = 0.27) and 1.58 (95% CI, 1.17 to 2.12; p = 0.003) respectively, relative to TT homozygote. For the CDKN2A/2B variant rs7020996, the heterozygous and homozygous carriers of allele ‘C’ were protective with odds ratios of 0.65 (95% CI, 0.51 to 0.83; p = 0.0004) and 0.62 (95% CI, 0.27 to 1.39; p = 0.24) respectively, relative to TT homozygote. This is the first study to report on the association of HHEX variant rs1111875 with EOT2D in this population.

Keywords: early-onset type 2 diabetes, genome-wide association studies, HHEX gene, MassARRAY genotyping, South Asians, TCF7L2 gene

Introduction

Early-onset type 2 diabetes (EOT2D) a relatively new phenomenon recognized in the past few decades and is caused by the complex interplay between genetic and environmental factors [1,2]. Currently 425 million people are living with diabetes worldwide and the number is expected to reach 629 million by 2045 with nearly 60% of the affected peo-
ple living in Asian countries [3,4]. Asians have an earlier age of diagnosis and a higher prevalence of diabetes for the same body mass index (BMI) than Europeans. India alone is presently home to 72 million people with diabetes [4,5]. Epidemiological studies performed in Indians showed a 25.3% increase in individuals developing type 2 diabetes (T2D) at < 40 years [6]. A recent nationwide population-based study estimating the national prevalence of diabetes and prediabetes in India also indicates 25–34 years as the take-off point for diabetes both in urban and rural areas [7].

Genome-wide association studies (GWAS) and subsequent meta-analyses of these studies have increased the list of T2D associated genetic variants to more than a hundred [8]. However, these variants identified by the large scale GWAS were mostly with the late-onset type 2 diabetes (LOT2D) subtype that develops after 40 years of age. Though the T2D subtype that develops at earlier ages (EOT2D) has a considerably larger heritable component, very few studies have looked at the genetic component of EOT2D. T2D develops at an earlier age a decade or two earlier in Asian Indians and often coincides with the monogenic form of diabetes namely maturity-onset diabetes of the young (MODY) [9,10]. Indeed previous studies have demonstrated association of some MODY variants also with EOT2D [11-13]. Genetic variants in TCF7L2 [14-16], HNF1A [17], ABCA1 [18], Dio2 [19], PCLO [20], Trib3 [21], AdipoQ, and Lepr [22] identified with LOT2D in various populations also showed association with EOT2D. Our group also replicated the association of the variants in TCF7L2, CDKN2A/2B and an intergenic single nucleotide polymorphism (SNP) on chromosome 1p31 identified with LOT2D in various population also with EOT2D in Asian Indians [23]. With this background, the present study was designed to study the association of 25 variants within 18 distinct gene loci, previously identified with the LOT2D subtype in various GWAS, on South Indians with EOT2D.

Methods

Study subjects
The study group comprised of 1,188 unrelated EOT2D subjects and 1,183 normal glucose tolerant (NGT) subjects recruited from Chennai Urban Rural Epidemiology Study (CURES) and from Dr. Mohan’s Diabetes Specialties Centre (DMDSC) tertiary diabetes center in Chennai in South India. Subjects for the study were selected based on the World Health Organization (WHO) criteria. NGT was defined as fasting plasma glucose < 100 mg/dL and 2-h post glucose value ≤ 140 mg/dL. Diabetes was diagnosed if the fasting plasma glucose was ≥ 126 mg/dL or 2-h post glucose value ≥ 200 mg/dL or if the participant was on drug therapy for diabetes after diagnosis by a physician. The following criteria were used for selection of EOT2D subjects: patients having early-onset diabetes if they were diagnosed before the age of 35 years, responding to oral hypoglycemic agents, fasting C-peptide > 1.0, stimulated C-peptide > 2.0 pmol/mL, and glutamic acid decarboxylase antibodies negative. Only unrelated individuals were included in this study. Subjects with ketoacidosis at diagnosis, exocrine pancreatic disease (fibrocalscular pancreatic diabetes), pregnant women and subjects known to have confirmed maturity-onset diabetes of the young, were excluded from the study. Written consent was obtained from all the individuals participating in the study and the study was approved by the Institutional Ethics Committee of the Madras Diabetes Research Foundation (RHN/Adhoc/19/2011-2012).

Anthropometric and biochemical measurements

Anthropometric measurements including weight, height, and waist measurements were obtained using standardized techniques. The BMI was calculated using the formula, weight (kg)/(height × height)(m²). Blood pressure (BP) was measured with a mercury sphygmomanometer (Diamond Deluxe BP apparatus, Pune, India) from the left arm in a sitting position. Fasting plasma glucose (glucose oxidase-peroxidase method), serum cholesterol (cholesterol oxidase-peroxidase-amidopyrine method), serum triglycerides (glycerol phosphate oxidase-peroxidase-amidopyrine method), and high-density lipoprotein cholesterol (direct method polyethylene glycol-pretreated enzymes) was measured using Hitachi-912 Auto analyzer (Hitachi, Mannheim, Germany). Low-density lipoprotein cholesterol was calculated using the Friedewald formula. Glycated hemoglobin was estimated by high-pressure liquid chromatography using the variant machine (Bio-Rad, Hercules, CA, USA) and the intra- and inter-assay coefficient of variation of glycated hemoglobin was less than 10%.

SNP selection

Twenty-five variants representing eighteen different gene loci identified in various GWAS studies including ADAMTS9 [24], CDC123 [25], CDKAL1 [26-28], CDKN2A/2B [24,29,30], COBLL1 [31], GRB14 [32], HNF1A [33], HNF4A [34], IGF2BP2 [26,28], JAZF1 [24,33], HHEX [35], PPARG [35], RBMS1 [36], SLC30A8 [27], TCF7L2 [37], THADA [24], TSPAN8 [34] and TSPAN8 [34] were selected for the study.

Genotyping

Genomic DNA was extracted from peripheral blood leukocytes by proteinase K digestion followed by phenol-chloroform method. Genotyping was done using MassARRAY system (Sequenom, San Diego, CA, USA) following the manufacturer’s instructions as
published elsewhere [38]. SpectroTYPEr software (Sequenom) automatically called the genotypes and only conservative and moderate calls were accepted for the study. Ten percent of the samples genotyped were replicated and discordance rate observed was less than 0.4% for the replicated samples. All the variants genotyped had call rate ranging between 90%–99%.

**Statistical analysis**

Hardy-Weinberg equilibrium (HWE) was performed by using Pearson χ² statistics in controls for each variant separately. Logistic regression analysis was performed assuming additive model to determine the association between variants and the risk for EOT2D, with and without adjusting for parametric confounders such as age, sex, and BMI using SPSS version 20.0 (IBM Corp., Armonk, NY, USA). The power of the study was estimated using PS Power and Sample Size program (Vanderbilt University, Nashville, TN, USA) calculations (with type I error probability α = 0.05). Linkage disequilibrium (LD) and haplotype frequencies were estimated using Haploview software (http://www.broad.mit.edu/mpg/haploview/) [39].

**Ethical approval**

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional ethics committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Results**

**Clinical and biochemical parameters of the study subjects**

Table 1 summarizes the clinical and biochemical parameters of the subjects studied. Mean age of the EOT2D and NGT subjects were 32 ± 6 and 31 ± 8 (mean ± SD), respectively. The fasting plasma glucose, 2-h post plasma glucose, and glycated hemoglobin were significantly (p < 0.001) higher among the EOT2D subjects when compared with the NGT subjects.

**Comparison of minor allele frequencies of the studied polymorphisms in South Indian population with frequencies from the 1000 Genomes Project populations**

Minor allele frequencies (MAF) of the SNPs studied in the present study were compared with the reported frequencies of 1000 Genomes Project (Global, European, and South Asian population), representative of the genetic diversity that exists within various population in the world and is shown in Supplementary Table 1. According to the 1000 Genome Project database, MAF of most of the studied SNPs in the present study was similar to the South Asian population, thus supporting the fact that this study population presented a high South Asian component. Similarly, all the SNPs included for the present study were common in South Asian population with allele frequency (MAF) ≥ 0.05.

**LD estimation**

LD analysis was performed for SNPs in IGF2BP2 (rs4402960, rs1470579, and rs6769511), CDKAL1 (rs4712523, rs4712524, and rs7754840), JAZF1 (rs868745 and rs849134), and CDKN2A/2B (rs564398, rs7020996, and rs2383208). Fig. 1 shows the r² values for the studied SNPs. The r² values were found to be at least 0.83 between the SNPs in IGF2BP2, CDKAL1, and JAZF1. Since the r² values between SNPs in CDKN2A/2B was less than 0.35, haplotypes were constructed and the difference in the haplotype frequencies between cases and controls were analyzed. For the SNPs within CDKN2A/2B, although the frequency of the TCA haplotype was higher in EOT2D subjects when compared with the NGT subjects (p = 0.027), the significance was lost after Bonferroni correction (p < 0.05/7 = 0.007). Table 2 shows the haplotype frequencies of the SNPs within CDKN2A/2B gene.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>NGT (n = 1,183)</th>
<th>EOT2D (n = 1,188)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (yr)</td>
<td>31 ± 8</td>
<td>32 ± 6</td>
<td>0.022*</td>
</tr>
<tr>
<td>Body mass index (kg/m²)</td>
<td>23.5 ± 4.8</td>
<td>26.0 ± 5.4</td>
<td>0.98</td>
</tr>
<tr>
<td>Fasting blood sugar (mg/dL)</td>
<td>86 ± 9</td>
<td>177 ± 62</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>2-h post plasma glucose (mg/dL)</td>
<td>99 ± 19</td>
<td>260 ± 85</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Glycosylated hemoglobin (%)</td>
<td>5.5 ± 0.5</td>
<td>8.9 ± 1.9</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Total cholesterol (mg/dL)</td>
<td>167 ± 42</td>
<td>173 ± 46</td>
<td>0.002</td>
</tr>
<tr>
<td>Log transformed serum triglycerides (mg/dL)</td>
<td>106 ± 65</td>
<td>168 ± 145</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>HDL cholesterol (mg/dL)</td>
<td>44 ± 16</td>
<td>39 ± 11</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Values are presented as mean ± SD.

NGT, normal glucose tolerant; EOT2D, early onset type 2 diabetes; HDL, high-density lipoprotein; SD, standard deviation.
**Association of studied SNPs with EOT2D**

Genotypic distributions of all the variants studied were in HWE and none of the variants studied showed monoallelic condition. As shown in Table 3, six SNPs within five distinct loci rs7020996 (CDKN2A/2B), rs7607980 (COBLL1), rs6769511, rs1470579, rs4402960 (IGF2BP2), rs4812829 (HNF4A), and rs1111875 (HHEX) and rs4506565 (TCF7L2) were found to be significantly associated (p < 0.05) with EOT2D in our South Indian population. However, after Bonferroni correction (p < 0.05/25 = 0.002) the association with EOT2D remained significant only for three SNPs within three distinct gene loci rs1111875 (HHEX: p = 2.0 × 10^{-4}), rs4506565 (TCF7L2: p = 1.0 × 10^{-5}) and rs7020996 (CDKN2A/2B: p = 6.0 × 10^{-4}). A tendency to association (p < 0.05) with EOT2D was also observed with variants in COBLL1 (rs7607980), IGF2BP2 (rs6769511, rs1470579, and rs4402960), and HNF4A (rs4812829) in the present study. The risk allele frequencies for all the variants in the EOT2D and NGT subjects are shown in Table 3.

Logistic regression analyses were performed under the additive model for the variants with significant association to EOT2D identified in the present study: rs1111875 (HHEX), rs4506565 (TCF7L2), and rs7020996 (CDKN2A/2B) after adjusting for potential confounders like age, sex, and BMI (Table 4). The heterozygous and homozygous carriers of allele ‘T’ of the TCF7L2 variant rs4506565 (A/T) had an odds ratio of 1.47 (95% confidence interval [CI], 1.17 to 1.83; p = 0.001) and 1.65 (95% CI, 1.18 to 2.28; p = 0.006) respectively relative to AA homozygote. In the case of the HHEX variant rs1111875 (T/C), heterozygous and homozygous carriers for allele ‘C’ had an odds ratio of 1.13 (95% CI, 0.91 to 1.42; p = 0.27) and 1.58 (95% CI, 1.17 to 2.12; p = 0.003) respectively relative to TT homozygote. However, for the CDKN2A/2B variant rs7020996 heterozygous carrier for the ‘T’ allele showed an association that was protective in nature with odds ratios of 0.65 (95% CI, 0.51 to 0.83; p = 0.0004) while the homozygous carrier showed no significant association (OR, 0.62; 95% CI, 0.27 to 1.39; p = 0.24) relative to the CC homozygote with EOT2D.

Table 5 shows the comparison of the clinical and biochemical characteristics of NGT subjects and the SNPs associated with EOT2D based on their genotype. For the rs4506565 of the TCF7L2 gene, NGT subjects homozygous for the ‘TT’ genotype had increased glycated hemoglobin levels (mean ± SD, 5.5 ± 0.4) when compared with the carriers of the ‘AA’ genotype (5.4 ± 0.4, p = 0.01). In case of the CDKN2A/2B variant rs7020996, carriers of

---

**Table 2. Haplotype frequencies of SNPs in CDKN2A/2B**

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Cases Freq</th>
<th>Controls Freq</th>
<th>(\chi^2)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDKN2A/2B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCA</td>
<td>0.621</td>
<td>0.590</td>
<td>4.901</td>
<td>0.027*</td>
</tr>
<tr>
<td>CCA</td>
<td>0.216</td>
<td>0.221</td>
<td>0.188</td>
<td>0.664</td>
</tr>
<tr>
<td>TTG</td>
<td>0.058</td>
<td>0.072</td>
<td>3.989</td>
<td>0.046</td>
</tr>
<tr>
<td>TCG</td>
<td>0.039</td>
<td>0.036</td>
<td>0.318</td>
<td>0.573</td>
</tr>
<tr>
<td>TTA</td>
<td>0.024</td>
<td>0.033</td>
<td>3.130</td>
<td>0.769</td>
</tr>
<tr>
<td>CTG</td>
<td>0.018</td>
<td>0.020</td>
<td>0.216</td>
<td>0.642</td>
</tr>
<tr>
<td>CTA</td>
<td>0.014</td>
<td>0.020</td>
<td>1.952</td>
<td>0.162</td>
</tr>
</tbody>
</table>

SNP, single nucleotide polymorphism.

*p < 0.05, statistically significant.*

---

**Fig. 1.** Linkage disequilibrium plot for the single nucleotide polymorphisms (SNPs) of IGF2BP2, CDKAL1, JAZF1, and CDKN2A/2B, genes. R^2^ values mentioned in the linkage disequilibrium (LD) plot. LD is seen only between SNPs of the same gene, not across the genes.
### Table 3. The allelic distribution of the variants genotyped in cases and controls

<table>
<thead>
<tr>
<th>No.</th>
<th>Variant</th>
<th>Gene/nearest gene</th>
<th>Chromosome No.</th>
<th>Risk allele</th>
<th>RAF (%)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs7607980 (T/C)</td>
<td>COBLL1</td>
<td>2</td>
<td>T</td>
<td>91.8</td>
<td>93.7</td>
</tr>
<tr>
<td>2</td>
<td>rs3923113 (A/C)</td>
<td>GRB14</td>
<td>2</td>
<td>A</td>
<td>82.6</td>
<td>84.6</td>
</tr>
<tr>
<td>3</td>
<td>rs7593730 (C/T)</td>
<td>RBMS1</td>
<td>C</td>
<td>77.2</td>
<td>79</td>
<td>0.14</td>
</tr>
<tr>
<td>4</td>
<td>rs7578597 (T/C)</td>
<td>THADA</td>
<td>C</td>
<td>89.7</td>
<td>89.3</td>
<td>0.07</td>
</tr>
<tr>
<td>5</td>
<td>rs1801282 (C/G)</td>
<td>PPARG</td>
<td>3</td>
<td>G</td>
<td>90.1</td>
<td>90.3</td>
</tr>
<tr>
<td>6</td>
<td>rs4607103 (T/C)</td>
<td>ADAMTS9</td>
<td>C</td>
<td>46.3</td>
<td>47.5</td>
<td>0.43</td>
</tr>
<tr>
<td>7</td>
<td>rs7578597 (T/C)</td>
<td>THADA</td>
<td>C</td>
<td>48.9</td>
<td>52.3</td>
<td>0.03</td>
</tr>
<tr>
<td>8</td>
<td>rs1470579 (A/C)</td>
<td>A</td>
<td>47.7</td>
<td>50.9</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>rs4402960 (G/T)</td>
<td>G</td>
<td>46.7</td>
<td>50.7</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>rs7754840 (G/C)</td>
<td>CDKAL1</td>
<td>6</td>
<td>C</td>
<td>28.2</td>
<td>28.9</td>
</tr>
<tr>
<td>11</td>
<td>rs4712523 (A/G)</td>
<td>G</td>
<td>29.1</td>
<td>30.7</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>rs4712524 (A/G)</td>
<td>A</td>
<td>26.3</td>
<td>27.8</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>rs864745 (T/C)</td>
<td>JAZF1</td>
<td>7</td>
<td>T</td>
<td>73.9</td>
<td>75.9</td>
</tr>
<tr>
<td>14</td>
<td>rs849134 (A/G)</td>
<td>A</td>
<td>74.2</td>
<td>76.3</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>rs13266634 (C/T)</td>
<td>SLC30A8</td>
<td>8</td>
<td>C</td>
<td>77.5</td>
<td>79.1</td>
</tr>
<tr>
<td>16</td>
<td>rs896854 (T/C)</td>
<td>TP53INP1</td>
<td>T</td>
<td>56.4</td>
<td>59.1</td>
<td>0.07</td>
</tr>
<tr>
<td>17</td>
<td>rs7020996 (C/T)</td>
<td>CDKN2A/2B</td>
<td>9</td>
<td>C</td>
<td>85.5</td>
<td>88.9</td>
</tr>
<tr>
<td>18</td>
<td>rs2383208 (A/G)</td>
<td>A</td>
<td>86.8</td>
<td>87.4</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>rs564398 (T/C)</td>
<td>C</td>
<td>78.7</td>
<td>79.5</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>rs1111875 (C/T)</td>
<td>HHEX</td>
<td>10</td>
<td>T</td>
<td>37</td>
<td>42.5</td>
</tr>
<tr>
<td>21</td>
<td>rs4506565 (A/T)</td>
<td>TCF7L2</td>
<td>T</td>
<td>30.1</td>
<td>36.4</td>
<td>1.00 × 10^-5</td>
</tr>
<tr>
<td>22</td>
<td>rs10906116 (A/G)</td>
<td>CDC123</td>
<td>G</td>
<td>48.4</td>
<td>50</td>
<td>0.32</td>
</tr>
<tr>
<td>23</td>
<td>rs1800574 (C/T)</td>
<td>HNF1A</td>
<td>12</td>
<td>C</td>
<td>7.6</td>
<td>8.7</td>
</tr>
<tr>
<td>24</td>
<td>rs4760790 (G/A)</td>
<td>TSPAN8</td>
<td>G</td>
<td>36.1</td>
<td>38.5</td>
<td>0.08</td>
</tr>
<tr>
<td>25</td>
<td>rs4812829 (A/G)</td>
<td>HNF4A</td>
<td>20</td>
<td>A</td>
<td>32.1</td>
<td>35.5</td>
</tr>
</tbody>
</table>

RAF, risk allele frequency; NGT, normal glucose tolerant; EOT2D, early onset type 2 diabetes.

* p-values with significance threshold after Bonferroni correction p = 0.05/25 = 0.002.

