Genomics & Informatics is indexed/tracked/covered by PubMed, PubMed Central, KoreaMed, KoMCI, ScienceCentral, CrossRef, BIOSIS Previews, DOAJ, and Google Scholar.
Aims and scope

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

It is published and distributed quarterly at the last dates of March, June, September, and December. All submitted manuscripts will be reviewed and selected for publication after single blind review process. All manuscripts must be submitted online through the e-submission system available from: http://submit.genominfo.org. It is an online-only peer reviewed open access journal. A free full text both in the XML and PDF formats is available from the journal homepage (https://genominfo.org). It has been indexed by or searchable from PubMed, PubMed Central, 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
Jeon 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.
Srijulnath, Kornsorn
Kasetart University, Thailand
Terwilliger, Joseph
Columbia University, U.S.A.
Valdes, Jorge
Centro de Genómica y Bioinformática, Chile
Van, Steen
Kristel University of Liège, Belgium
Zhang, Feng
Fudan University, China

ETHICS EDITOR
Chung, Yeun-Jun
The Catholic University, 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 18:1, 2020)
Taesung Park

Original articles

Validity of patient-derived xenograft mouse models for lung cancer based on exome sequencing data
Jaewon Kim, Hwanseok Rhee, Jhingook Kim, Sanghyuk Lee

Strong concordance between RNA structural and single nucleotide variants identified via next generation sequencing techniques in primary pediatric leukemia and patient-derived xenograft samples
Sonali P. Barwe, Anilkumar Gopalakrisnapillai, Nitin Mahajan, Todd E. Druley, E. Anders Kolb, Erin L. Crowgey

Associations between an MDM2 gene polymorphism and ulcerative colitis by ARMS-PCR
Mahsa Sadat Hashemi Doulabi, Reza Golejani Moghadam, Ali Salehzadeh

Bioinformatics services for analyzing massive genomic datasets
Gunhwan Ko, Pan-Gyu Kim, Youngbum Cho, Seongmun Jeong, Jae-Yoon Kim, Kyoung Hyoun Kim, Ho-Yeon Lee, Jiyeon Han, Namhee Yu, Seokjin Ham, Insoon Jang, Byunghee Kang, Sunguk Shin, Lian Kim, Seung-Won Lee, Dougu Nam, Jihyun F. Kim, Namshin Kim, Seon-Young Kim, Sanghyuk Lee, Tae-Young Roh, Byungwook Lee

Quantitative evaluation of the molecular marker using droplet digital PCR
Wonseok Shin, Haneul Kim, Dong-Yep Oh, Dong Hee Kim, Kyudong Han

Detecting outliers in segmented genomes of flu virus using an alignment-free approach
Mosaab Daoud

Research communication

A note on the distance distribution paradigm for Mosaab-metric to process segmented genomes of influenza virus
Mosaab Daoud

Application notes

PAIVS: prediction of avian influenza virus subtype
Hyeon-Chun Park, Juyoun Shin, Sung-Min Cho, Shinseok Kang, Yeun-Jun Chung, Seung-Hyun Jung

HisCoM-PCA: software for hierarchical structural component analysis for pathway analysis based using principal component analysis
Nan Jiang, Sungyoung Lee, Taesung Park

Clinical genomics

Accelerating next generation sequencing data analysis: an evaluation of optimized best practices for Genome Analysis Toolkit algorithms
Karl R. Franke, Erin L. Crowgey
Editor's introduction to this issue (G&I 18:1, 2020)

Taesung Park

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

In this issue, there are 10 articles: six Original Articles, one Research Communication, two Application Notes, and one article in the category of Clinical Genomics. The first two original articles deal with patient-derived xenografts (PDX). First, Kim et al. (Ewha Womans University, Korea) presented PDX mouse models for 132 lung cancer patients and performed whole-exome sequencing to compare tumor, normal, and xenograft tissues. Through a computational analysis of the somatic mutations and copy number variations, the authors showed that the genomic and histological results agreed well, with more than 90% of concordant cases. Their analyses demonstrate the potential usefulness of PDX mouse models in cancer studies.

Second, Barwe et al. (Alfred I. duPont Hospital for Children, USA) generated 25 primary pediatric acute leukemia samples and their corresponding PDX samples. The authors demonstrated that the primary samples and PDX samples showed a high level of concordance between single nucleotide variants and gene fusions, while other complex structural variants were not as consistent. The high concordance between single nucleotide variants and gene fusions confirms the utility of PDX models for preclinical drug testing.

The third article, by Doulabi et al. (Islamic Azad University, Iran), presents a case-control study of 174 ulcerative colitis biopsy samples and 82 control individuals. The authors performed a candidate gene association analysis of MDM2, which is a phospho-protein and a ubiquitin ligase for p53. The rs309 single nucleotide polymorphism (SNP) detected by the amplification-refractory mutation system PCR technique was shown to be associated with the occurrence of ulcerative colitis. A further study on the direct association of this polymorphism with carcinogenesis is warranted.

Ko et al. (Korea Bioinformation Center, KRBIB, Korea) presented a cloud computing-based system, Bio-Express, for handling a large amount of genomic data. Bio-Express provides user-friendly, cost-effective analysis of massive genomic datasets loaded with multi-omics data analysis pipelines including genome, transcriptome, epigenome, and metagenome pipelines. Bio-Express is a highly efficient cloud computing-based system.

Shin et al. (Dankook University, Korea) presented a platform for detection of the Hanwoo-specific structure variation using droplet digital PCR (ddPCR). The ddPCR platform is expected to provide more accurate quantification than PCR and can be applied for the quantitative evaluation of molecular markers.

The final Research Article is by Daoud, who proposed a new robust approach to detecting outliers in a set of segmented genomes of the influenza virus with feature extraction, an alignment-free distance measure, and a mapping into distance space to analyze a quantum of distance values. In his sequel article, the author presents a few technical notes about the distance distribution paradigm used to analyze composite data points in high-dimensional feature spaces. The integrated statistical learning pipeline to process segmented genomes of the influenza virus is illustrated as a sequential-parallel computa-
In this issue, there are two Application Notes. Park et al. (The Catholic University of Korea, Korea) developed a user-friendly tool, named prediction of avian influenza virus subtype (PAIVS). PAIVS is an analysis pipeline of next-generation sequencing (NGS)-based avian influenza virus (AIV) sequencing data that supports the pre-processing of NGS data, reference-guided AIV subtyping, de novo assembly, variant calling, and identifying the closest full-length sequences by BLAST, and then provides a graphical summary to the end user. Jiang et al. (Seoul National University, Korea) presented the HisCoM-PCA software for performing pathway analysis of SNP data using hierarchical structural component models. HisCoM-PCA is based on principal component analysis (PCA) for the dimensional reduction of SNPs in each gene, and a hierarchical structural component model for pathway analysis. The HisCoM-PCA software has several features.

Various selection criteria for the principal component scores in the PCA step can be specified by the user. Multiple public pathway databases and customized pathway information can be used to perform pathway analysis.

The one article in the Clinical Genomics section by Franke and Crowgey (Nemours Alfred I duPont Hospital for Children, USA) provides an evaluation of optimized best practices for genome analysis toolkit (GATK) algorithms, including Parabricks and Sentieon. The evaluation results would be highly informative for users to decide which algorithm of GATK to use to analyze large-scale human genomics datasets.

**ORCID**

Taesung Park: https://orcid.org/0000-0002-8294-590X
Patient-derived xenograft (PDX) mouse models are frequently used to test the drug efficacy in diverse types of cancer. They are known to recapitulate the patient characteristics faithfully, but a systematic survey with a large number of cases is yet missing in lung cancer. Here we report the comparison of genomic characters between mouse and patient tumor tissues in lung cancer based on exome sequencing data. We established PDX mouse models for 132 lung cancer patients and performed whole exome sequencing for trio samples of tumor-normal-xenograft tissues. Then we computed the somatic mutations and copy number variations, which were used to compare the PDX and patient tumor tissues. Genomic and histological conclusions for validity of PDX models agreed in most cases, but we observed eight (~7%) discordant cases. We further examined the changes in mutations and copy number alterations in PDX model production and passage processes, which highlighted the clonal evolution in PDX mouse models. Our study shows that the genomic characterization plays complementary roles to the histological examination in cancer studies utilizing PDX mouse models.

**Keywords:** copy number alteration, lung neoplasms, mutation, patient-derived xenograft, whole exome sequencing

---

**Introduction**

Lung cancer incidence and mortality rates are the highest worldwide, accounting for 11.6% of the total cases and 18.4% of the total cancer deaths in 2018 [1]. Traditional treatment for lung cancer has been surgery and radiochemotherapy, but targeted therapies are increasingly adopted for patients who have the druggable aberrations such as epidermal growth factor receptor (EGFR) mutations or gene fusions involving ALK, ROS1, and NTRK genes [2-5]. Targeted therapies usually show fast response with minimal side effects, but tumor recurs within a few months in many cases, thus necessitating additional therapies.

The main reasons for resistant and recurrent tumors are intrinsic heterogeneity and tumor cell evolution. Tumors may consist of multiple clones where targeted therapies kill only subsets of clones leaving residual clones, whose proliferation leads to resistance or recurrence eventually. Alternatively, tumor cells may undergo evolution after treatment to acquire *de novo* mutations overcoming the treatment effect of cancer drugs. Understand-
ing molecular mechanisms of resistance development is essential to identify follow-up treatment options in targeted therapy.

Preclinical models are extremely useful in the course of drug development, especially to test the drug efficacy in cost-effective ways. Animal models and organoids derived from patient tumors are two representative systems frequently adopted in cancer drug development. Patient-derived xenograft (PDX) mouse models where the patient’s tumor tissue is transplanted into immunodeficient mice have demonstrated their usefulness to recapitulate patient’s response to cancer agents in various types of cancers including breast, brain, colon, and lung tumors [6-10]. The banks of these ‘Avatar’ mice are valuable resources for preclinical tests. However, the fidelity of PDX mouse models to substitute patient’s tumor tissues has not been thoroughly studied in lung cancer models. Here we compare the genomic characteristics of PDX mouse, patient tumor, and patient normal tissues based on exome sequencing data to test the validity of PDX mouse models in lung cancer.