**Variants showing only borderline significance (p < 0.05).**

### Table 4. Association of variants with early onset type 2 diabetes with OR and CI (adjusted for age, sex, and BMI)

<table>
<thead>
<tr>
<th>Variant</th>
<th>Genotype</th>
<th>Genotype frequency, n (%)</th>
<th>OR (95% CI) (adjusted for age, sex and BMI)</th>
<th>p-value</th>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs4506565 (TCF7L2)</td>
<td>AA</td>
<td>506 (49.2)</td>
<td>457 (41.3)</td>
<td>Reference</td>
</tr>
<tr>
<td></td>
<td>rs4506565 (TCF7L2)</td>
<td>TA</td>
<td>412 (40.0)</td>
<td>490 (44.3)</td>
<td>1.47 (1.17–1.83)</td>
</tr>
<tr>
<td></td>
<td>rs4506565 (TCF7L2)</td>
<td>TT</td>
<td>110 (10.7)</td>
<td>158 (14.3)</td>
<td>1.65 (1.18–2.28)</td>
</tr>
<tr>
<td></td>
<td>rs1111875 (HHEX)</td>
<td>TT</td>
<td>438 (40.3)</td>
<td>408 (35.7)</td>
<td>Reference</td>
</tr>
<tr>
<td></td>
<td>rs1111875 (HHEX)</td>
<td>TC</td>
<td>493 (45.4)</td>
<td>505 (44.1)</td>
<td>1.13 (0.91–1.42)</td>
</tr>
<tr>
<td></td>
<td>rs1111875 (HHEX)</td>
<td>CC</td>
<td>156 (14.4)</td>
<td>231 (20.2)</td>
<td>1.58 (1.17–2.12)</td>
</tr>
<tr>
<td></td>
<td>rs7020996 (CDKN2A/2B)</td>
<td>CC</td>
<td>805 (73.4)</td>
<td>879 (79.3)</td>
<td>Reference</td>
</tr>
<tr>
<td></td>
<td>rs7020996 (CDKN2A/2B)</td>
<td>TC</td>
<td>266 (24.2)</td>
<td>214 (19.3)</td>
<td>0.65 (0.51–0.83)</td>
</tr>
<tr>
<td></td>
<td>rs7020996 (CDKN2A/2B)</td>
<td>TT</td>
<td>26 (2.4)</td>
<td>15 (1.4)</td>
<td>0.62 (0.27–1.39)</td>
</tr>
</tbody>
</table>

OR, odds ratio; CI, confidence interval; BMI, body mass index; NGT, normal glucose tolerant; EOT2D, early onset type 2 diabetes.

* p-values were adjusted for age, sex, and BMI.
the ‘CC’ genotype had significantly higher fasting plasma glucose levels (mean ± SD, 87 ± 9 mg/dL), compared to the carriers of the ‘TC’ genotype (85 ± 9 mg/dL, p = 0.03). None of the other biochemical parameters showed any significant differences among the genotypes in either the NGT or the diabetic subjects.

Discussion

There is a rapid increase in the number of subjects diagnosed with T2D below the age of 40 years. However, only few studies have investigated the association of genetic determinants of LOT2D with EOT2D. The present study aimed at investigating the association of 25 variants from 18 distinct gene loci with EOT2D in this South Indian population, has shown association of variants in TCF7L2, CDKN2A/2B, and HHEX with EOT2D with p-values of 1.00 × 10^{-5}, 6.00 × 10^{-4}, and 2.00 × 10^{-4} respectively with power ranging from 67%–83%.

Transcription factor-7-like 2 (TCF7L2) spans 217kb region on chromosome 10q25.3. TCF7L2 is a transcription factor involved in the Wnt signaling pathway and is expressed not only in the \( \beta \)-cells but also in other cell lineages and glucose-metabolizing tissues, including the liver [40]. TCF7L2 identified by Grant et al. [41] is the gene with greater susceptibility to LOT2D in various populations [42-48]. Association of the TCF7L2 variant rs4506565 (A/T) with LOT2D was initially reported by the Wellcome Trust Case Control Consortium with odds ratio of 1.88 (1.56–2.27, p = 5.1 × 10^{-10}) [37]. The association of rs4506565 (TCF7L2) with T2D was later replicated in Middle east [49,50], Tunisian Arabs [51], Lebanese [52], and Indian population [48,53,54]. While in Europeans, the TCF7L2 variant rs4506565 showed evidence for association with EOT2D exceeding genome-wide significance, thus clearly establishing TCF7L2 as a T2D susceptibility gene of substantial importance [42]. Table 6 shows the comparison of the p-value and odds ratio of the SNPs with association to EOT2D identified in the present study with p-value and odds ratio in other population with T2D [24,35,37,49,52,55,56]. The risk allele frequency of the TCF7L2 (rs4506565) in South Indian EOT2D subjects was observed to be 36.4%, compared with 37% in North Indian subjects [48], 49% in Saudi Arabian subjects [49], 44% in Tunisian Arab subjects [51], 46% in Lebanese subjects [52] and 39% in European subjects with LOT2D [37]. In the present study, we have shown a strong association of the TCF7L2 variant with EOT2D in the South Indian population. A previous study by Chidambaram et al. [23] has shown only marginal association of rs4506565 (TCF7L2) with EOT2D in Asian Indians. While, the limitation of the previous study was the small sample size, in the present study, we used a much larger sample size thus increasing the power of the study. Rs4506565 (TCF7L2) also showed a significant association with fasting glucose in non-diabetic subjects in European population [57]. A comprehensive pathway analysis with 529 of the 548 genes within 5 kb of a TCF7L2 binding site by Zhao et al. [58] has shown enriched metabolism-related pathway categories in genes bound by TCF7L2. Lyssenko et al. [59] using an adenovirus system showed 2-fold increased expression of TCF7L2 in human islets, associated with increased insulin gene expression and reduced glucose-stimulated insulin secretion compared with control islets. These studies thus provide evidence for increased expression of TCF7L2 in human islets with altered insulin but not glucagon secretion. TCF7L2 also plays a crucial role in coordinating the expression of proinsulin and its subsequent processing to form mature insulin [60]. In mouse models, removal of TCF4 from B cells in newborn Tgβ2L−/− mice and in adult B cell–specific Tcf7l2 mutants, challenged by fasting or by high-fat diet did not show any affect in their function [61].

**Table 5.** Clinical and biochemical characteristics of the NGT subjects stratified based on rs4506565 (A/T) and rs7020996 (C/T) genotypes

<table>
<thead>
<tr>
<th>Parameter</th>
<th>TCF7L2-rs4506565 (A/T)</th>
<th>CDKN2A/2B-rs7020996 (C/T)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AA</td>
<td>AT</td>
</tr>
<tr>
<td><strong>BMI (kg/m(^2))</strong></td>
<td>23.6 ± 4.8</td>
<td>23.7 ± 4.6</td>
</tr>
<tr>
<td><strong>Fasting plasma glucose (mg/dL)</strong></td>
<td>86 ± 8</td>
<td>88 ± 9</td>
</tr>
<tr>
<td><strong>2-h plasma glucose (mg/dL)</strong></td>
<td>98 ± 20</td>
<td>99 ± 19</td>
</tr>
<tr>
<td><strong>Total cholesterol (mg/dL)</strong></td>
<td>170 ± 51</td>
<td>168 ± 33</td>
</tr>
<tr>
<td><strong>Triglycerides (mg/dL)</strong></td>
<td>106 ± 61</td>
<td>111 ± 79</td>
</tr>
<tr>
<td><strong>HDL cholesterol (mg/dL)</strong></td>
<td>45 ± 16</td>
<td>45 ± 15</td>
</tr>
<tr>
<td><strong>LDL cholesterol (mg/dL)</strong></td>
<td>101 ± 30</td>
<td>101 ± 30</td>
</tr>
<tr>
<td><strong>Glycated hemoglobin (%)</strong></td>
<td>5.4 ± 0.4</td>
<td>5.4 ± 0.4</td>
</tr>
</tbody>
</table>

NGT, normal glucose tolerant; BMI, body mass index; HDL, high-density lipoprotein; LDL, low density lipoprotein.

\(^{a}\)Log transformed values.

\(^{b}\)p = 0.01 compared to AA (adjusted for age and sex).

\(^{c}\)p = 0.03 compared to TC (adjusted for age and sex).

https://doi.org/10.5808/GI.2020.18.3.e27
The present study has also confirmed the association of the CDKN2A/2B variant rs7020996 with EOT2D in the South Indian population, which was also earlier suggested by Chidambaram et al. [23] in Asian Indian population. The CDKN2A/2B locus at chromosome 9p21 was tagged as hot spot for association with LOT2D in a series of GWAS [24,26,30,35]. Zeggini et al. [24] in a meta-analysis study initially reported on the association of rs7020996 of the CDKN2A/2B gene with T2D in European population with OR 1.26 (1.15–1.38) (p = 1.8 × 10^{-7}). Though CDKN2A/2B was reported to influence diabetes risk across varied ethnicities, not many studies have replicated the association of the CDKN2A/2B variant rs7020996 with T2D. Replication of the CDKN2A/2B variant rs7020996 both with EOT2D and LOT2D by our own group has shown significant association [23,54].

The risk allele frequency of CDKN2A/2B (rs7020996) in South Indian EOT2D subjects was observed to be 88.9% in EOT2D subjects, compared with 88% in Asian Indian EOT2D subjects [23] and 91% in South Indian LOT2D subjects [53]. The CDKN2A/2B genes are expressed in adipocytes and pancreatic islets. CDKN2A and CDKN2B encodes p16INK4a and p15INK4b and inhibit the activity of CDK4 and CDK6, respectively. The p16INK4a encoded by CDKN2A is a tumor suppressor and inhibits CDK4 (cyclin-dependent kinase) influencing pancreatic β cell proliferation, through decreased cell mass and subsequent decreased insulin release. The increased insulin demand possibly increases the susceptibility to T2D [43]. In murine models, overexpression of Cdkn2a leads to decreased islet proliferation in aging mice and that of Cdkn2b leads to islet hypoplasia and diabetes [62]. Study by Kong et al. [63] investigating the mechanism through which the GWAS identified CDKN2A/2B variants increase the T2D risk showed the impact of CDKN2A/2B SNPs mediated through β-cell mass but not β-cell function.

**HHEX**, located on chromosome 10q23.33 encodes a 270 amino-acid protein and was identified to be strongly associated with LOT2D in European populations by Scott et al. [35] with odds ratio (OR) 1.13 (1.09–1.17), p = 5.7 × 10^{-10}. The association was later replicated in Danish [43], Japanese [55,64,65], Korean [56], Han Chinese [66], and in Tunisian population [67]. However, studies performed in Indians [53,54,68-70] and African American population [71] failed to replicate the association observed in various populations. The lack of association of the HHEX variant among various Indian populations Khatri Sikhs [68], Hyderabad population [69], an endogamous North Indian population [53] and South Indian population [54] could possibly due to the insufficient sample size, population stratification/admixture or due to confounders. Meta-analysis of 26 studies with 45,792 cases and 65,083 controls, also revealed a stronger association between

### Table 6: Comparison of association results for T2D loci in various population with the present study

<table>
<thead>
<tr>
<th>Gene</th>
<th>Present study (OR 95% CI)</th>
<th>European (OR 95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1111875</td>
<td>1.26 (1.15–1.38)</td>
<td>1.39 (1.20–1.63)</td>
<td>0.00013</td>
</tr>
<tr>
<td>rs7020996</td>
<td>1.30 (1.17–1.45)</td>
<td>1.31 (1.14–1.49)</td>
<td>0.0006</td>
</tr>
<tr>
<td>rs6568865</td>
<td>1.15 × 10^{-7}</td>
<td>1.10 × 10^{-7}</td>
<td>0.153</td>
</tr>
<tr>
<td>T2D, type 2 diabetes; OR, odds ratio, CI, confidence interval.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
rs1111875 and risk for T2D in East Asian population (OR, 1.19) than in white populations (OR, 1.15) and Indian population (OR, 1.13) [72]. Intriguingly, a meta-analysis by Chauhan et al. [73] performed in Indian population successfully replicated the association of the HHEX variant with LOT2D in North Indian population. The risk allele frequency of HHEX (rs1111875) in South Indian was observed to be 42.3% in EOT2D subjects, compared with 32% in Japanese [55], 36% in Korean [56], 32% in Han Chinese [66], and 52% in European [35] subjects with T2D. Giannini et al. [74] showed association of the HHEX variant rs1111875 with prediabetes among obese youth. In European and Finnish population, the rs1111875 (HHEX) showed association with lower birth weight providing evidence for the ‘fetal programming hypothesis’ suggestive of decreased insulin secretion or action with reduced intrauterine growth and thereby lower birth weight as well as susceptibility to LOT2D [75,76]. HHEX genes encodes a transcription factor involved in the Wnt signaling pathway and also plays important role in many biological processes including cell cycle regulation, organ development, and cell differentiation via both transcriptional activation and repression [77]. Recent functional studies have identified HHEX as the first transcription factor required for δ-cell maintenance mediated through paracrine regulation of β-cell activity. The same study also showed misregulated HHEX expression with paracrine control of insulin secretion, leading to accelerated β-cell exhaustion and failure [78]. In HHEX-null mice pancreatic β cells was defined for its involvement in β-cell differentiation and function and failure of ventral pancreas development [79]. However, the exact mechanism through which the TCF7L2, CDKN2A/2B, and HHEX exerts its effect on T2D is still unclear. Additionally, a recent study by Mohan et al. [80] has also suggested the predominant role of beta-cell dysfunction than insulin resistance in the pathogenesis of T2D among Asian Indian youth.

With regard to the other variants genotyped in the present study rs6769511, rs1470579, rs4402960 of IGF2BP2, rs7607980 of COBLL1, and rs4812829 of HNF4A showed only a nominal association (p < 0.05) in terms of the association with EOT2D in this South Indian population with power ranging from 34% to 46%. The nominal association of these variants rs6769511, rs1470579, rs4402960 (IGF2BP2), rs7607980 (COBLL1), and rs4812829 (HNF4A) observed with EOT2D in the present study may however be due to the relatively small sample size, which is one of the major limitations of this study. Moreover, we have replicated only 25 gene variants with EOT2D in the present study out of the several hundred gene variants identified with LOT2D.

The significance of EOT2D is that, due to the earlier onset of diabetes these individuals are at increased susceptibility to complications of diabetes includes neuropathy, retinopathy, and cardiovascular disease compared to LOT2D. Our results highlight the need for larger prospective studies to identify the effect of genetic variants implicated in the development of EOT2D. To conclude, the present study is the first study to confirm the association of gene variants associated with EOT2D in South Indian population, and shows the importance of the HHEX variants with EOT2D.

**ORCID**

Samuel Liju: https://orcid.org/0000-0003-1878-0319
Manickam Chidambaram: https://orcid.org/0000-0001-8750-8576
Viswanathan Mohan: https://orcid.org/0000-0001-5038-6210
Venkatesan Radha: https://orcid.org/0000-0002-4281-0250

**Authors’ Contribution**

Conceptualization: VR. Data curation: VR. Formal analysis: SL. Funding acquisition: VM, VR. Methodology: MC, SL. Writing – original draft: VR, SL. Writing – review & editing: VR, VM.

**Conflicts of Interest**

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

**Acknowledgments**

This study was supported by funding from Indian Council of Medical Research and University of Minnesota through the project ‘Diabetes: Genetic Susceptibility in the Asian Indian Population’ (RHN/Adhoc/19/2011-2012 dated 13.7.2011). We thank Dr.Myron Gross (University of Minnesota) for all the support. LS was supported by CSIR senior research fellowship.

**Supplementary Materials**

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

**References**

2. Wilmot E, Idris I. Early onset type 2 diabetes: risk factors, clinical...
impact and management. Ther Adv Chronic Dis 2014;5:234-244.
26. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Re-


37. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-678.


47. Bodhini D, Radha V, Dhar M, Narayani N, Mohan V. The rs12253372(G/T) and rs7903146(C/T) polymorphisms of the HHEX gene are associated with type 2 diabetes mellitus in the Indian population. Metabolism 2007;56:493-498.


19:3100-3106.


58. Zhao J, Schug J, Li M, Kaestner KH, Grant SF. Disease-associated loci are significantly over-represented among genes bound by transcription factor-7-like 2 (TCF7L2) in vivo. Diabetologia 2010;53:2340-2346.


Molecular characterization and functional annotation of a hypothetical protein (SCO0618) of *Streptomyces coelicolor* A3(2)

Nadim Ferdous¹, Mahjerin Nasrin Reza¹, Md. Tabassum Hossain Emon¹, Md. Shariful Islam², A. K. M. Mohiuddin¹, Mohammad Uzzal Hossain³*

¹Department of Biotechnology and Genetic Engineering, Faculty of Life Science, Mawlana Bhashani Science and Technology University, Tangail 1902, Bangladesh
²Laboratory of Reproductive and Developmental Biology, Hoikaido University, Sapporo 060-0808, Japan
³Bioinformatics Division, National Institute of Biotechnology, Savar, Dhaka 1349, Bangladesh

*Corresponding author:
E-mail: uzzalbge10044@gmail.com

**Streptomyces coelicolor** is a gram-positive soil bacterium which is well known for the production of several antibiotics used in various biotechnological applications. But numerous proteins from its genome are considered hypothetical proteins. Therefore, the present study aimed to reveal the functions of a hypothetical protein from the genome of *S. coelicolor*. Several bioinformatics tools were employed to predict the structure and function of this protein. Sequence similarity was searched through the available bioinformatics databases to find out the homologous protein. The secondary and tertiary structure were predicted and further validated with quality assessment tools. Furthermore, the active site and the interacting proteins were also explored with the utilization of CASTp and STRING server.

The hypothetical protein showed the important biological activity having with two functional domain including POD-like_MBL-fold and rhodanese homology domain. The functional annotation exposed that the selected hypothetical protein could show the hydrolase activity. Furthermore, protein-protein interactions of selected hypothetical protein revealed several functional partners those have the significant role for the bacterial survival. At last, the current study depicts that the annotated hypothetical protein is linked with hydrolase activity which might be of great interest to the further research in bacterial genetics.

**Keywords:** genome, hydrolases, hypothetical protein, modeling, *Streptomyces coelicolor*

---

**Introduction**

*Streptomyces coelicolor* A3(2) is one of the best studied representatives amongst other members of the genus *Streptomyces* [1]. Like the streptomyces genus in general, it lives in soil [2]. It is considered a model organism to study soil bacteria [3], which has been studied genetically for about 60 years [4]. They have the capability to degrade chitin and other compounds that are difficult to degrade which makes them especially important [5]. This bacterium produces a range of secondary metabolites, including actinorhodin, undecylprodigiosin, calcium-dependent antibiotic, methylenomycin A and perimycin [6]. Some of them have antifungal activities also. So, *Streptomyces coelicolor* A3(2) has the potential to make such secondary metabolites, and metagenomic analysis has revealed it has...
revealed it has tremendous quantities of significant biosynthetic gene sets \[7,8\]. These characteristics have elicited biotechnological interest in this bacterium and have aroused the interest of researchers in the past few years to investigate the different proteins involved in secondary metabolites production. As an example, it is recently found that albafavenone, germicidin A, and chalcone are produced during germination of \textit{Streptomyces coelicolor} \[9\] and the genes responsible for the biosynthesis of streptomycese secondary metabolites are generally clustered with high expression of regulation \[10\]. Another research shows that a group of mtbH-like genes in \textit{S. coelicolor} are necessary for some secondary metabolite production \[11\]. \textit{Streptomyces coelicolor} has three such genes, cloY is one of them \[11\]. When all three genes were absent, clorobiocin, an antibiotic which inhibits the enzyme DNA gyrase was produced only in very small amounts, but when cloY was restored, clorobiocin was produced at a more significant level \[11\].

\textit{Streptomyces coelicolor} A3(2) is reported to have 8,667,507 base pair linear chromosome, containing the largest number of genes so far discovered in a bacterium \[10\]. The genes so far predicted are 7,825 which include more than 20 clusters coding for known or predicted secondary metabolites \[10\]. However, there are many proteins of this bacterium which are considered hypothetical proteins as their structures and biological functions are not yet known. These proteins can be very important and their annotation can lead to knowledge about new structures, pathways, and functions. Thus, bioinformatics approaches can play an important role in predicting and analyzing various forms of structure of those hypothetical proteins, their biological functions as well as protein-protein interactions.

With the advancement of in-silico analysis, it became easier to annotate function to a hypothetical protein using various bioinformatic tools. Thus, the purpose of this study was to assign structural and biological function to the hypothetical protein SCO0618 (accession No. NP_624929.1) of \textit{S. coelicolor} for an improved understanding of the protein. Subcellular localization, secondary structure, and active site were predicted and protein-protein interaction was analyzed. Further, a good quality model of the SCO0618 was tried to generate using homology modeling techniques.

**Methods**

**Sequence retrieval and similarity identification**

The sequence information of the hypothetical protein (NP_624929.1) was retrieved from the NCBI database. The sequence was then collected as a FASTA format sequence and submitted to several prediction servers for the in-silico characterization (Table 1). To get the initial prediction about the function of the targeted hypothetical protein, similarity search was performed with the NCBI protein Database (https://www.ncbi.nlm.nih.gov/) against non-redundant and SwissProt \[12\] database to find out the proteins that might have structural similarities with that of the uncharacterized protein by using BLASTp program \[13\].

**Multiple sequence alignment and phylogeny analysis**

Multiple sequence alignment was performed using MUSCLE server of EBI (https://www.ebi.ac.uk/Tools/msa/muscle/) \[14\] and visualized using the CLC Sequence Viewer 7.0.2 (http://www.clcbio.com). The phylogeny analysis was done by using the webtool Phylogeny.fr (http://phylogeny.lirmm.fr/) \[15\].

**Physiochemical properties analysis**

The physical and chemical properties including molecular weight, theoretical pl, amino acid composition, atomic composition, extinction coefficient, estimated half-life, total number of negatively charged residues (Asp + Glu), total number of positively charged residues (Arg + Lys), instability index, aliphatic index, and grand average of hydropathicity (GRAVY) predictions, etc. were performed by the ProtParam (http://web.expasy.org/protparam/) \[16\] tool of ExPASy.

**Subcellular localization analysis**

Subcellular localization was predicted by CELLO \[17\]. Results were also cross-checked with subcellular localization predictions obtained from PSORTb \[18\], PSLpred \[19\], and SOSUIgramN \[20\]. TMHMM \[21\], HMMTOP \[22\], and CCTOP \[23\] were used for the topology prediction.

**Conserved domain, motif, fold, coil, family, and superfamily identification**

Search carried out at conserved domain database (CDD, available at NCBI) \[24\], for conserved domain. Protein motif search was carried out using Motif (Genome Net) server \[25\]. Pfam \[26\] and SuperFamily \[27\] database searches were done to assign the protein’s evolutionary relationships. For the detection of coiled-coil conformation within the protein, the COILS server \[28\] was employed. Protein sequence analysis and classification server InterProScan \[29\] was employed for the functional analysis of the protein. For protein folding pattern recognition, PFP-FunD SeqE server \[30\] was used. And STRING 10.0 \[31\] search was carried out for the identification of possible functional interaction network of the protein.

**Secondary structure prediction**

PSI-blast based secondary structure Prediction (PSIPRED) \[32\]
and self-optimized prediction method with alignment (SOPMA) servers were used for the prediction of the proteins' secondary structure \[33\].

Three-dimensional structure prediction
The three-dimensional structure was predicted by HHpred server (https://toolkit.tuebingen.mpg.de/tools/hhpred) \[34\] of the Max Planck Institute for Developmental Biology, Tübingen which is based on the pairwise comparison profile of hidden Markov models (HMMs). For higher accuracy, the 3D structure was predicted on the basis of best scoring template. Later the 3D structure was refined through YASARA energy minimization server \[35\].

Model quality assessment
Finally, PROCHECK (https://servicesn.mbi.ucla.edu/PROCHECK/) \[36\], Verify3D (http://nihserver.mbi.ucla.edu/Verify_3D/) \[37\], and ERRAT Structure Evaluation server (https://servicesn.mbi.ucla.edu/ERRAT/) \[38\] were used for quality assessment of the predicted three dimensional structure.

Active site detection
The active site of the protein was determined by the Computed Atlas of Surface Topography of Protein (CASTp) (http://sts.bengr.uic.edu/castp/) \[39\] which provides an online resource for locating, delineating, and measuring concave surface regions on three-dimensional structures of proteins.

Results and Discussion
The work-flow of the study was shown in Fig. 1.

Sequence and similarity information
The BLASTp result against non-redundant and SwissProt database showed homology with other hydrolase and sulfurtransferase proteins (Tables 2 and 3). Multiple sequence alignment (Supplementary Fig. 1) was considered the FASTA sequences of the hypothetical protein (SCO0618) and the homologous annotated proteins. For the confirmation of homology assessment between the proteins, down to the complex and subunit level, phylogenetic analysis was also performed. Phylogenetic tree was constructed based on the alignment and BLAST result which gives the similar concept about the protein (Fig. 2). The distances between branches are also included.

Physicochemical features
The protein consist of 461 amino acids, among the most abundant

Table 1. Tools used for the in-silico characterization of hypothetical protein SCO0618

<table>
<thead>
<tr>
<th>No.</th>
<th>Server name</th>
<th>Reference</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BLASTp</td>
<td>Johnson et al. (2008) [13]</td>
<td>Similarity search</td>
</tr>
<tr>
<td>2</td>
<td>protBLAST</td>
<td>Altschul et al. (1999) [40]</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>MUSCLE</td>
<td>Madeira et al. (2019) [14]</td>
<td>Multiple sequence alignment</td>
</tr>
<tr>
<td>4</td>
<td>ProtParam</td>
<td>Gasteiger et al. (2003) [16]</td>
<td>Physicochemical characterization</td>
</tr>
<tr>
<td>5</td>
<td>PSORTb</td>
<td>Yu et al. (2010) [18]</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>PSLpred</td>
<td>Bhasin et al. (2005) [19]</td>
<td>Subcellular localization prediction</td>
</tr>
<tr>
<td>7</td>
<td>CELLO</td>
<td>Yu et al. (2006) [17]</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SOSUIGramN</td>
<td>Imai et al. (2008) [20]</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>TMHMM</td>
<td>Moller et al. (2001) [21]</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>HHpred</td>
<td>Tusnady and Simon (2001) [22]</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>CCTOP</td>
<td>Dobson et al. (2015) [23]</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Motif</td>
<td>Kanehisa et al. (2002) [25]</td>
<td>Motif discovery</td>
</tr>
<tr>
<td>13</td>
<td>Pfam</td>
<td>Finn et al. (2014) [26]</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Superfamily</td>
<td>Wilson et al. (2007) [27]</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>COILS</td>
<td>Lupas et al. (1991) [28]</td>
<td>Coiled-coil motif identification</td>
</tr>
<tr>
<td>16</td>
<td>PFP-FunDSeqE</td>
<td>Shen and Chou (2009) [30]</td>
<td>Fold recognition</td>
</tr>
<tr>
<td>17</td>
<td>InterPro</td>
<td>Hunter et al. (2009) [29]</td>
<td>Functional classification</td>
</tr>
<tr>
<td>18</td>
<td>STRING</td>
<td>Szklarczyk et al. (2015) [31]</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>PSIPRED</td>
<td>McGuffin et al. (2000) [32]</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>SOPMA</td>
<td>Geourjon and Deleage (1995) [33]</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>HHpred</td>
<td>Zimmermann et al. (2018) [34]</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>PROCHECK</td>
<td>Laskowski et al. (1993) [36]</td>
<td>Structure verification</td>
</tr>
<tr>
<td>23</td>
<td>Verify3D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>ERRAT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
were Ala (92) followed by Val (51), Arg (42), Gly (41), Leu (40), Asp (32), Glu (30), Pro (26), Thr (21), Ser (19), His (17), Phe (11), Ile (10), Tyr (8), Trp (6), Asn (5), Gln (4), Met (4), and Cys (2). The calculated molecular weight was 48216.15 Da and theoretical pI was 5.27 indicating the protein to be negatively charged. Total number of positively charged residues (Arg + Lys) and the total number of negatively charged residues (Asp + Glu) were found to be 62 and 42, respectively. The computed instability index was 32.67 classifying the protein as stable one. Aliphatic index was 94.34 which gives an indication of proteins’ stability over a wide temperature range. The GRAVY was 0.053. Positive value of GRAVY indicates that the protein is polar. Protein half-life computed was found to be 30 h in mammalian reticulocytes (in vitro), > 20 hours in yeast (in vivo), > 10 h in Escherichia coli (in vivo). And the molecular formula of protein was identified as C_{2119}H_{3350}N_{636}O_{643}S_{6}.

**Functional annotation of the hypothetical protein**

The conserved domain search tool revealed that this hypothetical protein sequence was found to have two domains, MBL-fold metallo-hydrolase domain (accession No. cd07724) and rhodanese homology domain (RHOD) (accession No. cd00158). The result was also checked by two other domain searching tools namely InterProScan and Pfam. Pfam server predicted the rhodanese like domain at 362–444 amino acid residues with an e-value of 2.3e-05 and metallo-beta-lactamase superfamily domain at 16–171 amino acid residues with an e-value of 4.7e-07. InterproScan server predicted rhodanese like domain at 249–454 amino acid residues.

**Table 2.** Similar protein obtained from non-redundant UniProt KB/SwissProt sequences

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Organism</th>
<th>Protein name</th>
<th>Identity (%)</th>
<th>Score</th>
<th>e-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP_011027250.1</td>
<td>Streptomyces</td>
<td>MULTISPECIES: MBL fold metallo-hydrolase</td>
<td>100</td>
<td>889</td>
<td>0.0</td>
</tr>
<tr>
<td>WP_003978243.1</td>
<td>Streptomyces</td>
<td>MULTISPECIES: MBL fold metallo-hydrolase</td>
<td>99.57</td>
<td>886</td>
<td>0.0</td>
</tr>
<tr>
<td>WP_121713050.1</td>
<td>Streptomyces sp. ESN91</td>
<td>MBL fold metallo-hydrolase</td>
<td>99.13</td>
<td>884</td>
<td>0.0</td>
</tr>
<tr>
<td>WP_016325181.1</td>
<td>Streptomyces lividans</td>
<td>MBL fold metallo-hydrolase</td>
<td>99.35</td>
<td>883</td>
<td>0.0</td>
</tr>
<tr>
<td>WP_093455449.1</td>
<td>Unclassified Streptomyces</td>
<td>MULTISPECIES: MBL fold metallo-hydrolase</td>
<td>99.35</td>
<td>883</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Table 3.** Similar protein obtained from UniProt database

<table>
<thead>
<tr>
<th>Entry name</th>
<th>Organism</th>
<th>Protein name</th>
<th>Identity (%)</th>
<th>Score</th>
<th>e-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q88FF3.1</td>
<td>Pseudomonas putida KT2440</td>
<td>Hydroxyacylglutathione hydrolase</td>
<td>32.97</td>
<td>60.1</td>
<td>6e-09</td>
</tr>
<tr>
<td>B1JBN3.1</td>
<td>Pseudomonas putida W619</td>
<td>Hydroxyacylglutathione hydrolase</td>
<td>31.49</td>
<td>58.2</td>
<td>3e-08</td>
</tr>
<tr>
<td>B0KN02.1</td>
<td>Pseudomonas putida GB-1</td>
<td>Hydroxyacylglutathione hydrolase</td>
<td>30.77</td>
<td>57.8</td>
<td>3e-08</td>
</tr>
<tr>
<td>ASW167.1</td>
<td>Pseudomonas putida F1</td>
<td>Hydroxyacylglutathione hydrolase</td>
<td>31.32</td>
<td>55.8</td>
<td>1e-07</td>
</tr>
<tr>
<td>D3RPB9.1</td>
<td>Allochromatium vinosum DSM 180</td>
<td>Sulfurtransferase</td>
<td>33.33</td>
<td>51.6</td>
<td>3e-07</td>
</tr>
</tbody>
</table>

**Fig. 1.** A complete workflow of the study.

**Fig. 2.** Phylogenetic trees with true distance of different hydrolases proteins.

https://doi.org/10.5808/GI.2020.18.3.e28
no acid residues and metallo-beta-lactamas domain at 13–180 amino acid residues. Rhodanese like domain, lactamase-B and MreB-Mbl domains were also found by Motif server. Superfamily search revealed present of Metallo-hydrolase/oxidoareductase and rhodanese/cell cycle control phosphatase superfamily. β-Lactamases can catalyze the hydrolysis of a wide range of β-lactam antibiotics. Members of the MBL-fold metallohydrolase superfamily are mainly hydrolytic enzymes which carry out various biological functions. Both the active and inactive version of the Rhodanese domain in a variety of proteins including certain protein phosphatases, sulfide dehydrogenases, certain stress proteins and sulfuryl transferases, where they are thought to play a regulatory role in multidomain proteins (Fig. 3). All these results confirm the presence of hydrolytic enzyme containing domains in this protein.

Fold pattern recognition by PFP-FunDSeqE tool revealed the presence of a ‘(TIM)-barrel’ fold within the protein sequence. (TIM)-barrel structure is generally eight stranded α/β barrel. The x-axis of the graph represents the position in the protein of amino acid number (starting at the N-terminus) and the y-axis shows the coiled coil whereas ‘Window’ refers to the width of the amino acid ‘window’ that is scanned at one time (Fig. 4).

Subcellular localization nature
Subcellular localization analysis was predicted by CELLO and validated by PSORTb, SOSUIGramN, and PSLpred. The subcellular localization of the hypothetical protein was predicted to be a cytoplasmic protein (Table 4). Absent of transmembrane helices predicted by THMM and HMMTOP also emphasizes the result of being a cytoplasmic protein. Also, CCTOP server predicted that the query protein was not a transmembrane protein. All these results summarize the protein as a cytoplasmic one.

Secondary structure analysis
The SOPMA secondary structure prediction server analysis revealed the proportions of alpha helix, beta turn, extended strand, and the random coil of protein as 31.89%, 9.11%, 18.87%, and 40.13%, respectively (Supplementary Fig. 2).

Three-dimensional structure analysis
Prediction of 3D structure was done by HHpred server. The server

![Fig. 4. Coil depicts the heptads corresponding to the residue windows 14 (green), 21(blue), and 28 (red).](https://doi.org/10.5808/GI.2020.18.3.e28)
predicted 3D structure of the protein with 100% identity with the highest scoring template (PDB ID: 3TP9_A) (Fig. 5). 3TP9 is the crystal structure of *Alicyclobacillus acidocaldarius* protein with β-lactamase and rhodanese domains. This protein is a homo-dimer which has two chains (chain A and chain B) and the chain A was used as template to build the model. Validation of the predicted three-dimensional model was assessed by PROCHECK through Ramachandran plot analysis, where the distribution of φ and ψ angle in the model within the limits are shown (Table 5, Fig. 6). Residues in the most favored regions covered 90.9%, which is the quality of a valid model. Finally, the established model of 3D structure for the target sequence was verified by structure validation server Verify3D and ERRAT. In the Verify3D graph, 92.73% of the residues have averaged 3D-1D score ≥ 0.2 which indicates that the environmental profile of the model is good and the overall quality factor predicted by the ERRAT server was 69.0583 indicates a good model. The 3D structure was later modified by YASARA energy minimization server. The energy calculated before energy minimization was –77,930.2 kJ/mol whereas after energy minimization (through 3 round of steepest descent method), it was changed to far less value of –244,148.6 kJ/mol making the modeled structure more stable one.

**Table 4. Subcellular localization analysis**

<table>
<thead>
<tr>
<th>No.</th>
<th>Analysis</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CELLO 2.5</td>
<td>Cytoplasmic localization</td>
</tr>
<tr>
<td>2</td>
<td>PSORTb</td>
<td>Cytoplasmic localization</td>
</tr>
<tr>
<td>3</td>
<td>SOSUIGramN</td>
<td>Cytoplasmic localization</td>
</tr>
<tr>
<td>4</td>
<td>PSIPred</td>
<td>Cytoplasmic protein</td>
</tr>
<tr>
<td>5</td>
<td>TMHMM 2.0</td>
<td>No transmembrane helices present</td>
</tr>
<tr>
<td>6</td>
<td>HMMTOP</td>
<td>No transmembrane helices present</td>
</tr>
<tr>
<td>7</td>
<td>CCTOP</td>
<td>Not transmembrane protein</td>
</tr>
</tbody>
</table>

**Table 5. Ramachandran plot statistics of the hypothetical protein**

<table>
<thead>
<tr>
<th>Ramachandran plot statistics</th>
<th>No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residues in the most favored regions [A, B, L]</td>
<td>351 (90.9)</td>
</tr>
<tr>
<td>Residues in the additional allowed regions [a, b, l, p]</td>
<td>26 (6.7)</td>
</tr>
<tr>
<td>Residues in the generously allowed regions [a, b, l, p]</td>
<td>8 (2.1)</td>
</tr>
<tr>
<td>Residues in the disallowed regions</td>
<td>1 (0.3)</td>
</tr>
<tr>
<td>No. of non-glycine and non-proline residues</td>
<td>386</td>
</tr>
<tr>
<td>No. of end-residues (excl. Gly and Pro)</td>
<td>2</td>
</tr>
<tr>
<td>No. of glycine residues (shown in triangles)</td>
<td>41</td>
</tr>
<tr>
<td>No. of proline residues</td>
<td>25</td>
</tr>
<tr>
<td>Total No. of residues</td>
<td>454</td>
</tr>
</tbody>
</table>

**Fig. 5.** Predicted three-dimensional structure of the hypothetical protein.

**Fig. 6.** Ramachandran plot of modelled structure validated by PROCHECK program.

**Protein-protein interaction analysis**

STRING 10.0 search was carried out for the identification of possible functional interaction network of the protein [31]. The identified functional partners with scores were; SCO0619 (0.970), SCO0620 (0.743), SCO0621 (0.739), groES (0.568), SCO2899 (0.568), guaA (0.545), SCO6160 (0.520), pheT (0.508), SCO5178 (0.485), polA (0.473). Of them, SCO0619 is a possible membrane protein. The others are two hypothetical proteins, two chaperonins, GMP synthase, multifunctional fusion protein, phenylalanine tRNA ligase β subunit, putative sulfurylase, and DNA polymerase I (Fig. 7).

**Active site of the hypothetical protein**

The predicted active site of the protein found that 42 amino acids are involved in potent active site (Fig. 8). The best active site was
found in areas with 613.075 and a volume of 608.774 amino acids. The amino acid residues in the active site were shown in Supplementary Fig. 3.

**Conclusion**
The identification of protein functions is fundamental for the understanding of biological processes. So, this study was aimed to determine the structural and biological function of SCO0618, a hypothetical protein of this bacterium through an in-silico approach. The identified protein revealed several characteristics such as cytoplasmic nature, hydrolytic enzymes containing domain presence, ‘(TIM)-barrel fold presence, and hydrolase activity emphasize the significance of this protein. These characters of the hypothetical protein will strengthen basic knowledge on *S. coelicolor*. So, extended in-vitro research has to be carried out to experimentally validate the possibilities shown here and to find out the proteins’ role in biotechnology.

**Fig. 7.** String network analysis of the hypothetical protein, indicates as SCO0618.

**Fig. 8.** Active site of the hypothetical protein. Here the red sphere indicates the active site of the protein.

**ORCID**
Nadim Ferdous: https://orcid.org/0000-0003-4240-6829
Mahjerin Nasrin Reza: https://orcid.org/0000-0002-0331-1416
Md. Tabassum Hossain Emon: https://orcid.org/0000-0001-5521-4565
Md. Shariful Islam: https://orcid.org/0000-0002-7631-882X
A. K. M Mohiuddin: https://orcid.org/0000-0003-4188-6592
Mohammad Uzzal Hossain: https://orcid.org/0000-0002-9957-122X

**Authors’ Contribution**
Conceptualization: NF, MUH. Data curation: NF, MNR. Formal analysis: NF, MNR, MTHE, MUH. Methodology: NF, MNR. Writing - original draft: NF, MNR. Writing - review & editing: MUH, MTHE, MSI, AKMM.

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

**Acknowledgments**
We are grateful to the book of Gobeshonay Bioinformatics-1st Part.

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

**References**
6. Hobbs G, Obanye AI, Petty J, Mason JC, Barratt E, Gardner DC,


Introduction

Maize (Zea mays L.) is one of the most important crops in the world and is used as food, animal feed, and biofuel. Among the total production of maize in the world, about 20% is used for human consumption, while a high percentage of the rest is used as animal feed [1]. Until now, white and yellow color maize are more popular in the world than colored maize, such as those with black, red, purple, and blue kernel [2-4]. Although most consumers prefer white or yellow maize, colored maize contains phytochemicals and many secondary metabolites, such as phenolic compounds and flavonoids [2]. Thus, there has been increasing demand recently for the development of colored kernel maize because of its anti-cancer, anti-inflammatory, and anti-oxidant traits [5]. Due to such interest in colored maize, Korean maize breeding programs have been focusing on the development of functional purple maize varieties, resulting in the development of some purple waxy maize varieties, such as Miheugchal [6], Sekso2 [7], Heukjinjuchal [8], and Cheongchunchal [9]. Therefore, research into the genes associated with seed pigmentation is an essential step to developing breeding materials that can foster a number of varieties in colored maize breeding. Generally, the pigments of colored maize are determined by flavonoids, mainly anthocyanins, and especially phlobaphenes [2,10,11]. Structural flavonoid genes are responsible for the anthocyanin pigmentation in diverse tissues, such as leaves, stems, anthers, and ker-
nels [12]. Previous researches on anthocyanin pigmentation isolated a number of genes, such as c2 (colorless2), a1 (anthocyaninless1), a2 (anthocyaninless2), chi1 (chalcone flavanone isomerase1), fht1 (flavanone 3-hydroxylase1), pr1 (red aleurone1), bz1 (Bronze1), bz2 (Bronze2), and transcription factors, such as Sn1 (scutellar node color1), r1 (colored1), b1 (colored plant1), c1 (colorless1), p1 (Pericarp color1), pl (purple plant), and pac1 (pale aleurone color1) in the anthocyanin metabolism pathway of maize [10,13,14].