Methods

Producing whole exome sequencing data
We acquired the tumor and matched normal tissues from 132 lung cancer patients at the Samsung Medical Center in Seoul. This study was approved by Samsung Medical Center institutional review board (IRB 2018-03-110), and informed consent was obtained from each patient. Tumor tissues were transplanted into the NSG mouse, NOD.Cg-PrkdcscidIl2rgtm1Wjl/SzJ (Stock No. 005557) [11], purchased from the Jackson Laboratory to establish the PDX mouse models. All tumor tissues of the patient and PDX were prepared for formalin-fixed paraffin-embedded, and a pathologist examined histopathology of tissue through hematoxylin and eosin staining. Whole exome sequencing (WES) was performed using the Illumina TruSeq Exome kit and HiSeq 2500 sequencing platform (Illumina, San Diego, CA, USA). The WES data for patient tumor, patient normal, and PDX tumor tissues were deposited at the Korean Bioinformation Center (KOBIC) (ID 10050154).

Data processing and variant analysis
Preprocessing steps consist of adapter trimming, quality control, and filtering mouse reads. First, adapter sequences were removed and sequence reads whose quality score < 33 in more than 50% of bases were discarded using fastx-toolkit (ver. 0.0.14). After the trimming process, we removed the single reads and kept the paired-end reads only using cmpfastq perl program (http://compbio.brc.iop.kcl.ac.uk/software/cmpfastq.php). For WES data from PDX mouse tissues, we applied an additional step to filter out mouse-originated reads using Xenome software (ver. 1.0.1) [12] with the reference genomes of human (UCSC hg19 in https://genome.ucsc.edu/) and mouse (UCSC mm10). We kept the human-specific reads only for subsequent analyses.

Resulting reads were mapped to the human reference genome (UCSC hg19) using Burrows-Wheeler Alignment (BWA)-MEM alignment tool [13] with default options. After mapping, reads were sorted by samtools version 1.8 [14]. We performed Genome Analysis Toolkit (GATK, v4.0.7.0) [15] AddOrReplaceRead-Groups command for adding read group information, MarkDuplicates command for removing duplicated reads, BaseRecalibrator and ApplyBQSR commands for correcting realignment. Data processing summary statistics are given in Supplementary Table 1. Then, somatic single nucleotide variations (SNVs) and insertions/deletions were called using the GATK4-Mutect2 [16] pipeline. Filter-based annotation in ANNOVAR [17] was used for variant annotations. In addition, we calculated the copy number alterations using EXCAVATOR2 (v1.1.2) [18]. All statistical analysis and visualizations were performed using R version 3.6.1.

Results

Clinical and histopathological features of samples
We analyzed 132 lung cancer patients whose tumor and matched normal tissues were available and the PDX tumor samples were successfully harvested. The pathophysiological information of patients is summarized in Table 1. Briefly, we had 54 adenocarcinoma cases (41%), 48 squamous cell carcinoma cases (36%), four large cell carcinoma cases (3%), and 26 unclassifiable adenocarcinoma cases (20%). Our patient cohort was enriched with male (66%), smokers (64%), early stages (50%), non-recurrent (61%), and primary (62%) patients. In accordance with the previous reports [19], the success rate of establishing PDX mouse models was higher in squamous cell carcinoma than in adenocarcinoma. Histopathological examination of patient tumor and PDX tumor tissues concluded that tumors were consistent between patient and PDX mouse in 97 cases (73.5%). The discrepancy in histology maybe presumably ascribed to the lymphomagenesis that had been reported to occur frequently in NSG or NOG mice transplanted with Epstein-Barr virus infected tumor tissues [20,21]. This concordance and discrepancy were further investigated by comparing the mutation and copy number profiles between patient and PDX mouse tumors.

Statistics in mapping and variant calling procedure
We performed WES on tumor, normal, and PDX samples of 132 lung cancer patients, with a mean coverage of 30×. In order to compare the somatic mutation profiles of patient and PDX mouse
would be replaced with the mouse stromal cells during engraftment of the patient’s tumor tissue into immunodeficient mice [22]. As a result, tumor cells are enriched in the PDX mouse tumors, which leads to more somatic mutations in variant calling. To test this hypothesis, we examined the variant allele frequencies (VAFs) of major cancer genes (TP53, KRAS, and PIK3CA) identified in both patient and PDX tumors (Fig. 1C). All three genes showed that the VAFs of PDX mouse tumors are larger than those of patient tumors, which supports our hypothesis of clonal enrichment in PDX mouse tumors. However, the extent of clonal selection pressure varied for different genes. Interestingly, the VAFs of TP53 gene were close to 100% in PDX mouse tumors. The biological meaning of this observation warrants further studies.

Comparison of somatic mutations between patient and PDX mouse tumors

Next, we examined how well the mutations identified in patients were reproduced in PDXs. In all 132 samples, it was found that 63% of the exonic mutations in patients were also identified in PDXs on average (Fig. 2A). The portion of common mutations, however, varied tremendously from 0% to 98%. Low rates of common mutations were mostly observed in cases where patient and PDX tumors showed different pathology in histological analysis. Considering 92 cases with consistent histology, 78% of the exonic mutations in patients were also identified in PDXs on average.

We have also examined the overlap of functional mutations, including non-synonymous SNVs, stop-gain mutations, stop-loss mutations, frameshift insertions, and frameshift deletions, between patient and PDX mouse tumors using 517 cancer-related genes curated from OncoKB and other literatures (Fig. 2B) [23-25]. The most commonly mutated gene was TP53. For cases with consistent histopathological result, 78% of the exonic mutations in patients were also identified in PDXs on average.

Table 1. Clinical information of 132 lung cancer patients

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>No. (%) (n = 132)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>44 (34)</td>
</tr>
<tr>
<td>Male</td>
<td>88 (66)</td>
</tr>
<tr>
<td>Age (yr), median</td>
<td>65</td>
</tr>
<tr>
<td>Smoking status</td>
<td></td>
</tr>
<tr>
<td>Nonsmoker</td>
<td>48 (36)</td>
</tr>
<tr>
<td>Smoker</td>
<td>84 (64)</td>
</tr>
<tr>
<td>Clinical stage</td>
<td></td>
</tr>
<tr>
<td>Early stage (I–II)</td>
<td>66 (50)</td>
</tr>
<tr>
<td>Late stage (III–IV)</td>
<td>25 (19)</td>
</tr>
<tr>
<td>N/A</td>
<td>41 (31)</td>
</tr>
<tr>
<td>Subtype</td>
<td></td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>54 (41)</td>
</tr>
<tr>
<td>Squamous cell carcinoma</td>
<td>48 (36)</td>
</tr>
<tr>
<td>Large cell carcinoma</td>
<td>4 (3)</td>
</tr>
<tr>
<td>Unclassified</td>
<td>26 (20)</td>
</tr>
<tr>
<td>Recurrent</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>52 (39)</td>
</tr>
<tr>
<td>No</td>
<td>80 (61)</td>
</tr>
<tr>
<td>Metastasis</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>50 (38)</td>
</tr>
<tr>
<td>No</td>
<td>82 (62)</td>
</tr>
<tr>
<td>Death</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>33 (25)</td>
</tr>
<tr>
<td>No</td>
<td>99 (75)</td>
</tr>
</tbody>
</table>

N/A, not available.

tumors, it is essential to test the reliability of mutation calls in the PDX mouse tumors because human stromal cells are replaced with the mouse stromal cells during engraftment. Thus, we checked the portion of mouse-originated reads from the WES data of PDX tumors using the Xenome software to separate the human and mouse reads specifically. A summary of the Xenome alignment results is provided in Supplementary Table 2. The median portion of human reads was 95.1% and that of mouse reads was below 5% except a few cases (Fig. 1A). This implied that the PDX mouse tumors contained a sufficient amount of human cells. Thus we used the well-known MuTect2 algorithm after BWA-MEM mapping to call somatic mutations in PDX mouse tumors.

The mutation rates were significantly higher in PDX mouse tumors than in patient tumors ($p = 7.36e-06$). The median values of exonic mutations in PDX and patient tumors were 136 and 102, respectively (Fig. 1B). Since we removed the mouse-originated reads before calling somatic mutations, the difference can be attributed to clonal selection and evolutionary processes in establishing PDX mouse tumors. For example, human stromal cells
The main advantage of PDX tumor model is that it is possible to mouse tumors mutation and copy number profiles over passages of PDX examination. Thus, the mutation profile plays complementary roles to histological result is inconsistent, but when the histological examination gave a effect, but error-prone procedure. In conclusion, it is not necessary to common mutation was found if the patient and PDX tumors were truly of the same histology. Histological examination is not a perfect, but error-prone procedure. In conclusion, it is not necessary to to compare the mutation profiles if the histological examination result is inconsistent, but when the histological examination gave a good result, comparing mutation profiles can be useful in determining if the PDX tumor is truly identical to the patient tumor. Thus, the mutation profile plays complementary roles to histological examination.

**Mutation and copy number profiles over passages of PDX mouse tumors**

The main advantage of PDX tumor model is that it is possible to amplify the amount of tumor cells by engrafting tumor tissue into other immunodeficient mice. Molecular characteristics are usually expected to be preserved in the passage process, but detailed examination at the genome scale is quite rare.

We examined the mutation and copy number profiles for six cases where PDX tumor samples were available over several passages/generations (Fig. 3). We had four histologically consistent cases and two inconsistent cases. Both the mutation and copy number profiles were well reproduced throughout many generations in four good cases. Interestingly, the copy numbers showed much larger changes than somatic mutations especially between patient tumor and the first generation of xenograft tumor (Fig. 3B). It seems that PDX tumors harbor a number of new copy number losses that were maintained over many passages, which again implied that the clonal selection occurred in establishing the PDX mouse tumors.