As mentioned above, many genes are involved in the pathway of anthocyanins. A key means to study the mechanisms of pigmentation of maize seed is to confirm its gene expression level and its functions. Thus, identification of the differential expression genes (DEGs) is crucial in understanding the mechanism of seed pigmentation. Because of recent advances in next-generation sequencing, transcriptome analysis by RNA sequencing (RNA-Seq) allows cost-effective research on both the sequence and transcriptional variations for functional genomic analysis, especially in large and repetitive sequence-rich genomes such as maize [15,16]. The transcriptome is the total set of transcripts among different cells or tissue types, as well as their developmental and physiological stages. An analysis of transcriptome dynamics provides the opportunity to imply the function of unannotated genes, identify critical network hub-related genes, and interpret the cellular processes associated with development [17]. Furthermore, this method can obtain single nucleotide polymorphisms, which are of particular interest when studying allele-specific expression patterns [18]. RNA-Seq for whole-transcriptome researches has been used to identify the spatial and temporal expression patterns of transcriptome in different cells or tissues, along with the development stages in maize [19-23]. A comparison of the transcriptomes associated with seed pigmentation is important in understanding the process of accumulation of anthocyanin in different kinds of colored maize. However, there is still limited information on the seed pigmentation-related gene expression during seed development in maize. This study sequenced and assembled three maize inbred lines (one purple colored, CM22, and two white colored, CMS and CM19) to perform a comparative expression analysis between colored and colorless maize kernels. The profiling of comparative DEGs and the functional classification by the Gene Ontology (GO) between colored and colorless maize were obtained. This study’s results may then provide valuable information for future studies on colored and colorless maize breeding programs.

Methods

Plant materials
Three waxy maize inbred lines, CM22 (colored; purple), CMS, and CM19 (colorless; white), were grown in a field at the College of Agriculture and Life Science of Kangwon National University located in Chuncheon, Gangwon-do, Korea. Ears of each maize inbred line were covered before silk emergence to prevent contamination from immediate pollen. Because each maize inbred line had different tasseling and silking periods, kernels were generated by selfing in time for the flowering phase of each inbred line. At 10 days after pollination (DAP), the seeds of each inbred line were isolated and transferred into independent tubes before being frozen in liquid nitrogen and stored at –80°C until RNA extraction.

RNA isolation, cDNA library, and sequencing
The total RNA from three maize inbred lines (CM22, CMS, and CM19) was isolated and purified using the Hybrid-R kit (GeneAll Biotechnology Co., Seoul, Korea). The extracted total RNA were processed to quality control using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) for high throughput sequencing library. The total RNA of integrity number values with at least eight were used for the subsequent cDNA synthesis. The cDNA library construction and high-throughput mRNA sequencing using the Illumina sequencing platform HiSeq 2500 (Illumina, San Diego, CA, USA) were conducted at the National Instrumentation Center for Environmental Management (NICEM, Seoul National University, Seoul, Korea). The sequencing was generated with 150 bp × 2 of paired-end reads in accordance with the manufacturer’s protocol. The raw reads were filtered out with the trimming of adaptor nucleotides and low-quality nucleotides (reads containing more than 50% bases with Q-value ≤ 20) using trimmomatic [24]. The high-quality data were assembled by means of the Trinity (v2.4.0) de novo assembler software to check for newly expressed genes.

De novo assembly and DEG analysis
The new reference gene set was constructed by de novo assembly using Trinity assembler (v2.4.0) and finalized using Transdecoder (v.3.0.1) and CD-EST (v4.6) software. The gene was annotated by BlastX analysis using NCBI non-redundant protein DB. The DEG analysis between colored and colorless lines were analyzed by mapping to the gene set and comparing the expression level. Also, the data were aligned to the maize transcriptome reference B73 RefGen_v3 (https://www.maizegdb.org/) and were checked using RSEM software (https://deweylab.github.io/RSEM/) to comparing expression and reference derived from de novo assembly. The DEGs were calculated upon comparing the log2 fold change of the normalization of data of expressed reads using the Trimmed Mean of M values of the EdgeR v3.22.1 DEG analysis package software (https://bioconductor.org/packages/release/bioc/html/
The expressed genes between colored and colorless lines, which were significantly different statistically, were isolated using the false discovery rate < 0.05 value and log2 fold change value.

Functional and pathway enrichment analysis of DEGs
For functional annotation clustering, The Database for Annotation, Visualization, and Integrated Discovery (DAVID) Bioinformatics Resources v 6.8 software (https://david.ncifcrf.gov/) for the GO categories of cellular component (CC), biology process (BP), and molecular function (MF) was used. The statistically significant GO terms enrichment analysis for DEGs was conducted with thresholds of p < 0.01 and enrichment gene count > 2 in DAVID software. The enriched BINs of the DEGs were confirmed using PageMan analysis, which employs the ORA_FISHER test (cutoff value = 1.0) [25], whereas colored or colorless upregulated genes were assigned to metabolic pathways through the MapMan tool [26].

Results
Transcriptome sequencing and comparison to reference RNA sequence
To obtain the transcriptional profile in 10 DAP seeds from colored and colorless maize inbred lines, RNA-Seq for three samples, one colored maize inbred line (CM22) and two colorless maize inbred lines (CMS and CM19), was accomplished. For the 10 DAP seeds of each colorless and colored maize inbred line, a total of 75,257,086 (CM22) colored line, and 103,139,640 (CM5) and 66,978,958 (CM19) colorless line sequence reads were generated with read lengths of about 0.8, 1.0, and 0.7 billion bp, respectively. After filtering high-quality sequences with a Phred quality score of at least 20 and read length of more than 50 bp in the three maize inbred lines, 71.5% (CM22), 71.2% (CM5), and 71.8% (CM19) high-quality reads were gathered from the raw data (Table 1). For the mapping of each transcriptome in the mentioned colored and colorless maize inbred lines, a comparison was made with reference RNA sequences in a public DB (NCBI) which revealed mapped reads of 60%, 62%, and 62% in the expressed sequence tag (EST) references in CM22, CM5, and CM19, respectively (Table 1). The inbred lines revealed with different transcripts with at least 38% to the reference EST sequence.

De novo assembly and DEGs between colored and colorless maize inbred lines
This study assembled the high-quality reads generated from the CM22, CMS, and CM19 maize inbred lines, and a total of 63,870 (CM22), 82,496 (CM5), and 54,555 (CM19) contigs were collected with average lengths of 498, 457, and 518 bp, respectively (Table 2). In comparison with the GenBank data, any new genes among the contigs could not be revealed. To estimate the gene expression of unigenes in the three samples, about 50,000 genes (upwards of 50% of total reference genes) were expressed based on the RPKM (reads per kilobase of transcript per million reads mapped) criteria (over 0.25 value) in each of the CM22, CM5, and CM19 maize inbred lines (Table 3) among a total of 98,696 genes of the reference sequence in the NCBI.

From the DEG analysis, it was revealed that 7,044 genes were differentially expressed by at least two-fold, with 4,067 upregulated in colored maize inbred lines and 2,977 upregulated in colorless maize inbred lines (Fig. 1). This study identified more differentially expressed contigs in colored maize than in colorless maize at different fold change levels: 2–4 fold (3,112 up in colored and 912 up in colorless), 4–8 fold (646 up in colored and 1,003 up in colorless), 8–16 fold (172 up in colored and 514 up in colorless), and ≥16 fold (137 up in colored and 548 up in colorless), indicating changes in expression levels during seed development between colored and colorless maize inbred lines.

Functional enrichment analysis
A total of 45 GO functions were detected for colored and colorless

Table 1. Summary of RNA sequencing data, high quality and mapped reads to gene set in the three color and colorless maize inbred lines

<table>
<thead>
<tr>
<th>Type</th>
<th>Line</th>
<th>Raw sequence data derived from Illumina/HiSeq running</th>
<th>High-quality sequence</th>
<th>Mapped reads to the reference EST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Reads</td>
<td>Total length (bp)</td>
<td>Reads</td>
</tr>
<tr>
<td>Color</td>
<td>CM22</td>
<td>75,257,086</td>
<td>7,600,965,686</td>
<td>53,779,012</td>
</tr>
<tr>
<td>Color</td>
<td>CM5</td>
<td>103,139,640</td>
<td>10,417,103,640</td>
<td>73,476,942</td>
</tr>
<tr>
<td>Color</td>
<td>CM19</td>
<td>6,6978,958</td>
<td>6,764,874,758</td>
<td>48,122,664</td>
</tr>
</tbody>
</table>

EST, expressed sequence tag.
upregulated genes using DAVID software. The colored upregulated DEGs were enriched in 37 GO terms, while the colorless upregulated DEGs were enriched in 8 GO terms (Table 4). The genes for colored upregulated DEGs were related with enrichments of BP, such as in the anthocyanin-containing compound biosynthetic and metabolic process, carbohydrate metabolic process, cell wall macromolecule metabolic process, cellular carbohydrate metabolic process, cellular glucan metabolic process, cellular polysaccharide metabolic process, flavonoid biosynthetic and metabolic process, glucan metabolic process, photosynthesis-related process, and protein-chromophore linkage; CC, such as apoplast, chloroplast related functions, organelle subcompartment, photosynthetic membrane, photosystem related functions, plastid related functions, and thylakoid related functions; and MF, such as chlorophyll binding, glucosyltransferase activity, glycogen (starch) synthase activity, hydrolase activity related functions, and UDP-glucosyltransferase activity (Table 4).

Pathway analysis for colored and colorless maize inbred lines
To reveal specific genes between colored and colorless maize inbred lines, the transcriptional levels of mRNA in two combinations for three maize inbred lines (CM22/CM5 and CM22/CM19) were compared using MapMan software. Among 98,696 genes, a total of 404 genes were upregulated at least over two-fold in CM22 compared with CM5 or CM19 in colored maize inbred lines, whereas a total of 682 genes were upregulated at least over two-fold in CM5 and CM19 in colored maize inbred lines, whereas a total of 682 genes were upregulated at least over two-fold in CM5 and CM19 in colored maize inbred lines (Fig. 2).

The significant BINs for the representative functional pathways of DEGs were detected in the two combinations of colored and colorless maize inbred lines using Fisher exact test, with a cut-off value of 1.0 employing PageMan analysis (Fig. 3). The enriched BINs of the color upregulated genes were photosynthesis, major and minor CHO metabolism related functions, glycolysis, TCA/org. transformation, mitochondrial electron transport/ATP synthesis, cell wall, lipid metabolism related DEGs, N-metabolism,
Table 4. Functional enrichment clusters of color and colorless maize genes using the DAVID software

<table>
<thead>
<tr>
<th>Responses</th>
<th>No. of cluster</th>
<th>FES</th>
<th>Category</th>
<th>GO ID</th>
<th>Term</th>
<th>No. of genes</th>
<th>(%)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color-upregulated genes</td>
<td>1</td>
<td>6.53</td>
<td>BP</td>
<td>GO:0009765</td>
<td>Photosynthesis, light harvesting</td>
<td>12</td>
<td>0.99</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0018298</td>
<td>Protein-chromophore linkage</td>
<td>13</td>
<td>1.07</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0015979</td>
<td>Photosynthesis</td>
<td>25</td>
<td>2.05</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0019684</td>
<td>Photosynthesis, light reaction</td>
<td>13</td>
<td>1.07</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CC</td>
<td>GO:0009579</td>
<td>Thylakoid</td>
<td>29</td>
<td>2.38</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0009521</td>
<td>Photosystem</td>
<td>19</td>
<td>1.56</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0034357</td>
<td>Photosynthetic membrane</td>
<td>24</td>
<td>1.97</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0009522</td>
<td>Photosystem I</td>
<td>14</td>
<td>1.15</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0044436</td>
<td>Thylakoid part</td>
<td>24</td>
<td>1.97</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0009523</td>
<td>Photosystem II</td>
<td>16</td>
<td>1.31</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0042651</td>
<td>Thylakoid membrane</td>
<td>21</td>
<td>1.73</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0055035</td>
<td>Plastid thylakoid membrane</td>
<td>17</td>
<td>1.40</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0009535</td>
<td>Chloroplast thylakoid membrane</td>
<td>17</td>
<td>1.40</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0031976</td>
<td>Photosynthetic membrane</td>
<td>29</td>
<td>2.38</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0009522</td>
<td>Photosystem I</td>
<td>19</td>
<td>1.56</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0044434</td>
<td>Chloroplast part</td>
<td>23</td>
<td>1.89</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0044435</td>
<td>Plastid part</td>
<td>23</td>
<td>1.89</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0044448</td>
<td>Organelle subcompartment</td>
<td>17</td>
<td>1.40</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0026517</td>
<td>Chloroplast</td>
<td>32</td>
<td>2.63</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0009536</td>
<td>Plastid</td>
<td>32</td>
<td>2.63</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MF</td>
<td>GO:0016168</td>
<td>Chlorophyll binding</td>
<td>12</td>
<td>0.99</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.86</td>
<td>BP</td>
<td>GO:00095975</td>
<td>Carbohydrate metabolic process</td>
<td>59</td>
<td>4.85</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0016798</td>
<td>Hydrolase activity, acting on glycosyl bonds</td>
<td>36</td>
<td>2.96</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0004553</td>
<td>Hydrolase activity, hydrolyzing 0-glycosyl</td>
<td>31</td>
<td>2.55</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0030414</td>
<td>Peptidase inhibitor activity</td>
<td>8</td>
<td>0.66</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0009812</td>
<td>Flavonoid metabolic process</td>
<td>6</td>
<td>0.49</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0009813</td>
<td>Flavonoid biosynthetic process</td>
<td>6</td>
<td>0.49</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0009718</td>
<td>Anthocyanin–containing compound biosynthetic</td>
<td>3</td>
<td>0.25</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0046283</td>
<td>Anthocyanin–containing compound metabolic</td>
<td>3</td>
<td>0.25</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3.37</td>
<td>BP</td>
<td>GO:0006073</td>
<td>Cellular glucan metabolic process</td>
<td>14</td>
<td>1.15</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0044042</td>
<td>Glucan metabolic process</td>
<td>14</td>
<td>1.15</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0044262</td>
<td>Cell wall carbohydrate metabolic process</td>
<td>18</td>
<td>1.48</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0044036</td>
<td>Cell wall macromolecule metabolic process</td>
<td>10</td>
<td>0.82</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0044264</td>
<td>Cellular polysaccharide metabolic process</td>
<td>14</td>
<td>1.15</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CC</td>
<td>GO:0048046</td>
<td>Apoplast</td>
<td>8</td>
<td>0.66</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0004373</td>
<td>Glycogen (starch) synthase activity</td>
<td>5</td>
<td>0.41</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0035251</td>
<td>UDP-glucosyltransferase activity</td>
<td>10</td>
<td>0.82</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0046527</td>
<td>Glucosyltransferase activity</td>
<td>10</td>
<td>0.82</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MF</td>
<td>GO:0004867</td>
<td>Serine-type endopeptidase inhibitor activity</td>
<td>7</td>
<td>0.87</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0004866</td>
<td>Endopeptidase inhibitor activity</td>
<td>8</td>
<td>0.99</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0030414</td>
<td>Peptidase inhibitor activity</td>
<td>8</td>
<td>0.99</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0061134</td>
<td>Peptidase regulator activity</td>
<td>8</td>
<td>0.99</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0061135</td>
<td>Endopeptidase regulator activity</td>
<td>8</td>
<td>0.99</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0004857</td>
<td>Enzyme inhibitor activity</td>
<td>12</td>
<td>1.49</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GO:0030234</td>
<td>Enzyme regulator activity</td>
<td>12</td>
<td>1.49</td>
<td>0.009</td>
</tr>
</tbody>
</table>

FES, functional enrichment score; BP, biological process; CC, cellular component; MF, molecular function.
secondary metabolism, hormone metabolism, tetrapyrrole synthesis, stress, redox regulation, nucleotide metabolism, miscellaneous related functions, RNA, DNA, protein related functions, signaling, cell, development, transport, and C4 photosynthesis-related DEGs (Fig. 3). In the colorless upregulated genes, there was an overexpression of enriched BINs such as major and minor CHO metabolisms, gluconeogenesis/glyoxylate cycle, lipid metabolism, amino acid metabolism related functions, metal handling, secondary metabolism, hormone metabolism related DEGs, co-factor and vitamin metabolism, tetrapyrrole synthesis, stress, miscellaneous related functions, RNA, protein, signaling, and transport-related DEGs.

To confirm expressions of DEGs associated with metabolic pathways, DEGs for each of the two combinations of the three maize inbred lines (based on over two-fold change of expression level for each combination) were assigned to the corresponding BINs using the MAPMAN software (Fig. 4). Among the genes in each of the two combinations (CM22/CM5 and CM22/CM19) of the maize inbred lines, many DEGs in the 10 DAP colored and colorless seeds showed combination-specific (CM22/CM5 or CM22/CM19) expressions in independent metabolic pathways. However, 37 common colored-specific, expressed DEGs in CM22 were detected compared with CM5 and CM19 in metabolic pathways. Although the same genes for each combination had different expression levels, color-specific genes were involved in metabolic pathways (Fig. 4). Meanwhile, 22 common, colorless-specific, expressed DEGs in CM5 and CM19 were confirmed compared with CM22 in metabolic pathways (Fig. 4).

**Discussion**

The comparative analysis of the given colored and colorless maize inbred lines (CM22, CM5, and CM19) was performed to evaluate the DEGs’ profiling between two colored/colorless maize combinations (CM22/CM5 and CM22/CM19). All of the expressed genes were confirmed upon mapping the 98,696 genes of the B73 reference sequences at the NCBI. This study found a number of DEGs by comparing the gene expression between the colored and colorless maize inbred lines. Among the 7,044 DEGs with two-fold elevation in two common maize inbred line combinations (CM22/CM5 and CM22/CM19), 4,067 were upregulated in the
Fig. 4. Transcription levels of two-fold differentially expressed genes included in metabolic pathways. (A) CM22/CM5. (B) CM22/CM19.
colored maize inbred line CM22, while 2,977 were upregulated in the colorless maize inbred lines CM5 and CM19 (Fig. 1).

Based on the GO functional enrichment analysis, the GO terms of upregulated genes for colored or colorless maize inbred lines were differently clustered between color- and colorless-specific, expressed genes. Colored upregulated genes were clustered in many GO terms (Table 4). Photosynthesis-related GO terms had relatively higher functional enrichment score (FES) value (6.53) than other clusters. In previous studies, immature seeds of cereal crop were photosynthetically active in pericarp, which is the outer layer of the seed [27]. In addition to photoreceptors, photosynthesis contributes to the formation of anthocyanin [28]. For example, light-dependent anthocyanin accumulation was significantly inhibited by treatment with a photosynthetic inhibitor in non-chlorophyllous maize leaf [29]. Thus, the differential expression pattern of photosynthesis-related genes in seed pericarp between colored and colorless lines could suggest a dissimilar photosynthesis ability, which may directly affect anthocyanin pigmentation. Moreover, pigmentation-related GO terms were only detected in the third cluster of color upregulated genes that had a 3.37 score for FES (Table 4) while color upregulated genes were clustered in response to wounding, endopeptidase inhibitor and regulator activity, enzyme inhibitor and regulator activity, peptidase inhibitor and regulator activity, and serine-type endopeptidase inhibitor activity (Table 4). These results confirmed that genes directly associated with seed pigmentation only upregulated in colored maize inbred lines (GO:0009812, GO:0009813, GO:0009718, and GO:0046283).

This study also compared gene expression levels for each maize combination, CM22/CM5 and CM22/CM19, with a total of 98,696 maize reference unigenes used for each combination. A total of 1,743 and 455 color-related genes were detected as being upregulated in the colored line CM22 in comparison with each colorless line, CM5 and CM19 (Fig. 2A). Moreover, a total of 926 and 663 colorless related genes were found as upregulated for colorless lines in comparison with the colored line (Fig. 2B). Based on this result, numerous genes upregulated in the colored or colorless line showed a maize inbred specific gene expression pattern. In addition, the CM22/CM19 (upregulated in 455 colored and 663 colorless specific genes) combination had less DEGs than the CM22/CM5 (upregulated in 1,743 colored and 926 colorless specific genes) combination that shared the same genes more because the CM22 inbred line was generated by backcross with CM19.