**Discussion**

With the recent advances in anti-cancer drugs from unspecified cytotoxic agents to targeted therapy or immunotherapy, a better preclinical model that reflects the characteristics of each patient is required to realize the precision medicine in cancer. The PDX mouse model has emerged as a valuable preclinical model to overcome the
limitations of in vitro cell lines. Drug development and targeted therapy studies using PDX models reported that their responses to treatments are consistent with the clinical outcomes of actual patients. However, co-clinical studies are extremely rare and the fidelity of PDX mouse to patient tumors is not well defined.

In this study, we have found that PDX mouse models recapitulate the genetic characteristics of patients quite well. Although it is difficult to assume that PDX models perfectly represent patients’ genetic profiles due to the effects of clonal selection or evolution, much of the alterations identified in the patient were identified in PDX tumors as well. In addition, these alterations have been maintained for generations. Importantly, several PDX models have actionable alterations that can help drug development targeting those aberrations.

Although we confirmed that major portion of somatic mutations and copy number alterations were maintained in PDX establishment, we also observed many novel mutations and copy number alterations in PDX mouse tumors. The clonal selection and evolution are the main causes, resulting in different VAFs for several driver mutations and novel copy number losses. We also observed many additional somatic mutations emerge as a result of clonal enrichment. Understanding the details of clonal evolution should be important in interpretation of treatment response using PDX mouse models.

To the best of our knowledge, this is the largest scale of PDX study in lung cancer with matching trio samples. Over 100 cases of PDX biobank data were produced, even though there are cases where the histopathology of the patient and matched PDX are not consistent. By comparing the mutation and copy number profiles of patient and PDX tumors, we were able to show that the molecular characteristics are mostly in agreement with the histological results. But several cases were identified that molecular characteristics did not agree even though the histological examination results were identical in patient and PDX tumors. This highlights the complementary roles of molecular profiling in evaluating the PDX mice as a surrogate model in preclinical tests.
Fig. 3. Mutation and copy number profiles over the passage in patient-derived xenograft (PDX) mouse models. (A) Somatic mutations over PDX mouse passages. (B) Copy number profiles over PDX mouse passages. The circled numbers indicating patient cases are consistent with those of Fig. 2B. SNV, single nucleotide variation.

**ORCID**

Jaewon Kim: https://orcid.org/0000-0003-3418-4875  
Hwanseok Rhee: https://orcid.org/0000-0001-5118-9534  
Jhingook Kim: https://orcid.org/0000-0002-3828-0453  
Sanghyuk Lee: https://orcid.org/0000-0001-9230-7461

**Authors’ Contribution**

Conceptualization: Jhingook Kim, SL. Data curation: Jaewon Kim, HR. Formal analysis: Jaewon Kim. Funding acquisition: SL. Methodology: Jaewon Kim, HR. Writing - original draft: Jaewon Kim. Writing - review & editing: SL.

**Conflicts of Interest**

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

**Acknowledgments**

This work was supported by the grants from the Technology Innov-
Supplementary Materials

Supplementary data including two tables can be found with this article at https://doi.org/10.5808/GI.2020.18.1.e3.

References


Strong concordance between RNA structural and single nucleotide variants identified via next generation sequencing techniques in primary pediatric leukemia and patient-derived xenograft samples

Sonali P. Barwe1, Anilkumar Gopalakrisnapillai1, Nitin Mahajan2, Todd E. Druley2, E. Anders Kolb1, Erin L. Crowgey1*

1Alfred I. duPont Hospital for Children, Wilmington, DE 19803, USA
2Washington University School of Medicine, St. Louis, MO 63110, USA

Acute leukemia represents the most common pediatric malignancy comprising diverse subtypes with varying prognosis and treatment outcomes. New and targeted treatment options are warranted for this disease. Patient-derived xenograft (PDX) models are increasingly being used for preclinical testing of novel treatment modalities. A novel approach involving targeted error-corrected RNA sequencing using ArcherDX HemeV2 kit was employed to compare 25 primary pediatric acute leukemia samples and their corresponding PDX samples. A comparison of the primary samples and PDX samples revealed a high concordance between single nucleotide variants and gene fusions whereas other complex structural variants were not as consistent. The presence of gene fusions representing the major driver mutations at similar allelic frequencies in PDX samples compared to primary samples and over multiple passages confirms the utility of PDX models for preclinical drug testing. Characterization and tracking of these novel cryptic fusions and exonal variants in PDX models is critical in assessing response to potential new therapies.

Keywords: error-corrected sequencing, genomics, patient derived xenograft models, pediatric cancers, structural variants

Introduction

Genomic characterization of the somatic landscape is essential for the robust clinical evaluation and classification of pediatric leukemias [1]. Somatic variants can inform both diagnosis and prognostication, as well as guide therapy decisions [2]. The development and validation of new targeted therapies for pediatric leukemias is dependent on the availability of pre-clinical models capable of recapitulation of the disease. Patient-derived orthotopic xenograft models (PDX) are routinely used in disease modeling for preclinical drug evaluation [3]. Although several studies have been conducted to understand the stability and suitability of PDX models, the majority of these efforts have focused on adult-derived leukemias and the characterization of single nucleotide variants (SNVs) [4].

Chromosomal rearrangements generating gene fusions and other structural variants
(StVs) are more common in pediatric malignancies compared to adults [5]. These StVs and SNVs have demonstrated a different landscape for diagnostic, prognostic, and therapeutic value. Of note, pediatric leukemias are genomically heterogeneous and require a broad spectrum of molecular biology techniques to fully characterize. Additionally, StVs are difficult to identify via short read DNA-seq approaches, and recent research has demonstrated the power and utility of identifying SNVs in RNA molecules [6].

Acute lymphoblastic leukemia (ALL) is the most common type of cancer in children and adolescents. ALL represents 20% of all cancers diagnosed in individuals with less than 20 years of age [7]. In general, survival in ALL has improved significantly over the past 40 years with more than 90% of patients now surviving. Acute myeloid leukemia (AML) is the second most common type of leukemia diagnosed in children. AML has an overall survival rate that is less than 65%. In all children with AML, and many with ALL, survival comes at the expense of intensive chemotherapy. New strategies are needed, as are preclinical models that reflect the clinical disease.

The goal of this study was to characterize complex genomic variants in pediatric leukemias and describe and monitor these variants in preclinical PDX models in comparison with the primary samples. The ability to track complex genomic lesions in primary samples and across passage in PDX lines is essential in ensuring that the model can be used for biologic and therapeutic modeling. RNA next generation sequencing (NGS) techniques enable a sensitive and broad approach for analyzing complex genomic lesions and identifying clinically relevant novel somatic mutations associated with pediatric leukemias.

### Methods

#### Patient samples and consent

All samples used in this study were procured by the Nemours Biobank following written informed consent. For majority of samples, leukemic cells were isolated from human bone marrow aspirates with the exception of NTPL-59 and NTPL-109, which were isolated from apheresis products by Ficoll density gradient centrifugation and provided to us under an Institutional Review Board approved protocol (Nemours Office of Human Subjects Protection IRB# 267207). Summary of the subject’s characteristics are presented in Table 1.

#### Generation of PDX models

PDX models were generated as described previously [8] using a protocol approved by the Nemours Institutional Animal Care and Use Committee. Leukemic cells from patient samples were injected into immune-deficient NSG-B2m mice (stock no. 010636, Jackson Laboratories, Bar Harbor, ME, USA) via the tail vein. Disease progression was examined by determination of the percentage of human leukemic cells in mouse peripheral blood by flow cytometry. Mice were closely monitored for experimental endpoints such as increased leukemic burden, weight loss greater than 20% body weight, hunched back, and impaired mobility. Mice meeting end point criteria were euthanized using a method consistent with the guidelines of the American Veterinary Medical Association. Leukemic cells were isolated from the femurs and spleen post euthanasia and used for serial transplantation in a new cohort of mice. Biouathentication and validation of PDX sample with matching primary sample was performed by subjecting the DNA samples to AmpFISTR Identifiler PCR Amplification Kit (Applied Biosystems, Foster City, CA, USA).

#### Error-corrected sequencing library preparation and sequencing

To optimize detection of structural and copy number variants in RNA we prepared RNA–error-corrected sequencing libraries using the ArcherDX (Boulder, CO, USA) FusionPlex HemeV2 Kit (catalog no. AB0012) per manufacturer’s protocols. Total RNA was extracted using RNeasy Mini Kit (Qiagen, Hilden, Germany). Nucleic acid quantity and quality was then assessed using the Agilent (Santa Clara, CA, USA) TapeStation 4200 following the manufacturer’s protocol and using the High Sensitivity RNA ScreenTape (catalog no. 5067-5579). cDNA was made from 50 ng of RNA using the QIAseq kit. Each library was sequenced on the Il-

### Table 1. Summary of leukemic samples utilized

<table>
<thead>
<tr>
<th>Patient characteristic</th>
<th>AML</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Age (yr), median (range)</td>
<td>10 (1.5-14)</td>
<td>5.5 (1-16)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>40</td>
<td>55</td>
</tr>
<tr>
<td>Female</td>
<td>60</td>
<td>45</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>60</td>
<td>35</td>
</tr>
<tr>
<td>African American</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>Hispanic</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Samples collected at diagnosis</td>
<td>80</td>
<td>95</td>
</tr>
<tr>
<td>Cytogenetically normal (by karyotype analysis)</td>
<td>0</td>
<td>55</td>
</tr>
<tr>
<td>Bone marrow origin</td>
<td>100</td>
<td>90</td>
</tr>
<tr>
<td>Peripheral blood origin</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Average leukemic blast percentage</td>
<td>76</td>
<td>78</td>
</tr>
</tbody>
</table>

Values are presented as percentage unless otherwise indicated. AML, acute myeloid leukemia; ALL, acute lymphoblastic leukemia.
lumina NextSeq platform (San Diego, CA, USA). The gene fusion data produced by the Archer panel was initially correlated with diagnostic fluorescence in situ hybridization data available for each primary sample.