This study found many DEGs between colored and colorless maize inbred lines (Figs. 1 and 2). Among these DEGs, some colored and colorless upregulated genes with two-fold elevation in two maize combinations showed higher expression in common metabolic pathways and the same function using MapMan based on BIns, such as cell wall.cellulose synthesis//cell wall.modification//glycolysis.glyceraldehyde 3-phosphate dehydrogenase. However, many other genes upregulated in the colored and colorless maize were involved in independent pathways or different

### Table 5. The gene list and expression level involved in anthocyanin related pathway

<table>
<thead>
<tr>
<th>ID</th>
<th>Gene name</th>
<th>Description</th>
<th>CM22/CM5</th>
<th>CM22/CM19</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF041043</td>
<td>-</td>
<td>Putative dihydro flavonol 4-reductase</td>
<td>1.104</td>
<td>0.097</td>
</tr>
<tr>
<td>AM156908</td>
<td>myb42</td>
<td>Transcription factor MYB42 (myb42 gene)</td>
<td>2.467</td>
<td>2.503</td>
</tr>
<tr>
<td>AY103770</td>
<td>-</td>
<td>Cinnamoyl-CoA reductase</td>
<td>0.180</td>
<td>1.748</td>
</tr>
<tr>
<td>AY106481</td>
<td>-</td>
<td>Chalcone-flavanone isomerase</td>
<td>2.210</td>
<td>2.957</td>
</tr>
<tr>
<td>AY108508</td>
<td>chi1</td>
<td>Chalcone isomerase</td>
<td>-0.056</td>
<td>1.528</td>
</tr>
<tr>
<td>AY109395</td>
<td>chs</td>
<td>Chalcone synthase</td>
<td>5.471</td>
<td>7.769</td>
</tr>
<tr>
<td>AY135018</td>
<td>pl</td>
<td>Transcription factor, pl-bol3 allele</td>
<td>5.561</td>
<td>5.561</td>
</tr>
<tr>
<td>AY135019</td>
<td>pl</td>
<td>Transcription factor, pl-W22 allele</td>
<td>4.903</td>
<td>4.903</td>
</tr>
<tr>
<td>DQ335126</td>
<td>bz</td>
<td>UDPG-flavonoid 3-oxy glucosyl transferase, bronze-Bz1-McC allele</td>
<td>3.534</td>
<td>3.532</td>
</tr>
<tr>
<td>DT641997</td>
<td>-</td>
<td>Oxidoreductase, 2OG-Fe(II) oxygenase</td>
<td>1.777</td>
<td>2.458</td>
</tr>
<tr>
<td>EU955769</td>
<td>C1</td>
<td>Anthocyanin regulatory C1</td>
<td>1.889</td>
<td>2.439</td>
</tr>
<tr>
<td>EU966779</td>
<td>-</td>
<td>Chalcone-flavanone isomerase</td>
<td>5.343</td>
<td>4.404</td>
</tr>
<tr>
<td>EU973925</td>
<td>-</td>
<td>Leucoanthocyanid dioxgenase</td>
<td>2.362</td>
<td>1.945</td>
</tr>
<tr>
<td>M26227</td>
<td>Lc</td>
<td>Lc regulatory protein</td>
<td>8.076</td>
<td>6.596</td>
</tr>
<tr>
<td>U04434</td>
<td>fht</td>
<td>Flavanone 3-beta-hydroxylase</td>
<td>1.897</td>
<td>3.592</td>
</tr>
<tr>
<td>X15806</td>
<td>R-S</td>
<td>Maize anthocyanin regulatory R-S</td>
<td>8.264</td>
<td>7.380</td>
</tr>
<tr>
<td>X55314</td>
<td>a2</td>
<td>a2 gene</td>
<td>4.638</td>
<td>5.556</td>
</tr>
<tr>
<td>X57276</td>
<td>B-Peru</td>
<td>B-Peru gene for a regulatory protein</td>
<td>2.454</td>
<td>5.031</td>
</tr>
</tbody>
</table>
functions even in the same pathway (Figs. 3 and 4). These results indicate a difference in the gene function involved in the seed developmental mechanisms for colored and colorless maize that is in agreement with functional enrichment and PageMan analyses (Table 4, Fig. 3).

Seed pigmentation for purple, red, and black colored maize seeds generally occur in the accumulation of anthocyanins. Anthocyanin is synthesized via the flavonoid pathway [30]. Anthocyanin pigmentation in maize organs is controlled by at least 20 loci as structural genes or transcriptional factors [31]. The materials of this study are CM22 with purple color and CM5 and CM19 with white color. Therefore, it is predicted that CM22 had more anthocyanin content and more express gene-related anthocyanin than CM5 and CM19. This study detected the genes, only upregulated in color in-

Fig. 5. Transcription levels of total genes involved in secondary metabolism pathways in two combinations. (A) CM22/CM5. (B) CM22/CM19.
bred lines, for the anthocyanin synthesis pathways based on Map-
Man and gene description (Table 5, Fig. 5). As a result, a total of 18
genes were associated with anthocyanin synthesis in this study. Among these genes, 15 genes were upregulated in both combina-
tions, CM22/CM5 and CM22/CM19, whereas three genes (AF041043, AY103770, and AY108508) were upregulated in only one combination (Table 5). In general, the levels of anthocyanin content increased for color maize during maturation [32]. Moreover, anthocyanin is rapidly accumulated at the later development stages after 25 DAP [30]. The RNA in this study was extracted from the 10 DAP seeds of colored and colorless maize inbred lines. Although many genes upregulated for colored maize inbred lines in this study already have high expression levels, these genes may ex-
press themselves more prominently in later development stages.
In this study, the assembled transcriptome method was exploit-
ted to perform a comparative expression analysis between one col-
ored and two colorless maize inbred lines. At least a two-fold up-
regulation of 4,067 genes in colored maize inbred lines and 2,977
genes in colorless maize inbred lines were detected. Although
many DEGs were not aligned in a functional pathway by MapMan
software, the differences between colored and colorless maize in-
bred lines were identified in some of the DEGs involved in antho-
cyanin synthesis. Eventually, these comparative gene expression
results may provide valuable information for maize breeding pro-
grams.

ORCID

Kyu Jin Sa: https://orcid.org/0000-0002-8286-4851
Ilk-Young Choi: https://orcid.org/0000-0003-4168-0471
Ju Kyong Lee: https://orcid.org/0000-0002-2769-0799

Authors’ Contribution

Conceptualization: JKL, IYC. Data curation: IYC. Formal analysis:
KJS, JKL. Funding acquisition: KJS, JKL. Writing – original draft:
KJS, JKL. Writing – review & editing: IYC.

Conflicts of Interest

No potential conflict of interest relevant to this article was report-
ed.

Acknowledgments

This study was supported by the Cooperative Research Program
for Agriculture Science & Technology Development (project no.
P[013157012020 and P[0131572020), Rural Development Ad-
ministration (RDA), and the Golden Seed Project (No. 213009-
05-1-WT821, P[012650012017), Ministry of Agriculture, Food,
and Rural Affairs (MAFRA), Ministry of Oceans and Fisheries
(MOF), Korea Forest Service (KFS), Republic of Korea.

References

1. Iordan LH, Martura T, Alionate E, Lazar C. Protein, starch, amino
and fatty acids content in corn hybrids developed at NARDI
Fundulea in natural climatic conditions. Sci Pap Ser A Agron
2015;58:221-227.
compounds, carotenoids, anthocyanins, and antioxidant capacity
of colored maize (Zea mays L.) kernels. J Agric Food Chem
2012;60:1224-1231.
3. Ford RH. Inheritance of kernel color in corn: explanations and
4. Navarro A, Torres A, Fernandez-Aulis F, Pena C. Bioactive com-
pounds in pigmented maize. In: Corn: Production and Human
Health in Changing Climate (Amanullah, Fahad S, eds.). Lon-
Blueberry anthocyanins in health promotion: a metabolic over-
black waxy corn hybrid cultivar, ‘Miheugchal’ with good eating
anthocyanin-rich grain corn hybrid cultivar, ‘Sekso2’. In: Annual
Symposium of Korean Society of Breeding Science, 2016 Jun 29-
Jul 1, Cheongju, Korea. Suwon: The Korean Society of Breeding
black waxy hybrid corn, “Heuljinjuchal” with good eating qual-
9. Park KJ, Park JY, Seo YH, Ryu SH, Choi JK, Kim HY. Anthocy-
anin-rich purple waxy corn single cross hybrid ‘Cheongchunchal’.
10. Petroni K, Piliu R, Tonelli C. Anthocyanins in corn: a wealth of
11. Morohashi K, Casas MI, Falcone Ferreya ML, Falcone Ferreya
L, Mejia-Guerra MK, Pourcel L, et al. A genome-wide regulatory
framework identifies maize pericarp color1 controlled genes.
Plant Cell 2012;24:2745-2764.
of the maize B regulatory gene: different leader and promoter se-
grams.

https://doi.org/10.5808/GI.2020.18.3.e29


Computational analysis of SARS-CoV-2, SARS-CoV, and MERS-CoV genome using MEGA

Vipan Kumar Sohpal*

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

The novel coronavirus pandemic that has originated from China and spread throughout the world in three months. Genome of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) predecessor, severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) play an important role in understanding the concept of genetic variation. In this paper, the genomic data accessed from National Center for Biotechnology Information (NCBI) through Molecular Evolutionary Genetic Analysis (MEGA) for statistical analysis. Firstly, the Bayesian information criterion (BIC) and Akaike information criterion (AICc) are used to evaluate the best substitution pattern. Secondly, the maximum likelihood method used to estimate of transition/transversions (R) through Kimura-2, Tamura-3, Hasegawa-Kishino-Yano, and Tamura-Nei nucleotide substitutions model. Thirdly and finally nucleotide frequencies computed based on genomic data of NCBI. The results indicate that general times reversible model has the lowest BIC and AICc score 347,394 and 347,287, respectively. The transition/transversions bias for nucleotide substitutions models varies from 0.56 to 0.59 in MEGA output. The average nitrogenous bases frequency of U, C, A, and G are 31.74, 19.48, 28.04, and 20.74, respectively in percentages. Overall the genomic data analysis of SARS-CoV-2, SARS-CoV, and MERS-CoV highlights the close genetic relationship.

Keywords: Middle East respiratory syndrome, Molecular Evolutionary Genetic Analysis, National Center for Biotechnology Information, SARS-CoV, SARS-CoV-2

Introduction

Coronaviruses (CoVs) usually influence the respiratory tract of mammals that lead to mild to severe respiratory tract infections [1]. In the past two decades, two highly pathogenic human CoVs including severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), emerging from animal reservoirs, have led to global epidemics with high morbidity and mortality [2]. According to the World Health Organization (WHO), as of April 2020 CoV has had a total of 2,269,630 diagnosed cases causing 155,205 deaths, throughout the world [3]. SARS-CoV-2 has a similar incubation phase and a relatively lower fatality rate than SARS-CoV or MERS-CoV, but it is estimated that the reproductive number of SARS-CoV-2 is higher than that of SARS-CoV [4]. Whole-genome analysis revealed that SARS-CoV-2 used mutations and recombination as crucial strategies in different genomic regions to become a novel infectious agent [5].

Generally, the rates of nucleotide substitution of RNA viruses are faster and are this rapid evolution is mainly shaped by natural selection [6]. Numerous substitution models
are time-reversible and, the model does not care which sequence is the ancestor and which is the descendant so long as all other parameters are held constant. Generalized time reversible (GTR) is the most general neutral, independent, finite-sites, time-reversible model possible [7]. On other hand, there are 203 possible ways that the exchangeability parameters can be restricted to form sub-models of GTR, ranging from the JC69 and F81 models (where all exchangeability parameters are equal) to the SYM model and the full GTR model (where all exchangeability parameters are free) [8-10]. The Jukes-Cantor (JC or JC69) model assumes equal transition rates as well as equal equilibrium frequencies for all bases and it is the simplest sub-model of the GTR model [11]. Kimura 2 (K2) parameters model and three parameters model are conserved the strong/weak properties of nucleotides [12,13]. F81 and HKY (five parameters) models in which the substitution rate are corresponds to the equilibrium frequency of the target nucleotide [14]. Bayesian information criterion (BIC) and Akaike information criterion (AICc) statistical models is important tool in analysis biological data [15]. In addition to that there are several methods for estimating substitution rates from genome sequence data [16].

As an emerging virus, limited information is available to depict the genetic diversity and nucleotide substitution and rate. Hence the purpose of the present work is to assess the genomic relationship on the basis of statistical techniques between MERS-CoV, SARS-CoV, and SARS-CoV-2 with an objective to (1) maximize value of likelihood function of nucleotide substitution models, (2) transition/transversion bias and frequencies computation using maximum likelihood (ML) technique, (3) analyze the probability rate of substitution using ML. It is assumed MERS-CoV, SARS-CoV, and SARS-CoV-2 belong to same phylogeny due to respiratory syndrome, but the present manuscript on the basis of genomic data able to depict the biological relationship. The comparison of the genomic data with various substitutions techniques is presented to analyze the relationship.

**Methods**

The genomic data for substitution analysis of SARS-CoV-2 (NC_045512.2), MERS-CoV (NC_019843.3), and SARS-CoV (FJS88686.1) viruses were obtained from the National Center of Biotechnology Information (NCBI) using Molecular Evolutionary Genetic Analysis (MEGA) [17] bioinformatics tool. Filtration of NCBI database through general nucleotide collection used of Megablast to optimize highly similar sequences. Filter and Mask of Blast used for filtration of data: (a) Filter (low complexity region filter) and (b) Mask (Query masked on using to scan database). ML statistical method used to compute BIC score and AICc value of 24 different nucleotide substitution models. Mathematically BIC is function of \( f(n, k, L) \), AIC \( f(k, L) \), AICc \( f(AIC, k, n) \) is as mentioned in Eqs. (1), (2), and (3).

\[
BIC = \ln(n)k - 21n(L) \quad (1)
\]
\[
AIC = 2k - 21n(L) \quad (2)
\]
\[
AICc = AIC + \frac{2k^2 + 2k}{n - k - 1} \quad (3)
\]

\( L \) = the maximized value of likelihood function of model M, \( n \) = number of data point, \( K \) = number of parameters estimated by model. Frequencies and transition/transversion bias 24 different nucleotide substitution models also evaluated. Simulate the biological data to estimate the probability rate of substitution (\( r \)) using ML method for different nucleotide substitution models. Similarly, database used to assess the nucleotide base frequencies for each sequence as well as an overall average to assess the extent of relation.

**Results and Discussion**

**ML of different nucleotide substitution models**

BIC and AICc are the most important parameters for statistical analysis of ML to analyze the biological data. Both the BIC and AICc used to evaluate the best model among a finite set of models with penalty parameters. BIC based, on the likelihood function and AICc estimator of out-of-sample prediction error.

GTR model have lowest BIC and AICc score 347,395, 347,288 computed using MEGA with \( K = 11 \) shown in Fig. 1. In addition, rate of variation across sites (+G), the GTR + G model show BIC and AIC score slightly increase with respect to GTR. On further addition, a proportion of invariable sites (+I) and/or rate of variation across sites (+G), GTR + G + I model indicates 0.0072% elevation in BIC score and 0.00144% go up in AIC (\( K = 13 \)). HKY model (\( K = 7 \)) having lowest value for BIC 347473, AICc 347405, but higher than most appropriate GTR model. Similarly HKY + I + G model (\( K = 9 \)) simulated result shows the score get higher with respect to base model. Both the model JC + G + I and K2 + I (\( K = 5 \)) boast BIC and AICc criterion score highest. The deviation between GTR and K2+I models is for BIC, AICc scores 1.49% and 1.50%, respectively.

It indicates ML method accurately fits of 24 different nucleotide substitution models for biological data of SARS-CoV-2, MERS-CoV, and SARS-CoV under neutral evolution. As per information theory, lowest BIC score preferred due to Bayesian probability and
inference, while highest score criteria opted for AICc based on frequentist-based inference. Simulative investigation results reveal that differences between lowest and highest scores are around 1.5%, virtue of that SARS-CoV-2, MERS-CoV, and SARS-CoV data best fitted through GTR model. The corrected AIC model gives better results as compare to AIC value as correlated in Eq. (3).

Nucleotide frequencies ($f$) and rates of base substitutions rate ($r$) are also key factor to justify best nucleotide substitution model using ML technique. The nucleotide frequencies predicted for GTR model are $A = 0.28$, $U = 0.317$, $C = 0.195$ and $G = 0.207$ of biological data of SARS-CoV-2, SARS-CoV, and MERS-CoV. The frequencies of nitrogenous base remain constant for first 12 models from GTR, GTR + G, GTR + G + I, HKY, TN93, HKY + G, TN93 + G, TN93 + G + I, HKY + G + I, GTR + I, HKY + I to TN93 + I. The nucleotide frequencies for T92, T92 + G, T92 + G + I and T92 + I models are ($A = 0.299$, $U = 0.299$, $C = 0.201$, $G = 0.201$) remain steady, but varied from prior methods. JC, JC + G, K2, K2 + G, K2 + G + I, JC + I, JC + G + I, and K2 + I models replicated the same frequency at the rate 0.25 for all nitrogenous base as revealed in Fig 2. Base substitution rates are also dependent on nucleotide substitutions models, in GTR model $r$(AU), $r$(UA), $r$(CA), $r$(GA) substitutions are dominated. Fig. 3 replicate the min and max rate of rates of base substitutions irrespective of models are as follow, $r$(AU $0.077$, $0.122$), $r$(AC $0.05$, $0.084$), $r$(AG $0.77$, $0.107$), $r$(UA $0.074$, $0.115$), $r$(UC $0.06$, $0.101$), $r$(UG $0$, $0.086$), $r$(CA $0.073$, $0.126$), $r$(CU $0.079$, $0.119$), $r$(CG $0.05$, $0.124$), $r$(GA $0.079$, $0.132$), $r$(GU $0$, $0.099$), and $r$(GC $0.05$, $0.086$).

It has been observed that under the model of uniform substitution among site REV, TN93 HKY, the frequency parameters are free to exchangeability, while JC and K2 models have frequencies at uniform rate 1/4. Virtue of these statistical parameters, the models GTR, GTR + G, GTR + G + I, HKY, TN93, HKY + G, TN93 + G, TN93 + G + I, HKY + G + I, GTR + I, HKY + I, and TN93 + I shows similar results in term nucleotide frequencies. JC and K2 model rely on different frequency parameter, due to that JC, JC + G, K2, K2 + G, K2 + G + I, JC + I, JC + G + I, and K2 + I models replicate the result same mode. The estimates of transitional and transversional of substitution rates are of 1st + 2nd + 3rd position data using simulation of data. Fig. 3 confirms that the number of transversional are larger than the number of transitions. In broad, the transitional/transversional varies from 0.57 (GTR model) to 0.89 (T92 + G + I), higher values indicate proportion of invariable sites (+I) and/or rate of variation across sites (+G) are more dominating in T92 model for SARS-CoV-2, SARS-CoV, and MERS-
Fig. 2. Frequencies of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), Middle East respiratory syndrome coronavirus, SARS-CoV genome using nucleotide substitution with maximum likelihood approach.

Fig. 3. Substitution rate for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), Middle East respiratory syndrome coronavirus, SARS-CoV genome using maximum likelihood for different nucleotide substitution models.
ML to estimate of substitution matrix and transition/transversion bias

Probability rate of substitution (R) using ML depends upon the base frequency parameters and nucleotide substitution models. Base frequency parameters $\Pi_A = \Pi_C = \Pi_T = \Pi_U = 1/4$ for JC and K2 models and for GTR, HKY, TN93, T3 models have all $\Pi_i$ free to exchange. Six different nucleotide substitution models were simulated for biological sequence data of SARS-CoV, MERS-CoV and SARS-CoV-2.

JC substitution model shows the transitional and transversionsal substitutions rate 8.33, while transitional substitutions for all base are 9.32 and transversionsal substitutions is equal to 7.84 for K2 parameter model. In general, HKY, TN93 models having transitional substitutions are more dominating in C-U and transitional substitution G-U and A-U. GTR and T3 parameter models resultant of higher transition substitution for A-G, 11.24 and 12.13, respectively. The lowest value of transition in GTR and T3 models also lies for same base (C-U). The highest probabilities of transversionsal substitutions (A-U) are the models are 9.93 and 12.05 as shown in Table 1. In all models except than JC and K2, the lowest transitional substitutions observed C-U base pair. Overall transitional substitutions have higher hand as compared transversionsal substitutions in all models.

The estimated transition/transversion bias is 0.59 for K2-parameter model with codon positions included 1st + 2nd + 3rd + Noncoding that is not translated into a protein. There are a total of 43,053 positions in the final dataset. The transition/transversion bias for T93 and GTR equal to 0.56, while HKY and T3 parameter have transition/transversion bias is 0.57 as revealed in Fig. 4. The variations in the entire model are from 0.56 to 0.59, and overall consistent value for transition/transversion bias.