**Bioinformatics**

The data was processed via ArcherDX Analysis platform (v5.1.3), hosted in the cloud by Amazon Web Services, including fastq trimming, read deduplication, genome alignment, and variant detection and annotation. The analysis pipeline contains the following applications: ABRA [9], bamaddrg, bftools, bedtools, blast [10], bowtie2 [11], bwa, EMBOSS, fastqc, freeBayes [12], Lofreq [13], MiXCR [14], Muscle, samtools, VEP [15], Velvet [16], HTSeq [17], complete-striped-smith-waterman-library, JBrowse [18], JQuery DataTables, Django Solo, and plot.js.

Fastq files were analyzed via fastqc for library quality, and error corrected reads (hamming distance of 2) were aligned to the genome build hg19 using BWA and bowtie2, and alignment files were processed via GATK best practices [19]. SNVs and short In-Dels (±20 bp) were detected from the genomic alignments (forced reference mapping) by freeBayes and Lofreq, whereas large structural variants and cryptic fusions were detected via de novo assembly approaches. A minimum of three reads per unique molecular barcode (UMI) was required for the downstream process of de-duplication and error-correction [20]. Variants were filtered based on depth of error-corrected sequencing bins, minimum of 3, that supported the call. All regions in which variants called required a total read depth > 100×, and a minimal base quality score (phred) of 20 was applied. The ExAC database was used to annotate common variants.

Variant allele frequencies (VAF) were calculated for SNVs based on number reads mapped to that location supporting the alternative allele versus the total number of reads mapped to that genome location. VAFs for StVs are calculated by analyzing the number of reads supporting the wild type sequence/junction, compared to the number of reads supporting the novel junction. R statistics was used for making scatter plots, specifically ggplot2 [21]. Alignment files (.bam) were visualized via integrative genome browser (IGV). The fastq data is publicly available via short read archives under the number reads mapped to that genome location (will add upon acceptance of manuscript).

**Results**

**Comparison of RNA StVs and SNVs between primary and PDX AML samples**

To determine the concordance of RNA variants between primary and PDX samples for pediatric AML, a targeted RNA sequencing panel approach (HemeV2; ArcherDx) was utilized. In this report, we analyzed 5 AML primary-PDX sample pairs, and in total 31 allelic specific SNVs were identified with the following distribution: 1 frameshift, 11 missense, 2 splice region and 17 untranslated region (UTR) variants (Supplementary Table 1). Five UTR variants were present at a VAF of 1 in both primary and PDX AML samples. The absolute change in VAFs between primary and PDX samples was less than 0.2 for 27 SNVs. A few variants increased in VAF in the PDX (MYC, CDKN2A, and NOTCH1), other variants reduced in VAF in PDX samples (CCND3 and ABL2) (Fig. 1A, Supplementary Table 1).

VAFs for all RNA StVs including gene fusions and alternative exon usage variants were graphed between the primary and PDX AML samples and results are displayed (Fig. 1B). Four unique gene fusions (KMT2A-MLLT1, KMT2A-MLLT3, NUP-98-NSD1, and the reciprocal NSD1-NUP98) were identified in the primary AML samples and PDX samples. Additionally, 5 exon duplications/deletions were identified in CEBPA and IRF4 (Supplementary Table 2).

Multiple retained introns (n = 14) were identified in the 5 primary and PDX AML samples in the following genes: ZCCHC7, ABL1, JAK2, IRF8, TAL1, CEBPG, ETV6, KMT2A, MLLT10, KLF2, and PRDM16 (Supplementary Table 2). The SNVs were more concordant between primary and PDX samples compared to StVs (Pearson correlation coefficient, 0.91; p = 5.12e-13 and 0.43; p = 0.036 respectively). Among the StVs, fusions were identified at similar VAFs in primary and PDX samples, whereas the alternative exon usage variants showed greater variability. Interestingly, the 2 AML samples with KMT2A gene rearrangements (NTPL-146 and NTPL-377) showed higher level of concordance between VAFs for StVs as well as SNVs.

**Comparison of RNA StVs and SNVs between primary and PDX T-ALL and B-ALL samples**

To determine the concordance of RNA variants between primary and PDX samples for pediatric ALL, samples target RNA sequencing approach was utilized. The correlation coefficients of VAF between primary and PDX T-cell ALL (T-ALL) samples identified across 3 primary and PDX T-ALL samples were similar between SNVs and StVs (Pearson correlation coefficient, 0.88; p = 6.12e-10 and 0.73; p = 0.003 respectively) (Fig. 2). In total, 25 allelic specific RNA SNVs were identified in the primary and PDX T-ALL samples: 3 frameshift, 8 missense, and 14 UTR variants. Six UTR variants had VAF = 1 in primary and PDX T-ALL samples. Three variants had absolute VAFs greater than 0.25; 1 of these SNVs (NOTCH1 frameshift variant) reduced in VAF in PDX samples, while 2 (CEBPA missense variants) showed gains in PDX samples.
Allelic specific single nucleotide variants in AML samples

Structural RNA variants detected in AML samples

STIL-TAL1 gene fusion detected in T-ALL samples

Fig. 1. Summary of primary and xenograft RNA variants in acute myeloid leukemia (AML). (A) Allelic specific single nucleotide variants. Variant allele frequency (VAF) at time of diagnosis, x-axis is plotted versus the VAF in the xenograft model, y-axis. (B) Structural RNA variants. VAF at time of diagnosis, x-axis is plotted versus the VAF in the xenograft model, y-axis. PDX, patient-derived xenograft.

Allelic specific single nucleotide variants in T-ALL samples

Structural RNA variants detected in T-ALL samples

STIL-TAL1 gene fusion identified in 2 of the T-ALL samples.

Fig. 2. Summary of primary and xenograft RNA variants in T-cell acute lymphoblastic leukemia (T-ALL). (A) Allelic specific single nucleotide variants. Variant allele frequency (VAF) at time of diagnosis, x-axis is plotted versus the variant allele frequency in the xenograft model, y-axis. (B) Structural RNA variants. VAF at time of diagnosis, x-axis is plotted versus the VAF in the xenograft model, y-axis. (C) STIL-TAL1 gene fusion identified in 2 of the T-ALL samples. PDX, patient-derived xenograft.
samples. NTPL-454 had a strong correlation between SNV VAFs in the primary and PDX models (Fig. 2A; green line), whereas NTPL-59 and NTPL-300 were not as consistent with VAFs for SNVs.

In total 14 StVs were identified in the primary and PDX models for T-ALL samples; 4 unique fusions (STIL-TAL1, SPTAN1-ABL1), 7 retained introns (EIF4A, IRF8, KMT2A, NF1, SETD2), and 3 molecules with exon duplications (BCL11B and ZCCH7). NTPL-300 was the most concordant for VAF of StVs in primary and PDX T-ALL samples (Fig. 2B). Of interest, 2 of the T-ALL samples had a STIL-TAL1 gene fusion, which was recently published as a potential driver/founder event [22].

The correlation between VAF from primary to PDX samples was analyzed for RNA StVs and SNVs in 17 B-cell ALL (B-ALL) samples. In total 114 RNA SNVs were identified in the primary and PDX B-ALL samples, and of those variants 4 were frameshift, 25 missense, 5 splice region, 2 stop gained and the rest were UTR variants (Supplementary Table 1). Twenty-two UTR variants (RUNX1, IKZF3, CHIC2, CCND2, BCL2) were detected at identical VAF of 1 in primary and PDX B-ALL samples. Five variants (4.4%) had VAFs greater than 0.25; CHID1, ABL2 UTR variants showed decreased VAF, and BCR, CCND2, NOTCH1 SNVs showed increased VAF in PDX samples.

The correlation between SNV VAFs from primary to PDX B-ALL samples was higher than the correlation between StV VAFs (Pearson correlation coefficient, 0.93; p = 2.2e-16 and 0.5; p = 9.5e-8, respectively) (Fig. 3A, B). Eight of 17 samples possessed a gene fusion (BCR-ABL1, ETV6-RUNX1, P2RY8-CRLF2, RUNX1-MKL1, TCF3-HLF, TCF3-PBX1). The VAFs for StVs, especially the alternate exon usage variants, were more variable in these samples, similar to AML samples. Interestingly, 15 out of 18 of the B-ALL samples had a retained intron in ZCCHC7 involving intron 2, which was persistent in PDX samples (Fig. 3B). Two AML and 1 T-ALL sample also showed a similar retained intron variant (Supplementary Table 2). ZCCHC7 intron 2 has been mapped to hotspot for breakpoints in B-ALL [23].

Discussion

Sequencing of primary acute leukemia patient samples and matching PDX samples showed concordance between the detected variants and their allelic frequencies for the majority of variants tested. The percentage of all variants with absolute delta VAFs < 0.2 was 86.7%. This percentage was higher in SNVs (93.6%) compared to StVs (79.6%) across all primary and PDX samples analyzed. Among the different categories of StVs, the allelic frequencies of fusion
genes, which are considered to be driver mutations, matched most consistently between the primary and PDX samples (Fig. 5). Our data validate this novel sequencing approach for detection and tracking of diverse variants in primary leukemic samples and corresponding PDX lines.

We identified several SNVs, but no StVs, with sustained VAF = 1 in primary and PDX samples across all leukemia subtypes. These SNVs in genes ABL1, BCL2, CCND2, CHIC2, IKZF3, RUNX1, and MECOM, likely represent the germline mutations. Several germline variants, including UTR variants have been shown to be associated with disposition to hematological malignancies [24]. Future characterization of these variants will determine the relevance of these germline UTR variants.

Retained intron variants were detected in all samples except NTPL-59. Retention of introns serves as another mode of regulation of gene expression [25]. Alternative splicing of multi-exon genes in patients with AML compared to normal CD34+ cells has been observed [26]. Such alternative exon usage variants were associated with oncogene expression and drug resistance [27]. Further work is required to understand the biological and clinical significance of alternative exon usage variants.

As we have shown previously, error-correction via the introduc-

---

**Fig. 4.** Waterfall graph for single nucleotide variants (SNVs) and structural variants (StVs) detected in B-cell acute lymphoblastic leukemia samples. Genes with either a coding SNV or StV were plotted (y-axis) per sample (x-axis). Mutations are colored based on type.
A nucleic acid-specific UMI allows the removal of NGS errors, retaining only true mutations and significantly improving the sensitivity of NGS [28-30]. In this study, we paired the error-correction strategy with anchored-multiplexed PCR (AMP) chemistry for the quantitative detection of complex structural RNA variants. Recently Benayed et al. [31] published an RNA sequencing approach similar to the one outlined in this manuscript (MSK targeted RNA panel using ArcherDx), and demonstrated that their MSK-IMPACT DNA panel missed cancer-related and targetable mutations in greater than 15% of lung cancer patients. They leveraged ArcherDx FusionPlex technology (identical to our approach) to identify these cases. Additionally, several recent studies have demonstrated the use of AMP technology (ArcherDx) for identifying rare and complex structural variants in pediatric cancers [32,33].

Taken together, advanced sequencing techniques are required to accurately detect and annotate complex StVs that are commonly associated with pediatric leukemias. Such complex variants, including StVs, are not detectable using DNA and short read sequencing technology such as Illumina sequencing platform. Additionally, the RNA molecules that are generated from these complex genomic rearrangements can be difficult to capture. Using an RNA sequencing approach with AMP technology and short read sequencing platform described in this study, pediatric PDX models could be appropriately characterized and validated for concordance of somatic mutations with respect to primary samples. Such analysis is not feasible using standard DNA sequencing techniques. This is one of the first reports to describe pediatric PDX samples using an RNA sequencing approach.

ORCID
Sonali P. Barwe: https://orcid.org/0000-0003-4162-3004
Anilkumar Gopalakrisnapillai: https://orcid.org/0000-0002-0465-578X
Nitin Mahajan: https://orcid.org/0000-0002-0907-6410
Todd E. Druley: https://orcid.org/0000-0002-3245-7561
E. Anders Kolb: https://orcid.org/0000-0003-2854-9014
Erin L. Crowgey: https://orcid.org/0000-0002-2037-0389

Fig. 5. Comparison of variant allele frequencies of structural variants between primary bone marrow samples (x-axis) and matched xenograft sample (y-axis). (A) The variant allele frequencies for all gene fusions were plotted between the primary and xenograft model ($R^2 = 0.7634$). (B) The variant allele frequencies for all retained introns were plotted between the primary and xenograft model ($R^2 = 0.2906$). (C) The variant allele frequencies for all exon deletion were plotted between the primary and xenograft model ($R^2 = 0.0078$). (D) The variant allele frequencies for all exon duplications were plotted between the primary and xenograft model ($R^2 = 0.0118$).
**Authors’ Contribution**

Conceptualization: SPB, AG, EAK, ELC. Data curation: SPB, AG, NM, TED. Formal analysis: ELC. Writing - original draft: SPB, AG, NM, TED, EA, ELC. Writing - review & editing: SPB, AG, NM, TED, EA, ELC.

**Conflicts of Interest**

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

**Acknowledgments**

The authors would like to thank the Nemours Center for Cancer and Blood Disorders, Nemours Biobank, and the Nemours Biomedical Research Department for supporting this work. This work was supported by the NIH NCI CA211711-01 (PI Druley), Leukemia Research Foundation of Delaware (PI Kolb), and B+ Foundation (PI Barwe).

**Supplementary Materials**

Supplementary data including two tables can be found with this article online at https://doi.org/10.5808/2020.18.1.e6.

**References**


Associations between an \textit{MDM2} gene polymorphism and ulcerative colitis by ARMS–PCR

Mahsa Sadat Hashemi Doulabi$^1$, Reza Goleyjani Moghaddam$^2$, Ali Salehzadeh$^3$

$^1$Young Researchers and Elite Club, Tonekabon Branch, Islamic Azad University, Tonekabon 46841-61167, Iran
$^2$Department of Biology, Tonekabon Branch, Islamic Azad University, Tonekabon 46841-61167, Iran
$^3$Department of Biology, Rasht Branch, Islamic Azad University, Rasht 41476-54919, Iran

Ulcerative colitis is a form of inflammatory bowel disease characterized by chronic inflammation of the colon and rectum. The abnormal lesions in the digestive system caused by ulcerative colitis and intermittent colitis are of major clinical importance. \textit{MDM2} is a phospho-protein that functions as a ubiquitin ligase for p53. Recently, a T>G substitution in the promoter of the \textit{MDM2} gene ($\text{rs309}$) has been identified. In this case–control study, 174 ulcerative colitis biopsy samples and 82 control samples were collected from colonoscopy centers, hospitals, and clinics in Mazandaran and Gilan Provinces in Iran from October 2014 to May 2015. This \textit{MDM2} polymorphism was investigated in DNA samples (extracted from biopsy samples) by amplification-refractory mutation system polymerase chain reaction. The mean age of patients with ulcerative colitis was 46.5 years (range, 28 to 69 years) and that of control individuals was 45.3 years (range, 26 to 71 years). Seventy-eight patients (44.82\%) were men and 96 (55.18\%) were women. The distribution of the TT, TG, and GG genotypes was 17.93\%, 27.59\%, and 34.48\%, respectively, in the ulcerative colitis patients and 31.70\%, 24.40\%, and 43.90\%, respectively, in the control individuals (odds ratio of GG for ulcerative colitis, 7.142; 95\% confidence interval, 2.400 to 9.542; $p = 0.001$). It was found that a single-nucleotide polymorphism at $\text{rs309}$ in the \textit{MDM2} gene was associated with ulcerative colitis. A direct relationship was found between age and ulcerative colitis, while no relationship was found with sex. This finding is of note because the occurrence of intestinal inflammation and subsequent ulcers can precede the development of cancer.

\textbf{Keywords:} inflammatory bowel disease, malignant, \textit{MDM2}, polymorphism, ulcerative colitis

\section*{Introduction}

Inflammatory bowel disease (IBD) is a complex disease that results from an inappropriate immune system response to intestinal bacteria \cite{1}. IBD is generally subdivided into Crohn disease and ulcerative colitis. Ulcerative colitis is characterized by chronic inflammation of the colon and rectum, whereas Crohn disease can affect the entire digestive system. The role of genetic factors in these conditions was first raised by epidemiological studies that reported familial associations of these diseases \cite{2}.

Ulcerative colitis is a chronic inflammatory disease that affects the entire colon \cite{3}. In ulcerative colitis, inflammation is classically confined to the colon, is typically persistent, and begins in the rectum \cite{4}. The presence of abnormal gastrointestinal lesions in pa-
Patients with ulcerative colitis and intermittent colitis is of major clinical importance, and many gastroenterologists therefore perform upper endoscopy to obtain a definitive diagnosis in patients with IBD [5].

The risk factors for ulcerative colitis appear to be related to changes in the intestinal microbiome or disorders in the intestinal mucosa [6,7]. Intestinal infections, non-steroidal anti-inflammatory drugs, and antibiotics all contribute to the development of IBD [7,8].

MDM2 is a phospho-protein and a ubiquitin ligase for p53 that is responsible for inhibiting p53 activity and promoting its destruction [9]. Recently, a T > G substitution in the MDM2 gene promoter (rs309) has been identified. This substitution is associated with increased expression of MDM2, which accelerates the formation of several types of tumors, resulting in a tendency for them to occur at a younger age [10]. These findings underscore the importance of this polymorphism as an important factor that can affect the frequency of cancer in a population, the age of cancer in individuals, and individuals’ responses to cancer treatment [11].

The rs309 locus in the second promoter region of the MDM2 gene, which is associated with increased expression of this gene, may have potential as a molecular target for cancer susceptibility and as a suitable tumor marker. If a polymorphism is present at the rs309 position of the MDM2 gene (i.e., a T > G conversion in this promoter region), the binding affinity of the SP1 transcription factor activator to this region is significantly increased, which increases MDM2 gene expression. This means that an individual with a TT genotype for this polymorphism has a baseline expression level of the MDM2 gene, but in an individual with the TG genotype, the G allele increases MDM2 gene expression, and this expression is even more dramatically increased in individuals with a GG phenotype. Due to the inhibitory effect of MDM2 on p53, an increase in MDM2 expression leads to a decrease in the intracellular amount of p53 protein, which is a key regulator of the response to cellular damage. Under usual circumstances, levels of the p53 protein are increased 5- to 14-fold when cellular damage occurs, but they are reduced by 2 to 3 times if the G allele is present in the rs309 locus of MDM2, which leads to increased levels of the MDM2 protein [12]. The overall aim of this study was to investigate the rs309 polymorphism of MDM2 and its association with ulcerative colitis, and a secondary aim was to explore the association between this polymorphism and the risk of cancer.

Methods

Sample selection method

In this case-control study, 174 ulcerative colitis biopsy samples and 82 control samples were collected from colonoscopy centers, hospitals, and clinics in Mazandaran and Gilan Provinces, Iran from October 2014 to May 2015. Patients’ history, including age, sex, place of residence, and severity of illness was obtained, the diagnosis was confirmed by the treating physician, and a consent form and questionnaire were obtained. Then, in the colonoscopy procedure, some of the intestinal tissue was removed, transferred to sterile vials, and stored at –20°C until DNA extraction.

MDM2 rs309 polymorphism

The amplification-refractory mutation system polymerase chain reaction (ARMS-PCR) technique was used to study nucleotide changes in the MDM2 gene. In this technique, the reaction can be performed in a tube. In this study, the ARMS technique was used to determine the presence of the T > G point mutation with two primer pairs (Tables 1 and 2).