JC and K2 models belong to one class of base frequency parameters, virtue of that JC model demonstrates equal rate of transition/transversion bias. K2 model shows constant rate of transition 9.32 and transversionsal 7.84 substitution biases. On the other hand, T93, T3, HKY, and GTR model exchangeability are free, due to that transitional and transversionsal substitutions rate are different. Transition/transversion bias is approximately 0.5 when that indicates no bias towards either transitional or transversionsal substitution because two kinds of substitution are equally probable, there are twice as many possible transversions as transitions.

SARS-CoV-2 nucleobase has higher frequency of T as compared to SARS-CoV, and approximately equal to MERS-CoV. Cytosine frequency of SARS-CoV-2 is less than both the biological sequences of SARS-CoV and MERS-CoV as shown in Fig 5. The variation

### Table 1. Probability rate of substitution (R) using maximum likelihood statistical method

<table>
<thead>
<tr>
<th>Substitution rate</th>
<th>A</th>
<th>T/U</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juke Cantor Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>8.33</td>
<td>8.33</td>
<td>8.33</td>
</tr>
<tr>
<td>T/U</td>
<td>8.33</td>
<td>-</td>
<td>8.33</td>
<td>8.33</td>
</tr>
<tr>
<td>C</td>
<td>8.33</td>
<td>8.33</td>
<td>-</td>
<td>8.33</td>
</tr>
<tr>
<td>G</td>
<td>8.33</td>
<td>8.33</td>
<td>8.33</td>
<td>-</td>
</tr>
<tr>
<td>Tamura Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>9.45</td>
<td>6.36</td>
<td>7.4</td>
</tr>
<tr>
<td>T/U</td>
<td>9.45</td>
<td>-</td>
<td>7.4</td>
<td>6.36</td>
</tr>
<tr>
<td>C</td>
<td>9.45</td>
<td>11</td>
<td>-</td>
<td>6.36</td>
</tr>
<tr>
<td>G</td>
<td>11</td>
<td>9.45</td>
<td>-</td>
<td>6.36</td>
</tr>
<tr>
<td>Tamura-Nei Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>9.33</td>
<td>6.1</td>
<td>8.31</td>
</tr>
<tr>
<td>T/U</td>
<td>8.78</td>
<td>-</td>
<td>6.79</td>
<td>6.49</td>
</tr>
<tr>
<td>C</td>
<td>8.78</td>
<td>11.06</td>
<td>-</td>
<td>6.49</td>
</tr>
<tr>
<td>G</td>
<td>11.24</td>
<td>9.93</td>
<td>-</td>
<td>6.49</td>
</tr>
<tr>
<td>Kimura-2 Parameter Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>7.84</td>
<td>7.84</td>
<td>9.32</td>
</tr>
<tr>
<td>T/U</td>
<td>7.84</td>
<td>-</td>
<td>9.32</td>
<td>7.84</td>
</tr>
<tr>
<td>C</td>
<td>7.84</td>
<td>9.32</td>
<td>-</td>
<td>7.84</td>
</tr>
<tr>
<td>G</td>
<td>9.32</td>
<td>7.84</td>
<td>7.84</td>
<td>-</td>
</tr>
<tr>
<td>General Time Reversible Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>12.05</td>
<td>8.25</td>
<td>8.97</td>
</tr>
<tr>
<td>T/U</td>
<td>10.64</td>
<td>-</td>
<td>6</td>
<td>3.64</td>
</tr>
<tr>
<td>C</td>
<td>11.87</td>
<td>9.78</td>
<td>-</td>
<td>5.72</td>
</tr>
<tr>
<td>G</td>
<td>12.13</td>
<td>5.57</td>
<td>5.38</td>
<td>-</td>
</tr>
</tbody>
</table>

Transitional substitutions are shown in bold and transversionsal substitutions are shown in italics.

![Fig. 4. Transition/transversion bias for different nucleotide substitution models.](https://doi.org/10.5808/GI.2020.18.3.e30)
in cytosine base is around 9.6% with respect to SARS-CoV. The adenine nucleobase frequency is 29.896 of SARS-CoV-2 much higher than MERS-CoV and 5.75% modified from SARS-CoV. On the other hand, guanine frequency for current SARS-CoV-2 is much lesser than both the SARS-CoV and MERS-CoV. The average frequency of SARS-CoV-2, SARS-CoV, and MERS-CoV for U, C, A, and G are 31.74012, 19.48521, 28.04331, and 20.73135, respectively. Close value of nucleobase frequency (SARS-CoV-2, MERS-CoV, and SARS-CoV) reflects that SARS-CoV-2 is modified from previous respiratory syndrome virus.

**Conclusions**

On the basis of BIC and AICc score, it concluded that GTR model is more accurate for genome analysis of SARS-CoV & MERS-CoV and CoV-2 under non-uniform rates of evolution and invariable (+I). 0.03% difference found in BIC and AIC score for GTR model at penalty parameter of 11 signified that SARS-CoV-2 is closely to SARS-CoV and MERS-CoV both virus strains. The base frequency all 24-substitution model except JC and K2 are same, due to free exchangeability, resultant of that JC and K2 parameter observations trends are different from other substitution models. The results also indicate the close proximity of SARS-CoV-2 to SARS-CoV and MERS-CoV probability rate of substitution confirmed transitional substitutions are more dominant in all genomic sequences (NC_045512.2, NC_019843.3, and FJ588686.1) because two out of three single nucleotide polymorphisms are transitions retain in SARS-CoV, MERS-CoV, and SARS-CoV-2. Low frequency of nucleotide (0–0.35) and substitution rate (0–0.18) in all nucleotide substitution models support the result of closeness among the virus strain. 1st + 2nd + 3rd + noncoding simulated result for transition/transversion bias reflected the positive evolution that indicates towards of nonsynonymous substitutions. The outcome of A-T (62.14%) and G-C (37.86%) nucleobase frequencies for SARS-CoV-2 evidence that variation in genome with respect SARS-CoV & MERS-CoV. The G-C frequencies are 5.86% elevated in SARS-CoV-2 & MERS-CoV and A-T frequencies are 5.86% upward for SARS-CoV-2. Closer the nucleobase frequency also supports and affirms SARS-CoV-2 is closer resemblance of SARS-CoV and MERS-CoV.

**ORCID**

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

**Conflicts of Interest**

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

**References**


Confidence intervals for the COVID-19 neutralizing antibody retention rate in the Korean population

Catherine Apio¹, Md. Kamruzzaman², Taesung Park¹,²*

¹Interdisciplinary Program in Bioinformatics, Department of Statistics, Seoul National University, Seoul 08826, Korea
²Department of Statistics, Seoul National University, Seoul 08826, Korea

The coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has become a global pandemic. No specific therapeutic agents or vaccines for COVID-19 are available, though several antiviral drugs, are under investigation as treatment agents for COVID-19. The use of convalescent plasma transfusion that contain neutralizing antibodies for COVID-19 has become the major focus. This requires mass screening of populations for these antibodies. While several countries started reporting population based antibody rate, its simple point estimate may be misinterpreted without proper estimation of standard error and confidence intervals. In this paper, we review the importance of antibody studies and present the 95% confidence intervals COVID-19 antibody rate for the Korean population using two recently performed antibody tests in Korea. Due to the sparsity of data, the estimation of confidence interval is a big challenge. Thus, we consider several confidence intervals using Asymptotic, Exact and Bayesian estimation methods. In this article, we found that the Wald method gives the narrowest interval among all Asymptotic methods whereas mid p-value gives the narrowest among all Exact methods and Jeffrey's method gives the narrowest from Bayesian method. The most conservative 95% confidence interval estimation shows that as of 00:00 on September 15, 2020, at least 32,602 people were infected but not confirmed in Korea.

Keywords: antibody, confidence interval, COVID-19, retention rate

Introduction

Several unexplained pneumonia cases have been successively discovered in Wuhan, Hubei province of China, in early December 2019, which have been confirmed to be severe acute respiratory syndrome caused by a novel coronavirus 2 (severe acute respiratory syndrome coronavirus 2, SARS-CoV-2) and has spread rapidly across the globe [1]. The spread of coronavirus disease 2019 (COVID-19) became a global threat and the World Health Organization (WHO) declared COVID-19 a global pandemic on March 11, 2020 [2]. As of September 18, 2020, there were a total of 30,358,087 confirmed cases and 950,635 deaths from COVID-19 worldwide [3].

No specific therapeutic agents or vaccines for COVID-19 are available yet [4]. Accurate and fast diagnosis of the SARS-CoV-2 is one of the most important methods used to isolate patients with COVID-19 and stop the spread of the epidemic, as well as save people’s lives [5]. The current recommendations for laboratory diagnosis of COVID-19 from the Center for Disease Control (CDC) are that clinicians coordinate this testing with local public health authorities and/or the CDC. Viral nucleic acid detection using real-time
reverse-transcription polymerase chain reaction assay, like that developed for the diagnosis of SARS-CoV, was developed and used for rapid detection of SARS-CoV-2 and remains the gold standard for diagnosis of COVID-19 [6].

Although several antiviral drugs, such as the nucleotide analogue remdesivir and favipiravir, are under investigation as treatment agents for COVID-19, the antiviral efficacy of these drugs is not yet known [4,7]. In addition to vaccine development and approaches that directly target the virus or block viral entry, treatments that address the immunopathology of the infection have become a major focus. The use of convalescent plasma was recommended as an empirical treatment during outbreaks of Ebola virus in 2014 and the treatment of Middle East respiratory syndrome coronavirus in 2015 [8,9]. This approach was effective with other viral infections such as SARS-CoV, H5N1 avian influenza, and H1N1 influenza [10], through a single convalescent plasma transfusion [11-13]. Treatment of severe infections of patients of influenza A(H1N1) 2009 pandemic with convalescent plasma was associated with reduced respiratory tract viral load, serum cytokine response, and mortality [10]. Also, the study involving 80 patients with SARS showed that administration of convalescent plasma was associated with a higher rate of hospital for patients who received the convalescent plasma than with patients who did not receive the plasma [10,14].

In addition, high-throughput platforms, such as the large-scale single-cell RNA sequencing of B cells (enriched for B cells that produce antibodies directed at the SARS-CoV-2 spike glycoprotein) from patients who are convalescent, have allowed the identification of SARS-CoV-2-specific neutralizing antibodies [6]. Neutralizing antibodies is said to play an important role in virus clearance and have been considered as a key immune product for protection or treatment against viral diseases. Convalescent plasma containing identified SARS-CoV-2-specific neutralizing antibodies has already been used to treat a small number of patients with severe disease, and preliminary results show clinical improvement in 5 of 5 critically ill patients with COVID-19 who had developed acute respiratory distress syndrome [6,10]. All these findings raise the hypothesis that use of convalescent plasma transfusion is beneficial in patients infected with SARS-CoV-2 and solidifies the importance of large-scale antibody testing for COVID-19 [15].

A COVID-19 antibody test, also known as a serology test, is a blood test that can detect if a person has antibodies to SARS-CoV-2. An antibody test checks for antibodies (proteins made by the immune system to fight infections like viruses and may help to ward off future occurrences by those same infections) in the blood. Human bodies make antibodies when we catch an infection to help fight the infection [16,17]. If coronavirus antibody is present in the blood, it's likely he/she had the virus before. The value of antibody tests is currently limited to (1) answering the question of whether someone has had the virus before, (2) providing data and a greater understanding on the spread of the virus [18]. Also, given the unknown scale of asymptomatic infections (infected patients without symptoms), there is a pressing need for serological diagnosis to represent the real number of COVID-19 infected patients which determines the true extent of infection in a given country [19]. Serological tests are known to be in use in Europe, United States, Japan and other developed countries to figure out how many people are infected with the potentially deadly virus [20]. For example, results from Spain's final stage of a nationwide antibody study shows that Spain's antibody retention rate is believed to be 5% [21] while London, Stockholm, and Tokyo have a retention rate of 17%, 7.3%, and 0.1%, respectively [22]. While viral RNA-based testing for acute infection is the current standard, surveying antibody protection is a necessary for discovering the real extend of coronavirus infections in a population and for return to social normality [19,20].

Other importance for COVID-19 antibody testing include; to identify donors with high-neutralizing titers for convalescent plasma for therapy and define correlates of protection from SARS-CoV-2 [19].

**Methods**

**Materials**

Recently, Korea Centers for Disease Control and Prevention (KCDC) announced the discovery of neutralizing antibodies for COVID-19 from two investigative screening surveys carried in South Korea. The first antibody screening results reported; 0 neutralizing antibodies were discovered out of 1,555 serum samples collected for antibody titer from subjects who participated in the Korea National Health and Nutrition Examination Survey (KNHANES) from April 21 to June 16 from 14 cities and provinces in South Korea excluding Daegu (which was the city for the major COVID-19 outbreak in Korea in early March), Daejeon and Sejong (0/1,555) and social normality [19,20].

The above reported dataset only captures a sample proportion
but does not provide its confidence interval. This kind of result reporting can mislead the public especially to people without any statistical knowledge. Based on these sample datasets, we will use inferential statistics for deciding for whole population in Korea about COVID-19 antibody screening studies. Because our sample data is sparse, point estimation (i.e., sample proportion) gives some misleading interpretations, for example; 0/1,555 sample proportion shows that there are no neutralizing antibodies present in the Korean population, and so it is good to report point estimation along with proper interval estimation. In paper, we present intervals (95% confidence intervals) for the above reported Korean COVID-19 neutralizing screening antibody results using known Asymptotic, Exact, and Bayesian estimation methods.

Methods
As one method doesn’t give the optimal confidence interval range, we present confidence intervals calculated using different methods that can apply the point estimate results above. There are many methods available for calculation of confidence intervals for various parameters and those methods are mainly divided into three different type of estimation techniques, such as asymptotic estimation, exact estimation, and Bayesian estimation. In this section, we have reviewed the likelihood function for binomial proportion as well as the methods for confidence interval.

Likelihood function for binomial parameter
Let we conduct $n$ iid Bernoulli experiments with probability of success $\pi$ and find $y$ successes. Then the likelihood function can be defined as

$$L(\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y},$$

for $y = 0, \ldots, n$

and the log-likelihood function is

$$l(\pi) = \ln(\binom{n}{y}) + y \ln(\pi) + (n-y) \ln(1-\pi)$$

Maximum likelihood estimator of $\pi$ is $\hat{\pi} = y/n$, the sample proportion of success for $n$ trials, and the standard error of $\pi$ is $\hat{\sigma}(1-\hat{\pi})/\sqrt{n}$.

Asymptotic estimation (large sample approximation)
Confidence intervals can be obtained by inverting the association test [23]. For instance, a 95% confidence interval for the population proportion $\pi$ is the set of $\pi_0$ for which test of $H_0: \pi = \pi_0$ has p-value exceeding 0.05 [23]. Wald, Score, and Likelihood-ratio are the three main asymptotic methods for estimating confidence intervals as described below.

Wald confidence interval
The 100(1-$\alpha$)%wald confidence interval is

$$\left\{ \pi : \frac{\hat{\pi} - \pi}{SE} < z_{\alpha/2} \right\} \approx \left( \hat{\pi} - z_{\alpha/2} (SE), \hat{\pi} + z_{\alpha/2} (SE) \right),$$

where $z_{\alpha/2}$ is the $z$-score form the standard normal distribution with right tailed probability $\alpha/2$.

Likelihood-ratio-based confidence interval (LR)
The 100(1-$\alpha$)% likelihood-ratio-based confidence interval is

$$\left\{ \pi : 2 \left[ L(\hat{\pi}) - L(\pi) \right] < X^2_{1}(a) \right\},$$

where $L(\pi)$ is the maximized value of likelihood function under $H_0: \pi = \pi_0$ and $L(\hat{\pi})$ is the likelihood function calculate at the ML estimate $\hat{\pi}$, and $x^2(a)$ is the 100(1-$a$) percentile of the chi-square distribution with 1 degree of freedom.

Score-based confidence interval
The 100(1-$\alpha$)% score-based confidence interval is

$$\left\{ \pi : \frac{u(\pi)}{I(\pi_0)^{1/2}} < z_{\alpha/2} \right\},$$

where $u(\pi)$ is the score function derived from the log-likelihood function.

Exact estimation
For both small and large to moderate samples, population proportion inference can occur both near 0 or 1, and both are not good. In such cases, asymptotic methods may have inadequate performance and provide quite different confidence intervals [23]. Therefore, we use alternative estimation techniques, such as exact sample inference and Bayesian sample inference. For exact estimation, we use the Clopper-Pearson and the mid p-value methods. Clopper-Pearson interval is based on inverting the tailed binomial tests for forming confidence intervals [24]. For a binomial data with parameter $\pi$ (success), the endpoints are the solution of

$$\sum_{k=0}^{y} \binom{n}{k} \pi^k (1-\pi)^{n-k} = \frac{a}{2}, \quad \sum_{k=0}^{y} \binom{n}{k} \pi^k (1-\pi)^{n-k} = \frac{a}{2},$$

and the lower bound is 0 when $y=0$ and the upper bound is 1 when $y=n$. Unfortunately, with the discrete probability distributions, it is usually not possible for a p-value to achieve the desired significance level exactly [24], so we use the mid p-value

Mid p-value:
For small samples of discrete data, it seems sensible to use adjustment of exact methods based on the mid p-value \cite{25}. For a test statistic $T$ with observed value $t_0$ and one-sided alternative hypothesis, the mid p-value is obtained by

$$
\text{Mid } p - \text{value} = \frac{1}{2} p(T = t_0) + p(T > t_0),
$$

where probability, $p$ calculated from the null distribution. The two-sided mid p-value is

$$
\text{Mid } p - \text{value} = 2\left[ \frac{1}{2} p(T = |t_0|) + p(T > |t_0|) \right].
$$

**Bayesian estimation**

Aforementioned (asymptotic estimation and exact estimation) approaches are known as the frequentist approach and requires a random process that produces the observed data \cite{26}. That is, the parameter value is assumed to be unknown, but a fixed quantity and obtained from the observed data. Recent years have seen with increasing popularity of this Bayesian approach, which considers the parameter is a random quantity and whose value can be described by a probability distribution, known as prior distribution and fixed data. Bayesian approach combines the prior with observed data to create a posterior for the parameters using the Bayes equation;

$$
p(\pi|y) = \frac{p(y|\pi)p(\pi)}{p(y)},
$$

where $\pi$ is the unknown parameter, $y$ is the observed data and the denominator $p(y)$ is the marginal probability function of the data, which is a constant with respect to $\pi$. The Bayes equation then simplifies to

$$
p(\pi|y) \propto p(y|\pi)p(\pi),
$$

The above equation gives the posterior probability of $\pi$, $p(\pi|y)$, as a function of likelihood $p(y|\pi)$ and prior $p(\pi)$. Therefore, we need to choose the prior information, which is the most difficult aspect in Bayesian approach. If there is lack of prior information one can use uniform prior, which can be got from literature if a pilot study has been conducted, which turns out to be a non-informative prior. The most popular non-informative choice of prior is Jeffrey’s prior, defined as

$$
\left[-E\left(\frac{\partial^2 \ln f(y|\pi)}{\partial \pi^2}\right)\right]^{1/2}.
$$

In binomial setting, the Jeffrey’s prior for binomial data is $\pi \sim \text{Beta} \left(\frac{1}{2}, \frac{1}{2}\right)$. Alternatively, when there is no prior information $\text{Beta}(1,1)$ prior known as uniform prior $\text{U}(0,1)$ can be considered.

Because the parameter $\pi$ is a random variable in Bayesian techniques, it allows for making ideal statements concerning the probability of the parameter and confidence intervals. This confidence interval contains most of the posterior distribution and is known as the posterior interval or credible interval.

**Results**

Table 1 presents the 95% confidence intervals for the first antibody results using only the KNHANES samples and for the total population in South Korea. The first two columns show the methods and the next two columns show the 95% confidence interval for anti-

### Table 1. 95% Confidence intervals (CIs) for the first antibody results using the KNHANES samples only and for the total population in South Korea

<table>
<thead>
<tr>
<th>Method</th>
<th>Sample 95% CI (antibody retention rate)</th>
<th>Total population 95% CI (antibody carriers)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>Asymptotic estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald*</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>Score</td>
<td>0.000000</td>
<td>0.000246</td>
</tr>
<tr>
<td>Likelihood ratio</td>
<td>0.00011</td>
<td>0.00123</td>
</tr>
<tr>
<td>Exact estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exact</td>
<td>0.000000</td>
<td>0.000236</td>
</tr>
<tr>
<td>MidP</td>
<td>0.000000</td>
<td>0.000193</td>
</tr>
<tr>
<td>Bayesian estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td>0.000000</td>
<td>0.000237</td>
</tr>
<tr>
<td>Jeffrey’s</td>
<td>0.000000</td>
<td>0.000161</td>
</tr>
</tbody>
</table>

KNHANES, Korea National Health and Nutrition Examination Survey.