A proliferation fragment of 224 bp should be seen in all samples as an indicator of the accuracy of PCR. The expected results of ARMS-PCR in this study included the normal genotype (TT), as well as the TG and GG mutant genotypes. After PCR, the products were separated on an agarose gel by electrophoresis and then bands were observed using ultraviolet visualization (Fig. 1).

Statistical analysis

The statistical analysis was conducted using SPSS version 22 (IBM Corp., Armonk, NY, USA) and p-values of < 0.05 were considered to indicate statistical significance.

Table 1. The sequences of the primers used in this study

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Direction</th>
<th>Primer sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDM2(M)</td>
<td>Forward (1)</td>
<td>5’-GGGGCCCGGGCTGGGCCTTTG3’</td>
</tr>
<tr>
<td>MDM2(M)</td>
<td>Reverse (1)</td>
<td>5’-TGGCCCATGGAACCGCGGACCTT3’</td>
</tr>
<tr>
<td>MDM2(C)</td>
<td>Forward (2)</td>
<td>5’-GGCGCATCAGGCGGGGCGG3’</td>
</tr>
<tr>
<td>MDM2(C)</td>
<td>Reverse (2)</td>
<td>5’-ACCACATGACCATCGGACCTCGG3’</td>
</tr>
</tbody>
</table>

Table 2. Thermocycler program for MDM2 gene amplification

<table>
<thead>
<tr>
<th>No.</th>
<th>Stage</th>
<th>Temperature (°C)</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Initial denaturation</td>
<td>95</td>
<td>15 min</td>
</tr>
<tr>
<td>2</td>
<td>Denaturation</td>
<td>95</td>
<td>45 s</td>
</tr>
<tr>
<td>3</td>
<td>Annealing</td>
<td>64</td>
<td>45 s</td>
</tr>
<tr>
<td>4</td>
<td>Extension</td>
<td>72</td>
<td>1 min</td>
</tr>
<tr>
<td>5</td>
<td>Final extension</td>
<td>72</td>
<td>7 min</td>
</tr>
<tr>
<td>2-4</td>
<td>Cycles</td>
<td>35</td>
<td></td>
</tr>
</tbody>
</table>
The characteristics of the participants with ulcerative colitis and the control sample are summarized in Table 2. The mean age of the patients with ulcerative colitis was 46.5 years (range, 28 to 69 years) and that of the control individuals was 45.3 years (range, 26 to 71 years).

The presence of ulcerative colitis was significantly related with age (p < 0.05 [chi-square test]). However, it was not significantly related with sex (p > 0.05 [chi-square test]) (Table 3).

Table 3. Demographic characteristics of patients with ulcerative colitis and controls

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Case</th>
<th>Control</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (n = 256)</td>
<td>174</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>Age (yr)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \leq 50 )</td>
<td>110 (63.2)</td>
<td>48 (58.5)</td>
<td>0.006</td>
</tr>
<tr>
<td>&gt; 50</td>
<td>64 (36.8)</td>
<td>34 (41.5)</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>78 (44.8)</td>
<td>28 (34.1)</td>
<td>0.4</td>
</tr>
<tr>
<td>Female</td>
<td>96 (55.2)</td>
<td>54 (65.9)</td>
<td></td>
</tr>
</tbody>
</table>

Values are presented as number (%).

Association of the MDM2 polymorphism with the risk of ulcerative colitis

Table 4 shows the allelic frequency of the MDM2 rs309 polymorphism and the distribution of genotypes. The G allele was present in 48.27% of the ulcerative colitis patients and in 56.09% of the control individuals. The distribution of the MDM2 genotype in the ulcerative colitis patients was as follows: TT, 37.93%; TG, 27.59%; and GG, 34.48%. This was significantly different from the distribution in the control individuals (GG genotype: 34.48% vs. 43.90%; p < 0.05).

People with the GG phenotype of the MDM2 gene were more prone to ulcerative colitis (odds ratio, 7.142; 95% confidence interval, 2.400 to 9.542) than those with the TT genotype. The heterozygous genotype of this polymorphism did not show a clear relationship with the risk of ulcerative colitis, but we could nonetheless identify the G allele as risky (Table 4).

Discussion

In 2005, Sotamaa et al. [13] conducted a study on the MDM2 gene polymorphism at rs309 in patients with intestinal cancer that included 93 patients and 100 controls. The allelic frequencies of polymorphisms in the patients and control individuals showed Hardy-Weinberg equilibrium, and there was no significant relationship between age and occurrence of this polymorphism [13].

In our study, the polymorphism at this locus was investigated using ARMS-PCR, and a significant relationship was found between the presence of the GG genotype and the incidence of ulcerative colitis disease. A significant relationship was also found between age and ulcerative colitis, but no significant relationship was found for sex.

In 2014 study by Enokida et al. [14] on the rs309 MDM2 gene polymorphism in lung cancer, the distribution of genotypes showed no significant difference between lung cancer patients and controls (patients: TT, 20.1%; TG, 49.7%; and GG, 30.2%; con-
trols: TT, 21.7%; TG, 47.9%; and GG, 30.4%).

In our study, we found that there was a significant relationship between age and genotype, our results were inconsistent with some previous studies. The distribution of the MDM2 genotype in ulcerative colitis patients was as follows: TT, 37.93%; TG, 27.59%; and GG, 34.48%. This distribution was significantly different from that observed in controls (TT, 31.70%; TG, 24.40%; and GG, 43.90%).

Mutations in the P53 gene have been identified in most human cancers, as well as in its downstream signaling pathways, which are mediated by the P21 and MDM2 genes; therefore, proper functioning of all three genes is important for the normal function of cells. Consequently, when mutations in any of these genes disrupt critical signaling pathways, they can result in malignancies in human cells [11].

Many studies have found MDM2 gene mutations in the intestinal system to be associated with cancer [15]. In the current study, the overall aim was to investigate the polymorphism of this gene at rs309 and its association with ulcerative colitis, but a secondary goal was to explore the association between this polymorphism and the risk of cancer.

Since the distribution of the MDM2 polymorphism in individuals with ulcerative colitis was approximately the same as, it can be concluded that ulcerative colitis precedes the development of ulcers into malignancies.

In this study, it was found that the T > G polymorphism at the rs309 locus of the MDM2 gene was associated with ulcerative colitis through a statistical analysis. A direct relationship was found between age and ulcerative colitis, while no relationship was found for sex.

Since this gene is directly associated with carcinogenesis (mutation and loss of function), it can be concluded that the occurrence of intestinal inflammation and subsequent ulceration lays the groundwork for subsequent cancer.

**ORCID**

Mahsa Sadat Hashemi Doulabi: https://orcid.org/0000-0003-3161-8789

Reza Goleyjani Moghaddam: https://orcid.org/0000-0002-6832-0401

Ali Salehzadeh: https://orcid.org/0000-0003-4238-0999

**Authors’ Contribution**

Conceptualization: RGM, AS. Data curation: MSHD. Formal analysis: RGM, AS. Funding acquisition: MSHD. Methodology: RGM, AS, MSHD. Writing – original draft: MSHD, AS. Writing – review & editing: MSHD, AS.

**Conflicts of Interest**

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

**References**


Bioinformatics services for analyzing massive genomic datasets

Gunhwan Ko1, Pan-Gyu Kim2, Youngbum Cho2, Seongmunch Jeong2, Jae-Yoon Kim2, Kyoung Hyoun Kim2, Ho-Yeon Lee2, Jiyeon Han3, Namhee Yu3, Seokjin Ham5, Insoon Jang5, Byunghee Kang4, Sunguk Shin5, Lian Kim6, Seung-Won Lee7, Dougu Nam8, Jihyun F. Kim8, Namshin Kim2, Seon-Young Kim10, Sanghyuk Lee3, Tae-Young Roh4,11, Byungwook Lee1

1Korea Bioinformation Center (KOBIC), KRIBB, Daejeon 34141, Korea
2Genome Editing Research Center, KRIBB, Daejeon 34141, Korea
3Department of Bioinformation Science, Ewha Womans University, Seoul 03760, Korea
4Department of Life Sciences and Division of Integrative Biosciences & Biotechnology, Pohang University of Science & Technology (POSTECH), Pohang 37673, Korea
5Department of Systems, Biology Division of Life Sciences, and Institute for Life Science and Biotechnology, Yonsei University, Seoul 03722, Korea
6Bioposh Inc., Daejeon 34016, Korea
7SeqGenesis, Daejeon 34016, Korea
8School of Life Sciences, Ulsan National Institute of Science and Technology, Ulsan 44919, Korea
9Strategic Initiative for Microbiomes in Agriculture and Food, Yonsei University, Seoul 03722, Korea
10Genome Structure Research Center, KRIBB, Daejeon 34141, Korea
11SysGenLab Inc., Pohang 37613, Korea

The explosive growth of next-generation sequencing data has resulted in ultra-large-scale datasets and ensuing computational problems. In Korea, the amount of genomic data has been increasing rapidly in the recent years. Leveraging these big data requires researchers to use large-scale computational resources and analysis pipelines. A promising solution for addressing this computational challenge is cloud computing, where CPUs, memory, storage, and programs are accessible in the form of virtual machines. Here, we present a cloud computing-based system, Bio-Express, that provides user-friendly, cost-effective analysis of massive genomic datasets. Bio-Express is loaded with predefined multi-omics data analysis pipelines, which are divided into genome, transcriptome, epigenome, and metagenome pipelines. Users can employ predefined pipelines or create a new pipeline for analyzing their own omics data. We also developed several web-based services for facilitating downstream analysis of genome data. Bio-Express web service is freely available at https://www.bioexpress.re.kr/.

Keywords: analysis pipeline, cloud computing, genomic data, web server, workflow system

Introduction

Next-generation sequencing (NGS) technology has revolutionized the researches in biology and medicine during the last decade. It is routinely used in genomics field, and explosive growth of NGS data has resulted in ultra-large-scale datasets and various computa-
tional problems [1]. Public archives for sequencing data such as the Sequence Read Archive have grown rapidly and now exhibit a doubling time of 10–18 months [2]. In Korea, genomic data have been increasing rapidly in recent years. As of February 2020, approximately 277 TB of genomic data have been deposited in Korea Bioinformation Center (KOBIC) database.

It is not easy for typical researchers to analyze these massive genomic datasets. To obtain results from the data, researchers need to use high-performance computing (HPC) environments with sufficient storage space and CPU cores. In addition, the difficulties in creating complicated computational pipelines and maintaining software packages tend to overwhelm bench biologists and prevent them from attempting to analyze their own genomic data [3]. Despite the availability of a vast set of computational tools and methods for genomic data analysis in public, it is still challenging for a genomic researcher to organize these tools, integrate them into workable pipelines, find accessible computational platforms, configure the computing environment, and perform the actual analysis.

A promising solution to address this computational challenge is cloud computing, where CPUs, memory, and storage are accessible in the form of virtual machines [4]. The cloud computing, by definition, refers to the on-demand delivery of IT resources and applications via the Internet [5]. The Software as a Service (SaaS) cloud service for applications provides the perfect solution for the analysis of massive genomic datasets. SaaS is a method of software delivery in the IT field that allows data to be accessed from any device with an Internet connection and web browser. In recent years, cloud computing has rapidly emerged as a viable option for quickly and easily acquiring computational resources and pipelines for large-scale NGS data analyses [6].

The parallelism techniques in HPC infrastructure are used to process all the produced data in a feasible time [7]. Parallel computing is a type of computation in which many calculations or the execution of processes are carried out simultaneously. However, it is still challenging to integrate bioinformatics experiments with parallel techniques in the HPC environments. Many applications developed for the analysis of genomic data are either tools running only on a parallel platform, such as a MapReduce platform, or general-purpose (mainly Linux-based) programs. It is crucial to integrate these two types of platform-based applications on a single pipeline.

In this study, we present Bio-Express, a software package for deploying an on-demand computing cloud with minimal user intervention. The goal of Bio-Express is to provide a web-based analysis environment in which all genomic researchers, including those with limited or no programming knowledge, can easily analyze their own genomic data. The Bio-Express Graphic User Interface (GUI) provides a workflow editor in which users can simply use a predefined analysis pipeline or create a multistep analysis pipeline using preinstalled programs. The analysis pipelines on Bio-Express are exactly reproducible, and all analysis parameters and inputs are permanently recorded. Bio-Express makes it simple to perform a multistep analysis using simple drag and drop functionality. We also developed several web-based services for facilitating downstream analysis of genome data such as gene-set enrichment analysis.

**Methods**

**Hardware**

All runs of analysis pipelines on Bio-Express are performed on a cluster of five master nodes and 33 data nodes. The hardware system of Bio-Express consists of 800 core CPUs, 2 TB of memory, and 800 TB of disk storage in total. Each node has an Intel Xeon E502690 v2 3.0 GHz CPU, 96 GB of memory, and 28 TB of disk storage. The data node HDD configuration consists of the Hadoop Distributed File System and a solid-state drive (SSD) cache. The node manager handles the individual data nodes in a Hadoop cluster.

**Graphic User Interface (GUI)**

The GUI workspace of Bio-Express consists of eight panels: the user’s projects, the file explorer, the canvas, the analysis programs of the current pipeline, the program parameter settings, the pipeline panel, the program panel, and the job execution history (Fig. 1). Among these panels, the canvas is the most important panel and is used for creating and modifying workflows by arranging and connecting activities to drive processes. The canvas provides the working surface for creating new workflows or editing predefined ones. The canvas makes it simple to perform multistep analyses using drag and drop functionality.

**Pipelines**

The workflows, or analysis pipelines, in the canvas are commonly depicted as directed acyclical graphs, in which each of the vertices has a unique identifier and represents a task to be performed. Additionally, each of the tasks in a workflow can receive inputs and produce outputs. The outputs of a task can be directed through another task as an input. An edge between two vertices represents the channeling of an output from one task into another. The edges determine the logical sequence. A task can be executed once all of its inputs can be resolved. If one of user pipeline programs fails, users can select the program of the pipeline to view more detailed information on errors, and resume the whole pipeline from the failed program after fixing the errors.
The transfer of data

The bottleneck of cloud computing is the transfer of data into clouds. Therefore, we developed a fast file transfer tool, Gbox, for uploading massive genomic datasets to the cloud server from the user's local computer and for downloading the resulting files to the local. The client program of Gbox can be downloaded from the website and be installed on the user’s computer. Gbox has a file transfer at a rate of approximately 10 Gigabits per second, capable of dealing with big data over the web. Currently, Gbox has no file size limitations and storage limit on the Bio-Express cloud server.

Scalability

Scalability is one of the most attractive prospects of cloud computing and provides a useful safety net when a user’s needs or demands change. The resource and job manager of Bio-Express distributes computing resources to user jobs within a parallel computing infrastructure. Its aim is to satisfy user’s demands for computation and achieve a good performance in overall system’s utilization by efficiently assigning jobs to resources. The resource and job manager analyzes the application performance during runtime and predicts the demand for load balancing, i.e., when to add/remove resources or redistribute workload. Thus the scalability of Bio-Express improves the execution speed of job by efficient assignment of computing resources.

Results

The analysis pipelines can be divided into two types: predefined and user-created. As of February 2020, Bio-Express contains approximately 170 analysis tools and 57 predefined analysis pipelines for genome, transcriptome, epigenome, and metagenome data. Users can employ a predefined pipeline suitable for their data by selecting a pipeline in the pipeline panel. If users want to create a new analysis pipeline, they can build their own pipeline either from scratch or by modifying a predefined pipeline. The following sections describe representative predefined analysis pipelines in the pipeline panel.

Genome pipeline

For the analysis of genome data and high-density single nucleotide polymorphism (SNP)-arrays, we developed 25 pipelines and programs, which can be grouped into seven categories: (1) discovery
of variants for human, animal and plant data, (2) discovery of candidate genes from whole-exome sequencing data of rare diseases, (3) identification of somatic mutations, SNPs and short INDELs from cancer genomes, (4) clonality and evolutionary analysis of cancer genomes, (5) structural variation and copy-number analysis of whole-genome sequencing data, (6) population genomic analysis of whole-genome sequencing data, and (7) association studies and genomic predictions from the high-density SNP-arrays. We also developed two dockerized workflows that can be used for the discovery of SNPs, short INDELs, or copy-number variations from germline and somatic sample data, and for population genomics analysis in evolutionary studies. The two dockerized workflows were developed using the Workflow Description Language, developed on the Data Sciences Platform at the Broad Institute.

We developed two tools using genome data: GenoCore [8] and SEXCMD [9]. GenoCore is a new method for selecting a core collection using modified statistical measures related to genetic allele coverage and diversity. It can be used to select core subsets from plant genotype datasets, which is important for increasing cost-effectiveness and shortening the time required for the analyses of genome-wide association studies (GWAS), genomics-assisted breeding of crop species, etc. SEXCMD is a pipeline that can extract sex marker sequences from reference sex chromosomes and rapidly identify the sex of individuals from whole-exome/genome and RNA sequencing (RNA-Seq) data.

Transcriptome pipeline
The analyzing an organism’s transcriptome is important for understanding the functional elements of a genome [10]. RNA-Seq is a deep-sequencing technique that can be used to explore and profile the entire transcriptome of any organism [11]. Fig. 2 shows a typical schematic overview of the RNA-Seq analysis pipeline on the canvas. The pipeline, often referred to as the tuxedo pipeline, includes five analysis tools: TopHat 2.1.1 [12], Cufflinks 2.1.1 [13], Cuffmerge 2.1.1, Cuffdiff 2.1.1, and limma voom 1.0 [14]. TopHat is a fast splice junction mapper that is used to align RNA-Seq reads to large genomes and analyze the mapping results to identify splicing junctions between exons. Cufflinks is used to assemble these alignments into a parsimonious set of transcripts and then estimate the relative abundances of these transcripts. The main purpose of Cuffmerge is to merge several Cufflinks assemblies, making it easier to produce an assembly GTF file suitable for use with Cuffdiff. Cuffdiff is then used to identify significant changes in transcript expression, splicing, and promoter use. Finally, voom robustly estimates the mean-variance relationship and generates a precision weight for each individual normalized observation, which can be used to calculate differentially expressed genes from transcript ex-

Fig. 2. Screenshot of the RNA-sequencing (RNA-Seq) schematic diagram and its pipeline. The RNA-Seq pipeline was implemented on the canvas.
pression levels. Several other pipelines for RNA-Seq data analysis are available at Bio-Express, including MapSplice2-RSEM [15], Bowtie-EMSAR [16], STAR-HTSeq [17], and STAR-RSEM [18].

**Epigenome pipeline**

Epigenetic changes, including histone modifications and DNA methylation, provide a differential gene regulatory mechanism without altering DNA sequences [19]. Histone modifications occur mostly at histone tails by acetylation, methylation, phosphorylation, and ubiquitination. The accurate mapping of the called peaks of these modification sites is a critical step for understanding epigenetic transcriptional regulation. A popular, fast applicable pipeline for histone modification mapping was established by comparing various peak calling programs such as CisGenome [20], MACS1 and MACS2 [21], PeakSeq [22] and SISSRs [23], RSEG [24], SICER [25], hiddenDomains [26], BroadPeak [27], PeakRanger-CCAT, and PeakRanger-BCP [28]. For the best performance to define the exact binding sites of proteins in DNA, we tested 12 histone modifications using different peak calling programs, and we suggest the MAC2 program for narrow peak identification and PeakRanger-BCP for broad peak identification. The analysis pipeline for histone modifications is summarized in Fig. 3; the input files in fastq format are preprocessed by cudapt, fastq_quality_filter, and paired_sequence_match.py and then read quality is tested with FastQC. After mapping reads onto the reference genome, peak calling or domain calling is followed by application of MACS2 or PeakRanger-BCP. The final output is produced with annotation information. This simple pipeline is open to the public under the Bio-Express portal provided by KOBIC.