\*Fails to provide CI.
body retention rate in the samples. The final two columns represent the estimated 95% confidence interval of antibody carriers in Korean population by multiplying the total number of Korean population (51,780,579 people) with the antibody ratio (the proportion of samples with neutralizing antibodies provided as confidence intervals) from September 19, 2020. Note that this estimation was derived from a simple random sampling assumption, while the antibody sample does not represent the total Korean population.

From Table 1, the Wald method fails to provide confidence interval. LR gives the minimum upper bound which is 63,690 and it also provides narrower confidence intervals among all types of confidence intervals’ methods. Score, Exact and Uniform methods gives similar interval.

Table 2 presents the 95% confidence intervals for the first antibody results using only the Seoul samples and for the total population in South Korea. Table 3 presents the 95% confidence intervals for the first antibody results using both Seoul and KNHANES samples (1,555 + 1,500) and the total population in South Korea. Table 4 presents the 95% confidence intervals for the second antibody results using only the KNHANES samples (1,440) and for the total population in South Korea. Table 5 presents the 95% confidence intervals for the sum of the first and second antibody results using the KNHANES + Seoul + KNHANES samples and for the total population in South Korea. The first two columns show the methods and the next two columns the 95% confidence interval antibody ratio in the total samples (1,555 + 1,500 + 1,440).

For all cases, the Wald method gives the narrowest interval and smallest upper bound values, except in Table 1 (102,008, 50,227, 104,079, and 54,887, respectively) among all the asymptotic estimation methods. mid p-value method among exact estimation and Jeffrey’s method among Bayesian estimation, all give the smallest interval. Jeffrey’s and mid p-value methods have similar intervals.

Table 2. 95% Confidence intervals (CIs) for the first antibody results using only samples from Seoul area and for the total population in South Korea

<table>
<thead>
<tr>
<th>Method</th>
<th>Sample 95% CI (antibody retention rate)</th>
<th>Total population 95% CI (antibody carriers)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>Asymptotic estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td>0.00000</td>
<td>0.00197</td>
</tr>
<tr>
<td>Score</td>
<td>0.00012</td>
<td>0.00377</td>
</tr>
<tr>
<td>Likelihood ratio</td>
<td>0.00005</td>
<td>0.00292</td>
</tr>
<tr>
<td>Exact estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exact</td>
<td>0.00009</td>
<td>0.00472</td>
</tr>
<tr>
<td>MidP</td>
<td>0.00010</td>
<td>0.00310</td>
</tr>
<tr>
<td>Bayesian estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td>0.00016</td>
<td>0.00371</td>
</tr>
<tr>
<td>Jeffrey’s</td>
<td>0.00000</td>
<td>0.00311</td>
</tr>
</tbody>
</table>

Table 3. 95% Confidence intervals (CIs) for the first antibody results using both KNHANES + Seoul sample size only and for the total population in South Korea

<table>
<thead>
<tr>
<th>Method</th>
<th>Sample 95% CI (antibody retention rate)</th>
<th>Total population 95% CI (antibody carriers)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>Asymptotic estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td>0.00000</td>
<td>0.00097</td>
</tr>
<tr>
<td>Score</td>
<td>0.00000</td>
<td>0.00185</td>
</tr>
<tr>
<td>Likelihood ratio</td>
<td>0.00000</td>
<td>0.00144</td>
</tr>
<tr>
<td>Exact estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exact</td>
<td>0.00000</td>
<td>0.00232</td>
</tr>
<tr>
<td>MidP</td>
<td>0.00010</td>
<td>0.00160</td>
</tr>
<tr>
<td>Bayesian estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td>0.00000</td>
<td>0.00182</td>
</tr>
<tr>
<td>Jeffrey’s</td>
<td>0.00000</td>
<td>0.00153</td>
</tr>
</tbody>
</table>

KNHANES, Korea National Health and Nutrition Examination Survey.

https://doi.org/10.5808/GI.2020.18.3.e31
while Score and Uniform methods have similar interval results. Bayesian Jeffrey’s prior gives better interval than Uniform Prior.

From Tables 1–5, confidence intervals are quite different depending on the method used. We think this inconsistency might be due to sparsity of data and small sample sizes. It is well known that sparsity causes the parameter estimates to fall near the boundary of the parameter space (for example, in proportion parameter value near to 0 or 1). As a result, the asymptotic methods such as the Wald method suffered from the convergence problem. In this sparse situation, either exact methods or Bayesian methods provide more reasonable confidence intervals. With modern computational power, it is not difficult to use exact inference for confidence interval directly from the binomial distribution without using large sample approximation to normality \[23\]. Due to discreteness, two exact methods provide different result. In Bayesian methods, Jeffrey’s prior provided better results than the uniform prior because it uses prior information for the scales of measurement, while the uniform prior does not use any prior information at all. Also, note that when the sample size was largest (Table 5), all methods provided more similar confidence intervals. In other words, as the sample size increases, all methods are expected to provide quite consistent confidence intervals.

In summary, as the sample size increases, the confidence interval become narrower. That is, as the sample size increases, more accurate estimation of antibody ratio is possible. In the confidence interval, the lower bound can be replaced by the number of confirmed patients through an actual test. Among the upper bound, the smallest value provides a conservative interpretation while the largest value does a more aggressive interpretation. Subtracting to today’s cumulative number of confirmed cases from the smallest upper limit can be interpreted as the minimum number of cases that were infected but not confirmed. As of 00:00 on September 15, 2020, at least 32,602 ( = 54,887–22,285) people were infected but not confirmed. This should be interpreted as having a high probability of cumulative infection.

### Table 4. 95% Confidence intervals (CIs) for the second antibody results using only KNHANES sample portion and for the total population in South Korea

<table>
<thead>
<tr>
<th>Method</th>
<th>Sample 95% CI (antibody retention rate)</th>
<th>Total population 95% CI (antibody carriers)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>Asymptotic estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td>0.00000</td>
<td>0.00201</td>
</tr>
<tr>
<td>Score</td>
<td>0.00001</td>
<td>0.00392</td>
</tr>
<tr>
<td>Likelihood ratio</td>
<td>0.00000</td>
<td>0.00308</td>
</tr>
<tr>
<td>Exact estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exact</td>
<td>0.00000</td>
<td>0.00491</td>
</tr>
<tr>
<td>MidP</td>
<td>0.00010</td>
<td>0.00310</td>
</tr>
<tr>
<td>Bayesian estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td>0.00016</td>
<td>0.00386</td>
</tr>
<tr>
<td>Jeffrey’s</td>
<td>0.00000</td>
<td>0.00324</td>
</tr>
</tbody>
</table>

KNHANES, Korea National Health and Nutrition Examination Survey.

### Table 5. 95% Confidence interval (CI) using the sum of both the first and second antibody results sample size and for the total population

<table>
<thead>
<tr>
<th>Method</th>
<th>Sample 95% CI (antibody retention rate)</th>
<th>Total population 95% CI (antibody carriers)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>Asymptotic estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td>0.00000</td>
<td>0.00106</td>
</tr>
<tr>
<td>Score</td>
<td>0.00001</td>
<td>0.00162</td>
</tr>
<tr>
<td>Likelihood ratio</td>
<td>0.00000</td>
<td>0.00137</td>
</tr>
<tr>
<td>Exact estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exact</td>
<td>0.00001</td>
<td>0.00178</td>
</tr>
<tr>
<td>MidP</td>
<td>0.00010</td>
<td>0.00110</td>
</tr>
<tr>
<td>Bayesian estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td>0.00014</td>
<td>0.00161</td>
</tr>
<tr>
<td>Jeffrey’s</td>
<td>0.00000</td>
<td>0.00143</td>
</tr>
</tbody>
</table>
Discussion

Statistical inference is a kind of process that helps us in making decisions about unknown population based on information contained in a sample taken from the population. There are two types of statistical inference: point estimation and hypothesis testing. The most important aspect of statistical inference is estimation of the unknown parameter, which is a procedure for finding the value of the unknown parameter by using the sample observations. For example, the sample means are used to estimate the population mean, sample proportion are used to estimate the population proportions, etc. An estimate of population parameter can be expressed in two ways: point estimate and interval estimate. A point estimate is a single number that is used to estimate an unknown population parameter. It is not much useful unless some information regarding possible error of estimate is associated with the estimate. For example, the sample proportion \( \hat{p} \) and we expect that this sample proportion is a good estimate of the population proportion. But the sample proportion vary from sample to sample and thus sampling error may be associated with the estimate. For a given sample proportion, the amount of sampling error is not known, so the standard error can be used as an estimate for the average amount of error in sample proportion. The total proportion cannot be identified with certainty based on only sample proportion without standard error. Therefore, instead of a point estimate, an interval or a range of values which is likely to contain the population parameter, which is known as interval estimation, must be provided. The major advantage of using interval estimation is that it provides a range of values with known probability of capturing the population parameter. Because we recognize sampling error, the point estimate has low confidence while interval estimates overcome this problem by using interval estimation techniques which is based on point estimate and margin of error. Thus, it is important to provide a point estimate along with its standard error or confidence intervals. Due to having more advantage of interval estimation relative to simple point estimation. In this article, we report different type of confidence interval for the sparse COVID-19 antibody test results from the Korean population.

Unfortunately, our study has several limitations. The current sample size for Korean studies is too small for accurate estimation of antibody test of Korean population. For estimation of antibody carriers in Korean population, we simply assumed that our antibody samples were derived from a simple random sampling assumption, because any detailed information of the samples is not available. Furthermore, the antibody samples may not represent the total Korean population well. To make more accurate estimation of antibody carriers in the Korean population, a much large sample size is required to represent the Korean population well. Then, with a further detailed demographic information of subjects and sampling information, more accurate sampling design-based inference can estimate the total number of antibody carriers in Korean population along with its 95% confidence intervals. In addition, there is no detailed information available for antibody test kits. Thus, in our estimations, it is expected that the kits should have high specificity, to accurately diagnose the subject without the COVID-19 antibody as negative, rather than high sensitivity to diagnose the subjects with the antibody as positive. No sensitivity and specificity information was considered in our analysis.

ORCID

Apio Catherine: https://orcid.org/0000-0003-0240-794X
Md. Kamruzzaman: https://orcid.org/0000-0002-3847-2666
Taesung Park: https://orcid.org/0000-0002-8294-590X

Authors’ Contribution

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

Conflicts of Interest

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

Acknowledgments

The authors thank Dr. Asaph Young Chun, Ph.D., Director-General, Statistics Research Institute Statistics, Korea for his extremely helpful comments.

References

3. Daily number of new reported cases of COVID-19 by country worldwide. Solna: European Centre for Disease Prevention and


Introduction

The reba carp, *Cirrhinus reba* (Hamilton, 1822) is an important freshwater labeonine fish of the family Cyprinidae and the order Cypriniformes. The species is geographically distributed over the Indian sub-continent including Bangladesh, India, Nepal, and Pakistan [1]. In Bangladesh, it is locally known as reba carp, tatkini, raik or bhagna bata and the natural habitats of the fish include the rivers, small creeks, natural depressions, and floodplains [2]. Siltation, conversion of wetlands, drying-up, and fragmentation of habitats, aquatic pollution, over fishing and climate changes have posed an alarming threat to the existence of many fish species including *C. reba*. Consequently, among 54 threatened freshwater fishes, *C. reba* is categorized as a near threatened (NT) species in Bangladesh [3]. In India, *C. reba* has been reported as vulnerable species based on International Union for Conservation of Nature (IUCN) criteria [4]. A few reports on the induced-breeding [5], stock identification [2], cryogenic preservation of sperm [1] have potentiated for the hatchery-based selective breeding, identification of the population...
status and the conservation of *C. reba*.

The mitochondrial genome is composed of a circular, double-stranded DNA with semi-autonomous central dogma machinery and situated in the mitochondria of each vertebrate cell outside the nuclear genome [6]. It possesses 37 genes in specific order that code for 13 polypeptides necessary for the oxidative phosphorylation (OxPhos) system [7]. The mitochondrial DNA of most metazoan species is predominantly inherited maternally [8]. The clonal inheritance, coupled with a substitution rate is typically 5–10 times more than that of nuclear DNA [9]. Mitochondrial DNA is an essential raw material as molecular marker for the investigation of population genetics [10], phylogenetic and evolutionary histories [11], genetic barcoding, and biodiversity of species [12], genetic disorder or mitochondrial diseases [13].

In the present study, we characterized the complete mitochondrial genome of *C. reba* using next-generation sequencing (NGS) with Illumina MiSeq platform. Sequence of the full mitochondrial genome would be an essential source to design specific molecular markers, for instance, single nucleotide polymorphisms studying population genetic structure. Moreover, detailed decoding of the mitochondrial genome would be a valuable insight for resolving any conservation and management strategies of the culture promising but NT fish.

**Methods**

**Specimen collection, genomic, and mitochondrial DNA isolation**

The specimen was collected from Khulna, Bangladesh (22°50′44.3″N, 89°32′27.7″E) in March 2017. The specimen is stored at the Department of Biotechnology, Sher-e-Bangla Agricultural University, Sher-e-Bangla Nagar, Dhaka 1207, Bangladesh. The genomic DNA was extracted from the fish tissue using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's manual. The mitochondrial partial COI sequence was amplified for the species confirmation using the BCL-BCH primers [14]. The specimen was confirmed based on both morphological characteristics and its COI sequence with 100% identity to the GenBank database (GenBank no. AP013325). The mitochondrial DNA was extracted with a commercially available mitochondrial DNA isolation kit (Abcam, Cambridge, UK) followed by the manufacturer's protocol and DNA concentration was checked by qubit fluorometer.

**Library construction and NGS**

In advance to the library preparation, the purified mitochondrial DNA was fragmented as 300–350 bp by Covaris M220 Focus-Ultrasonicator (Covaris Inc., Woburn, MA, USA). A library was prepared by TruSeq RNA library preparation kit V2 (Illumina, San Diego, CA, USA) which needed a couple of steps including end repair reaction, adenylation at 3′ ends, adaptor ligation, and enrichment of DNA fragments by PCR. To have better quality DNA, the enriched DNA library templates were purified using DNA purification kit (RBC Bioscience, Jerusalem, Israel) and quantified using qubit fluorometer. Size, purity, and quality of the constructed DNA libraries were further validated by the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Finally, the NGS was performed using the Illumina MiSeq platform 2 × 300 bp pair ends (Illumina).

**Sequence assembly and annotation of the mitochondrial genome**

The MiSeq raw reads were assembled using the Geneious Prime 2020.0.3 software [15] to construct a complete mitogenome of *Cirrhinus reba*. The gene order, sequences, and sizes of each of the 13 protein-coding genes and 2 ribosomal RNA genes were organized with the ORF finder program ([https://www.ncbi.nlm.nih.gov/orffinder/](https://www.ncbi.nlm.nih.gov/orffinder/)) following a reference mitogenome of *Cirrhinus reba* (GenBank no. AP013325). All transfer RNA genes were identified, the positions of their anticodons were predicted by using the ARWEN software [16].

**Gene mapping and phylogenetic tree construction**

The circular gene map of *C. reba* was drawn by OGDRAW software ([https://chlorobox.mpimp-golm.mpg.de/OGDraw.html](https://chlorobox.mpimp-golm.mpg.de/OGDraw.html)). The phylogenetic tree was constructed by MEGA7 program using the Minimum Evolution (ME) algorithm [17] performing with 1,000 bootstrap replications.

**Results and Discussion**

**Genome organization**

The complete mitochondrial genome of *C. reba* (GenBank no. MN862482) is a closed circular molecule of 16,597 bp in size. It possesses the typical combination of 37 genes including 13 protein-coding genes, 22 tRNA genes, two genes for ribosomal RNA subunits (12SrRNA and 16SrRNA), and two non-coding regions (control region, D-loop, and origin of light strand, OL) ([Table 1, Fig. 1](#)). Twenty-eight genes were located on the heavy (H) strand, while the ND6 gene and eight tRNA genes were transcribed from the light (L) strand ([Fig. 1](#)).

The overall nucleotide composition of *C. reba* mitogenome was found biased toward A + T contents (58.93%, A = 32.91% and T = 26.02%) over the G + C contents (41.07%, G = 15.16% and C...
Table 1. Organizational characteristics of the complete mitochondrial genome of *Cirrhinus reba*

<table>
<thead>
<tr>
<th>Gene/element</th>
<th>Strand</th>
<th>Nucleotide position</th>
<th>Size (bp)</th>
<th>Codon</th>
<th>Anti-codon</th>
<th>Intragenic nucleotides</th>
<th>A + T (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tRNA&lt;sup&gt;Phe&lt;/sup&gt;</td>
<td>H</td>
<td>1</td>
<td>70</td>
<td>70</td>
<td>-</td>
<td>-</td>
<td>GAA</td>
</tr>
<tr>
<td>12S rRNA</td>
<td>H</td>
<td>71</td>
<td>1,024</td>
<td>954</td>
<td>-</td>
<td>-</td>
<td>TAC</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Val&lt;/sup&gt;</td>
<td>H</td>
<td>1,025</td>
<td>1,096</td>
<td>72</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>16S rRNA</td>
<td>H</td>
<td>1,097</td>
<td>2,783</td>
<td>1,687</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Leu&lt;/sup&gt;</td>
<td>H</td>
<td>2,784</td>
<td>2,860</td>
<td>77</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ND1</td>
<td>H</td>
<td>2,861</td>
<td>3,835</td>
<td>975</td>
<td>ATG</td>
<td>TAA</td>
<td>3</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Val&lt;/sup&gt;</td>
<td>H</td>
<td>3,839</td>
<td>3,912</td>
<td>74</td>
<td>-</td>
<td>-</td>
<td>GAT</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Glu&lt;/sup&gt;</td>
<td>L</td>
<td>3,910</td>
<td>3,980</td>
<td>71</td>
<td>-</td>
<td>-</td>
<td>TG</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Met&lt;/sup&gt;</td>
<td>H</td>
<td>3,982</td>
<td>4,050</td>
<td>69</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ND2</td>
<td>H</td>
<td>4,051</td>
<td>5,096</td>
<td>1,046</td>
<td>ATG</td>
<td>TA-</td>
<td>-1</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Leu&lt;/sup&gt;</td>
<td>H</td>
<td>5,096</td>
<td>5,167</td>
<td>72</td>
<td>-</td>
<td>-</td>
<td>TAA</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Met&lt;/sup&gt;</td>
<td>H</td>
<td>5,169</td>
<td>5,237</td>
<td>69</td>
<td>-</td>
<td>-</td>
<td>GTC</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Ala&lt;/sup&gt;</td>
<td>H</td>
<td>5,239</td>
<td>5,311</td>
<td>73</td>
<td>-</td>
<td>-</td>
<td>GTT</td>
</tr>
<tr>
<td>OL</td>
<td>-</td>
<td>-</td>
<td>5,312</td>
<td>5,345</td>
<td>34</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Cys&lt;/sup&gt;</td>
<td>L</td>
<td>5,346</td>
<td>5,412</td>
<td>67</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Tyr&lt;/sup&gt;</td>
<td>H</td>
<td>5,485</td>
<td>5,167</td>
<td>1,551</td>
<td>ATG</td>
<td>TA-</td>
<td>-1</td>
</tr>
<tr>
<td>COX1</td>
<td>H</td>
<td>5,169</td>
<td>5,197</td>
<td>156</td>
<td>ATG</td>
<td>TAA</td>
<td>1</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Ser&lt;/sup&gt;</td>
<td>L</td>
<td>5,197</td>
<td>5,236</td>
<td>70</td>
<td>-</td>
<td>-</td>
<td>GTA</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Asp&lt;/sup&gt;</td>
<td>L</td>
<td>5,236</td>
<td>5,345</td>
<td>70</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>COX2</td>
<td>H</td>
<td>5,345</td>
<td>5,387</td>
<td>1,532</td>
<td>ATG</td>
<td>TAA</td>
<td>1</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Leu&lt;/sup&gt;</td>
<td>H</td>
<td>5,387</td>
<td>5,483</td>
<td>70</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ND3</td>
<td>H</td>
<td>5,483</td>
<td>5,487</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Val&lt;/sup&gt;</td>
<td>H</td>
<td>5,487</td>
<td>5,544</td>
<td>1,551</td>
<td>ATG</td>
<td>TAA</td>
<td>1</td>
</tr>
<tr>
<td>ATP8</td>
<td>H</td>
<td>5,544</td>
<td>5,605</td>
<td>61</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ATP6</td>
<td>H</td>
<td>5,605</td>
<td>5,648</td>
<td>43</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>COX3</td>
<td>H</td>
<td>5,648</td>
<td>5,691</td>
<td>42</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Gly&lt;/sup&gt;</td>
<td>L</td>
<td>5,648</td>
<td>5,705</td>
<td>68</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ND4</td>
<td>H</td>
<td>5,705</td>
<td>5,773</td>
<td>69</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ND4L</td>
<td>H</td>
<td>5,773</td>
<td>5,825</td>
<td>52</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ND5</td>
<td>H</td>
<td>5,825</td>
<td>5,879</td>
<td>54</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ND6</td>
<td>H</td>
<td>5,879</td>
<td>5,931</td>
<td>52</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Pro&lt;/sup&gt;</td>
<td>L</td>
<td>5,931</td>
<td>5,988</td>
<td>58</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CYTB</td>
<td>H</td>
<td>5,988</td>
<td>6,041</td>
<td>54</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Leu&lt;/sup&gt;</td>
<td>H</td>
<td>6,041</td>
<td>6,095</td>
<td>55</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>tRNA&lt;sup&gt;Phe&lt;/sup&gt;</td>
<td>L</td>
<td>6,095</td>
<td>6,149</td>
<td>54</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Control region (D-loop)</td>
<td>-</td>
<td>15,668</td>
<td>16,597</td>
<td>930</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The percentage of G was 25.91%, indicating a strong anti-guanine bias commonly observed in fishes [18]. This pattern was almost the same as the Cyprinid species, which has 15.70% of G content [19]. A total of 31 bp overlapping region was found in 11 different locations of the *C. reba* mitogenome. Besides, 11 intergenic spacers were present in the mitogenome involving a total size of 34 bp. The longest spacer included a nucleotide sequence of 15 bp which located between the tRNA Asp and COX2 genes. As shown in Table 1, the positive (+) numbers indicate intergenic space/gap between genes, the negative (–) numbers indicate overlapping between genes, and the zero (0) revealed either overlap or space does not exist between genes in the *C. reba* mitogenome.
Protein-coding genes

The mitochondrial genome of *C. reba* is consisted of 13 canonical protein-coding genes (PCGs) of 11,412 bp in length, accounted for 68.76% of the total mitogenome. With exception to the reading frame of ND6 gene, which was oriented clock-wise direction on the light strand, reading frames of all other PCGs were oriented counter-clock wise direction on the heavy strand (H). The shortest PCG was ATP8 (165 bp), whereas the longest one was ND5 (1,824 bp). Except for the COX1 gene with an unusual and alternative start codon, GTG (Valine), other twelve PCGs started with the typical translation initiation codon, ATG (Methionine). The open reading frame of ATP8 used TAG as stop codon, whereas, the remaining seven PCGs (ND1, COX1, ATP6, COX3, ND4L, ND5, and ND6) terminated with the typical stop codon (TAA), as in other vertebrates [20]. Four PCGs, COX2, ND3, ND4, CYTB ended with an incomplete termination codon (T--), while the ND2 gene ended with “TA-” (Table 1). These incomplete stop codons are assumed to be completed by using the posttranscriptional polyadenylation, poly-A tail [21]. These kinds of variations in the termination codons are not uncommon among vertebrate mitochondrial genes and previously reported in other bony fishes [22].

Analysis of the sequences of 13 PCGs showed that overall percentage of A and T contents (58.93%) also reflected in the codon usage of PCGs where the first, second and third position of the codons estimated the A plus T contents as 49.6%, 58.9%, and 69%, respectively (Table 2). The G content (15%) of all PCGs presents obvious anti-guanine characteristics as similar to other bony fishes [23]. Here, we observed that the 2nd and 3rd position of the codons were dominated by T (40.4%) and A (46.7%) contents, respectively over the C and G contents (Table 2). A very strong bias against the guanine contents (4.9%) was also demonstrated at the wobble or third position of codons across all 13 PCGs.

Ribosomal RNA and transfer RNA genes

The mitogenome of *C. reba* possesses two genes encoding two ribosomal RNA subunits, a small (12S) and a large (16S) which were typical as in other mitogenomes. Both the ribosomal RNA genes were located on the H-strand and consisted of 15.91% (2,641 bp) of the total circular mitogenome. The 12S rRNA (946 bp in length) located between the tRNA^{Val} and tRNA^{Val}, whereas...
the 16S rRNA was 1,695 bp in length and was located between tRNA Val and tRNA Leu. The A + T content (55.24%) was higher than the G plus C content (44.76%) in the ribosomal RNAs where the nucleotide, adenine was more prevalent (35.25%) followed by cytosine (24.50%) as similar to the previous reports in other bony fishes [24].

The 22 tRNA genes were identified in the mitochondrial of C. reba which demonstrated a varied size of 67–78 bp in length, estimating a total length of 1,580 bp (~9.6% of the total mitogenome) (Table 1). Among all tRNA genes, eight tRNAs were located on the L-strand, whereas the remaining fourteen tRNA genes were on the H-strand (Fig. 1). The highest content of A and T was observed in tRNA Ala and tRNA His (69.6%), whereas that of the lowest content was observed in tRNA Met (42%) which was consisted of previously published reports on other teleosts [24]. All these 22 tRNA genes were predicted to capable of folding typical cloverleaf secondary structures and showed great similarity as it is in other vertebrate mitogenomes [23].

Non-coding regions
Like other animal mitochondrial genome, C. reba contained no introns and possessed two non-coding regions, an OL and a control region or displacement loop (D-loop) region. The OL region, consisted of 34 nucleotides was situated between tRNA Aln and tRNA Cys and was oriented on the L-strand in a cluster of five tRNA genes (WANCY region). D-loop region of C. reba consisted of 930 nucleotides which represented 5.6 % of the total mitogenome. The control region was overpresented by A plus T content (67.6 %). The AT-rich D-loop contains promoters and an origin of replication of mtDNA which are essential for transcription and replication of mtDNA, respectively [25]. The D-loop region is very flexible to size variation.

Phylogenetic relationship analysis
The phylogenetic relationship of C. reba with other labeonine fishes was constructed using 13 complete mitochondrial genomes of the subfamily Labeoninae by MEGA 7.0 program. The mitogenome of our species of interest, C. reba (GenBank no. MN862482) showed 99.6% sequence identity with different haplotype of C. reba (GenBank no. AP013325, deposited from Japan), followed by 90.18% identity with Labeo bata (AP011198). Slight variations between two haplotypes of C. reba might reflect different populations of the same species. However, two haplotypes of C. reba was clearly isolated from other 11 species of labeonine fishes and formed a separate clade (Fig. 2). Two close relatives of C. reba, L. pangusia, and C. cirrhosis have already been considered globally as
Fig. 2. Phylogenetic tree of *Cirrhinus reba* in the subfamily Labeoninae. The phylogenetic relationship was analyzed by MEGA7 program with Minimum Evolution algorithm and 1,000 bootstrap replications. GenBank accession number of each species is shown next to its scientific name. *Reported in the present study.*

NT and vulnerable species, respectively [3]. Therefore, we alarmingly recommend further studies relating to conservation of the labeonine fish including reba carp.

**ORCID**

Mohammad Nazrul Islam: https://orcid.org/0000-0001-5692-3566  
Shirin Sultana: https://orcid.org/0000-0002-0607-1497  
Md. Jobaidul Alam: https://orcid.org/0000-0002-3594-8147

**Authors’ Contribution**

Conceptualization: MNI, MJA. Data curation: MNI, MJA. Formal analysis: MNI, MJA. Funding acquisition: MNI. Methodology: MNI, MJA. Writing - original draft: MNI, SS. Writing - review & editing: SS, MJA.

**Conflicts of Interest**

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

**References**


1971.


Application note

This paper describes a community effort to improve earlier versions of the full-text corpus of Genomics & Informatics by semi-automatically detecting and correcting PDF-to-text conversion errors and optical character recognition errors during the first hackathon of Genomics & Informatics Annotation Hackathon (GIAH) event. Extracting text from multiple column biomedical documents such as Genomics & Informatics is known to be notoriously difficult. The hackathon was piloted as part of a coding competition of the ELTEC College of Engineering at Ewha Womans University in order to enable researchers and students to create or annotate their own versions of the Genomics & Informatics corpus, to gain and create knowledge about corpus linguistics, and simultaneously to acquire tangible and transferable skills. The proposed projects during the hackathon harness an internal database containing different versions of the corpus and annotations.

Keywords: biomedical text mining, corpus, text analytics

Introduction

For biomedical text mining, it is necessary to use a corpus, which refers to a large and structured set of texts that have been electronically stored and processed. The full text of Genomics & Informatics (G&I) has been archived since 2003 as PDF files [1], and the content of the journal is available immediately upon publication without an embargo period. Even though the full-text publications of recent volumes are available as XML files, only scanned images or PDF files are available for earlier versions of publications, necessi-
tating the conversion of images into machine-encoded text.

Thus, to build an initial version of the G&I corpus 1.0, we wrote a simple Python-based web crawler to directly browse and download PDF files from the G&I archives; then, we converted the PDFs into plain text files using PDFMiner or other optical character recognition (OCR) tools [2]. In this way, a prototype version of the full text-corpus of G&I 1.0 was recently archived in the GitHub repository, in 2018 [3].

Unfortunately, earlier versions of the G&I corpus 1.0 are of poor quality, and the noise induced by these errors present thorny issues for downstream standard text analysis pipelines, including tokenization, sentence boundary detection, and part-of-speech (POS) tagging, that would be used to develop the next version of the G&I corpus. Consequently, it was impossible to directly employ the obtained results for subsequent tasks without costly manual editing.

It was necessary to obtain motivated volunteers. To address this problem, the first event of Genomics & Informatics Annotation Hackathon (GIAH) was organized at Ewha Womans University, Korea to join forces for biomedical text mining with the goal of improving G&I; a hackathon is typically an event in which computer programmers and others involved in software development collaborate intensively over a short period of time on software projects [4].

Accurately extracting texts from PDF files has been an important issue for decades in the area of natural language processing and text mining. Nonetheless, we still do not have a definitive solution. In that sense, this hackathon tackled an important and not-yet-solved problem. Thus, our aim in the present paper is to describe a community effort to construct enhanced versions of the G&I corpus, in a consistent machine-readable format. We describe and summarize a collection of corpus projects reflecting achievements from this hackathon.

Patterns of PDF-to-Text Conversion Errors

ASCII text and HTML text are human-readable formats. Text often comes in human unreadable formats, such as PDF files, that can only be opened using specialized software. Third-party libraries such as Adobe Acrobat Reader or PDFMiner provide access to these formats [2]. However, PDF conversion tools and OCR tools are still imperfect, as they occasionally misrecognize letters and falsely identify text, leading to misspellings and linguistic errors in the output text.

Most OCR conversion errors occur at line boundaries, where words are divided at the nearest break point between syllables, and a hyphen is inserted to indicate that the letters form a word fragment, rather than a full word. Thus, a word can be incorrectly separated (e.g., “se-parated” vs. “separated”). Many of these hyphenation errors could have been corrected, automatically, by applying some pattern-matching rules to these cases of hyphenation.

However, converting a PDF to a text file produces some odd and serious errors that need to be manually fixed. Thus, many errors need to be corrected manually, especially due to the fact that G&I contains many biomedical terms, many of which even contain special characters.

The first event of the GIAHHackathon was held at the ELTEC College of Engineering of Ewha Womans University, 2020, with 76 participants, to enhance the G&I 1.0 corpus [4]. The word hackathon is combined from the words “hack” and “marathon,” where “hack” is used in the sense of exploratory and investigative programming. A meeting was held as a symposium to exchange and publicize the activities and ideas of improving the earlier volumes of the G&I corpus 1.0 (Vol. 1 to Vol. 9), explaining various issues and problems, as shown in Fig. 1. The participants worked on implementing their ideas with collaboration with other participants during a 2-week period.

Most of the teams initially applied regular expressions, correcting hyphenation, single-error misspellings, and a certain class of double-error misspellings, which are the major source of inaccuracies [5]. The corpus was processed and upgraded in several separate stages: manual editing by individuals, automatic editing by writing new pattern matching rules, and a checking and update loop to enhance the corpus, in an iterative cycle.
Various strategies were proposed based on composite machine-learning methods. Linguistic context-based error correction techniques were also used by most of the teams to detect and correct OCR errors with respect to their grammatical and semantic context [6–8]. Some participants proposed a method of automating the correction of misspelled words using on-line spell checkers [9]. This solution consists of using a lookup dictionary to search for misspelled words and correcting them suitably. Several teams used word embedding and deep learning techniques, such as Word2Vec, and BERT, with the idea of using context based on linguistic categories [10–14]. Still, this semi-automatic procedure is considered laborious and error-prone, as humans may miss some mistakes.

Many versions of the corpus were submitted. However, comparison of the performance of each project was difficult, as evaluation requires additional manual labor. Instead, we used several text comparison programs (open-source differencing and merging tools). These programs are highly useful for determining what has changed between different corpus versions, and then merging changes between versions.
Fig. 2. A screenshot of text comparison software (WinMerge) used to search for differences between two versions of texts (G&I 1.0 and the improved version) in order to highlight corrections made in G&I Vol. 7 No. 2 [15].

Fig. 3. Five representative versions of the Genomics & Informatics corpus constructed during the hackathon are available through the subfolders of “G&I Hackathon 2020” of GitHub (https://github.com/Ewha-Bio/Genomics-Informatics-Corpus): raw1, raw2, raw3, raw4, and raw5.
Table 1. Number of files and updated lines in five folders of the GIAH hackathon archives

<table>
<thead>
<tr>
<th>Folder</th>
<th>No. of files in the folder</th>
<th>No. of updated lines</th>
<th>Average No. of updated lines per file</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw1</td>
<td>183</td>
<td>8,513</td>
<td>46.5</td>
</tr>
<tr>
<td>raw2</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>raw3</td>
<td>487</td>
<td>11,124</td>
<td>22.8</td>
</tr>
<tr>
<td>raw4</td>
<td>337</td>
<td>14,707</td>
<td>43.6</td>
</tr>
<tr>
<td>raw5</td>
<td>487</td>
<td>10,233</td>
<td>21</td>
</tr>
</tbody>
</table>

The statistics of the raw2 folder were unavailable for technical reasons. GIAH, Genomics & Informatics Annotation Hackathon.

Fig. 2 shows a WinMerge [15] screenshot of error corrections, where a search was made for differences between two versions of texts (G&I 1.0 and the improved version) in order to highlight corrections made in G&I Vol. 7 No. 2. For example, 59 corrections were detected in the modified version of gni-7-2-97 (https://doi.org/10.5808/gi.2009.7.2.097) in the raw1 folder [16]. Among them, 30 were manual edits, and 29 were automatic edits. Likewise, 54 corrections were detected in the modified version of gni-7-2-111 file (https://doi.org/10.5808/gi.2009.7.2.111) in the raw1 folder [17]. Among them, 30 are manual edits, and 22 were automatic edits.

Among all the submitted hackathon archives, the five best-performing versions of modified G&I corpus were selected and uploaded to subfolders of “G&I Hackathon 2020” on GitHub as shown in Fig. 3: raw1, raw2, raw3, raw4, and raw5.

Table 1 shows the number of files and updated lines in each of the five folders of GIAH hackathon archives. Among them, the raw1 folder (submitted by two participants, Sunho Kim and Royoung Kim) showed the best overall performance based on the number of manual corrections, the number of automatic corrections, documentation, and file coverage. We manually checked the error correction rate of randomly chosen files in the raw1 folder, and on average, 30.3 occurrences of manual corrections and 24.1 occurrences of automatic corrections could have been detected per article, which are slightly larger numbers than were automatically detected by software in Table 1. Thus, the release of these improved corpora could potentially be a meaningful contribution.

Conclusion

In this paper, we listed issues associated with upgrading the G&I corpus, and discussed methodological strategies to develop the next version of the G&I corpus based on a semi-automatic approach. Besides manual corrections, the outcome using pattern matching techniques and machine learning methods was noteworthy, and it greatly improved the error correction rate.

This is a progress report, and the current debate regarding our post-processing procedures focuses on how to ensure the quality of this semi-automatically modified corpus. It is taken as axiomatic that any correction must be confirmed by at least two, and usually more, people acting independently, so that their modification decisions can be compared. We suggest that a couple more rounds of the GIAH hackathon be organized to construct the future G&I 2.0 corpus. A semi-automatic method should be designed to build and improve the corpus, with a diminishing amount of manual checking.

Orcid

Sunho Kim: https://orcid.org/0000-0002-0824-5040
Royoung Kim: https://orcid.org/0000-0002-8698-800X
Hee-Jo Nam: https://orcid.org/0000-0001-6184-6737
Ryeo-Gyeong Kim: https://orcid.org/0000-0001-9506-2877
Enjin Ko: https://orcid.org/0000-0001-6327-4293
Han-Su Kim: https://orcid.org/0000-0002-5525-5911
Jihe Shin: https://orcid.org/0000-0002-8593-0607
Daeun Cho: https://orcid.org/0000-0003-3336-3150
Yurhee Jin: https://orcid.org/0000-0002-8896-7739
Seoyeon Bae: https://orcid.org/0000-0001-9282-7707
Ye Won Jo: https://orcid.org/0000-0002-0195-9550
San Ah Jeong: https://orcid.org/0000-0001-8045-1597
Yena Kim: https://orcid.org/0000-0002-9487-7617
Seoyeon Ahn: https://orcid.org/0000-0001-9362-0521
Bomi Jang: https://orcid.org/0000-0002-8166-0136
Jiheyon Seong: https://orcid.org/0000-0002-7764-8193
Yujin Lee: https://orcid.org/0000-0003-3380-5940
Si Eun Seo: https://orcid.org/0000-0002-4944-0942
Yujin Kim: https://orcid.org/0000-0001-5307-6357
Ha-Jeong Kim: https://orcid.org/0000-0003-1823-510X
Hyeji Kim: https://orcid.org/0000-0001-8711-178X
Hye-Lynn Sung: https://orcid.org/0000-0002-9229-4362
Hyoyoung Lho: https://orcid.org/0000-0002-4222-899X
Jaywon Koo: https://orcid.org/0000-0002-5539-5244
Jion Chu: https://orcid.org/0000-0002-3165-0992
Juwon Lim: https://orcid.org/0000-0003-0444-5886
Youngju Kim: https://orcid.org/0000-0003-2998-3699
Kyungyoon Lee: https://orcid.org/0000-0002-3793-2616
Yuri Lim: https://orcid.org/0000-0002-2390-6273
Meongeun Kim: https://orcid.org/0000-0002-1549-1362
Seonjeong Hwang: https://orcid.org/0000-0002-1196-2040
Shinhye Han: https://orcid.org/0000-0001-8157-3676
Sohyeun Bae: https://orcid.org/0000-0002-7873-2207
Sua Kim: https://orcid.org/0000-0003-4458-3319

https://doi.org/10.5808/GI.2020.18.3.e33
Authors’ Contribution

Conceptualization: HSP. Data curation: HJN, SK (Sunho Kim), RK, RGK, EK, HSK, JS (Jihye Shin), DC, YJ, SB (Soyeon Bae), SAJ, YK (Yena Kim), SA, BJ, JS (Jihyeon Seong), YL (Yujin Lee), SES, YK (Yujin Kim), YWJ, HJK, HK, HLS, HL, JK, JC, JL, YK (Youngju Kim), KL, YL (Yuri Lim), MK, SH (Seonjeong Hwang), SH (Shinhye Han), SB (Soheyun Bae), SK (Sua Kim), SY, YS (Yeongjeong Seo), YS (Yerim Shin), YK (YuJin Kim), YWJ, HJK, HH, HC. Methodology: HJN. Writing – original draft: HSP. Writing – review & editing: HJN, HSP, DYK, JHO.

Conflicts of Interest

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

Acknowledgments

This work was supported by a National Research Foundation of Korea grant (NRF-2019R1F1A1058858) funded by the Korean government (MSIT).

References


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

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

**Research and publication ethics**

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

**Statement of Informed Consent**

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

**Statement of Human and Animal Rights**

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

**Authorship**

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

**Corresponding author and first author**

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

**Correction of authorship after publication**

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

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

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

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

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

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

**Preparation of manuscripts**

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

**Manuscript type**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Examples of references are given below:

Journal article

Books

Book sections

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

**Conference paper**

**Dissertation/Thesis**

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

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

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

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

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

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

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

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

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

Gender equity (Described according to ICMJE recommendation available from http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html)
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 cannot be publicized, it can be negotiated with the editor. If there are any inquiries on depositing data, authors should contact the editorial office.

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

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

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

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

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

Statement of Informed Consent

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

Statement of Human and Animal Rights

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

Authorship

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

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

Corresponding author and first author

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

Correction of authorship after publication

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

Conflict of Interest Statement

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

Originality and Duplicate Publication

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

Secondary Publication

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

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

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


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

☐ 3. Abstract in unstructured format within 250 words.

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

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

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

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

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

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

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