**Metagenome pipelines**

The analysis of metagenome data can be categorized into three parts (Fig. 4): whole metagenome shotgun sequence data analysis (shotgun metagenomics), whole transcriptome shotgun sequence data analysis (RNA-Seq), and 16S rRNA gene amplicon sequence data analysis (16S sequencing). In shotgun metagenomics, there are three pipelines: the assembly-based gene profiling, scaffold-binning, and reference-guided analysis pipelines. In the assembly-based gene profiling pipeline, sequence reads are assembled using SOAPdenovo-63mer [29]; gene regions in the assem-

![Fig. 3. Workflow for the histone modification analysis pipeline. ChIP-Seq, chromatin immunoprecipitation sequencing.](https://doi.org/10.5808/GI.2020.18.1.e8)
bled sequences are predicted using MetaGeneMark [18], and the functions of the gene regions are assigned by the BLAST program with the COG and GenBank nr databases.

In the scaffold-binning pipeline, the coverage and GC content of the scaffolds are calculated, and taxonomic identifiers are assigned to the scaffolds using MEGAN [30] and HMMER 3.0 [31]. In the reference-guided analysis pipeline, sequence reads are mapped with the BWA program with reference genes or genomes. In the RNA-Seq category, sequence reads are mapped and normalized, statistical analyses are performed to identify differentially abundant genes, and finally, the results are annotated.

The 16S sequencing category is composed of three modules in sequential order: automatic platform-specific quality control (QC), community analysis, and statistical analysis and graphics. We developed a program, AutoQC, for the automatic platform-specific QC module. AutoQC uses platform-specific conditions to efficiently remove erroneous reads. AutoQC is freely available at https://sourceforge.net/projects/autoqc/. The community analysis module mainly reveals the microbial diversity and classification of microbes using Mothur. In the statistical analysis and graphical statistical analyses like pMANOVA test [32] are performed and the analysis results are visualized.

Creating custom (user defined) pipelines

Users can create a new pipeline to analyze their own data on the canvas. To create a new pipeline, users click the 'New Pipeline' button in the top menu and select an analysis pipeline type. Users will have only the [Start] and [End] modules on the canvas immediately upon creating a pipeline after selecting a ‘new analysis pipeline design’ in the project type. Users can drag and drop their desired analysis programs from the list of analysis programs on the right of the canvas. After the positioning of a desired analysis program on the canvas, when the users place the mouse over the edge of the analysis program icon, a connection mark will be created that can be drawn to the module. Starting from the mark, the connector must be dragged until the icon of the next analysis program to be connected becomes translucent. Users can make connections to the start module, the analysis program and the end module using this method to perform the analysis. The path for the output file is automatically a sub-path of the project in setting the input data. Finally, the analysis pipeline is executed with a message that the analysis has started. The status of the project is displayed on a

Fig. 4. Simplified workflow diagram of the metagenomics pipelines.
real-time basis in three modes: Complete, Execute, and Wait.

Users can see the final results by clicking the ‘Results’ icon on the menu and downloading them to the user’s local computer by clicking the ‘Download’ button on the menu bar. Bio-Express also allows users to view files in various formats including text, HTML, and PNG on the screen without having to download the files (Fig. 5).

**Web servers**

We have developed traditional web servers in which the input is a small amount of data such as a gene list. The traditional web servers do not provide automatic scalability to the applications which is the major feature of the cloud server [33]. The developed web servers are ADGO2 [34], ExPathNet [35], GSA-SNP [36], and Barcas [37].

ADGO2 provides biological interpretations of microarray data (gene-set enrichment approach) and a list of genes (gene list over-representation approach) via composite annotation. ADGO2 also supports gene- or sample-permuting gene-set enrichment analysis for RNA-Seq count data. ExPathNet provides network-weighted gene-set clustering that incorporates both gene-set overlap and protein-protein interaction networks. GSA-SNP is standalone software that provides widely used GSA methods for SNP and GWAS data. GSA-SNP2 [38] is an improved version of GSA-SNP that provides fast high-power computation by incorporating the random set model and SNP-count adjusted gene scores. GSA-SNP2 can also visualize protein interaction networks within and across the significant pathways. Barcas is pharmacogenomics data analysis software developed for the mapping and analysis of multiplexed barcode sequencing data. Barcas employs a trie data structure for fast mapping with mismatches allowed and provides many functions, including quality control, data analysis and visualization. Table 1 shows the web servers used for gene-set, pathway, and pharmacogenomic data analysis.

**Comparison between Bio-Express and Galaxy**

We compared Bio-Express with Galaxy, an open source system that is the most widely used pipeline system and empowers non-computational users to do computational biology. We performed a comparison experiment between Bio-Express and Galaxy with the same data and the same RNA-Seq pipeline. We used

---

**Table 1**

<table>
<thead>
<tr>
<th>Web Server</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADGO2 [34]</td>
<td>Biological interpretations of microarray data (gene-set enrichment approach) and a list of genes (gene list over-representation approach) via composite annotation.</td>
</tr>
<tr>
<td>GSA-SNP [36]</td>
<td>Standalone software that provides widely used GSA methods for SNP and GWAS data.</td>
</tr>
<tr>
<td>GSA-SNP2 [38]</td>
<td>Improved version that provides fast high-power computation by incorporating the random set model and SNP-count adjusted gene scores.</td>
</tr>
<tr>
<td>Barcas [37]</td>
<td>Pharmacogenomics data analysis software developed for the mapping and analysis of multiplexed barcode sequencing data.</td>
</tr>
</tbody>
</table>

---

*Fig. 5.* Screenshot of Bio-Express results. Users can view files in various formats, including text, HTML, and PNG on the web.
an RNA-Seq case-control sample data set: 42,112,235 paired-end case reads and 40,975,645 paired-end control reads. The total sample size of the case and the control reads is approximately 42 GB. We assigned four CPU cores and 16 GB of memory for a single RNA-Seq job. The same machine was used for the comparison. The execution of the RNA-Seq pipeline on the sample data using Bio-Express takes a total of 3 h 44 min. The execution time using Galaxy was 6 h 11 min, showing Bio-Express has approximately 1.7 times better performance than Galaxy in the execution of the RNA-Seq pipeline. There are two main reasons for the difference in runtime between the two systems. As Galaxy internally processes intermediate data for data conversion after finishing each pipeline program, the execution time is slightly increased due to the internal process of each step. Secondly, Bio-Express has fast access to input and output data by fully utilizing the function of a SSD cache, compared to the Galaxy system.

### Discussion

The substantial decrease in the cost of NGS techniques in the past decade has dramatically reshaped the genome research and has led to its rapid adoption in biological research. Nowadays, massive amount of data can be generated quickly using NGS platforms. These data range from the function and regulation of genes, the clinical diagnosis and treatment of diseases, to the omics profiling of individual patients for precision medicine. With the exponential increase in volume and complexity of NGS data, cluster or HPC systems are essential for the analysis of large amounts of NGS data. But the associated costs with the infrastructure itself and the maintenance personnel will likely be prohibitive for small institutions or laboratories.

Cloud-based applications and resources have been developed specifically to address the computational challenges of working with very large volumes of data generated by NGS technology. Cloud computing has changed how we manage computational resources. Increasingly cloud computing is also changing how large computational resources are organized and how scientists in genomics collaborate and deal with vast genome data sets.

We presented a Hadoop based distributed computational framework for large-scale genomic analysis, called Bio-Express, which incorporates a variety of tools and methods. Our system offers a variety of services to researchers. Firstly, Bio-Express allows genomic researchers without informatics or programming expertise to perform complex large-scale analysis with only a web browser using drag and drop functionality. Secondly, Bio-Express is a hybrid system that enables users to use both analysis programs providing traditional tools and MapReduce-based big data analysis programs simultaneously in a single pipeline. Lastly, we also developed a high-speed data transmission solution, Gbox, to transmit a large amount of data at a fast rate.

In the future work, we continuing to add powerful pipelines and programs including the most popular sequence and genome analysis algorithms, and to enable accessible and reproducible genomic science. Secondly, we plan to create a framework with both client-side and server-side components that simplifies the development of web-based visual applications. Visualization and visual analysis are important tools in high-throughput genomics experiments because large datasets do not need to be downloaded. Lastly, we will create a standalone installation package of Bio-Express. The increasingly large size of many datasets and moving the huge datasets is one particularly challenging aspect of current and future genomic science. Hence, local Bio-Express installations near the data are likely to become more prevalent because it makes more sense to run Bio-Express locally as compared to moving the data to a remote Bio-Express server.

### ORCID

Gunhwan Ko: https://orcid.org/0000-0002-3570-9074
Pan-Gyu Kim: https://orcid.org/0000-0002-6069-1611
Youngbum Cho: https://orcid.org/0000-0002-0335-5961
Seongmun Jeong: https://orcid.org/0000-0002-0038-461X
Jae-Yoon Kim: https://orcid.org/0000-0002-8557-0998
Kyoung Hyoun Kim: https://orcid.org/0000-0002-7652-1879
Ho-Yeon Lee: https://orcid.org/0000-0002-1026-1744
Jiyeon Han: https://orcid.org/0000-0002-6001-1540

---

### Table 1. Web servers for gene-set, pathway, and pharmacogenomic data analysis

<table>
<thead>
<tr>
<th>Tool</th>
<th>Main function</th>
<th>Web address</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADGO2</td>
<td>Gene-set analysis of microarray data</td>
<td><a href="http://www.btool.org/ADGO2">http://www.btool.org/ADGO2</a></td>
</tr>
<tr>
<td>GSA-SNP2</td>
<td>Gene-set analysis of GWAS summary data</td>
<td><a href="https://sites.google.com/view/gsasnp2">https://sites.google.com/view/gsasnp2</a></td>
</tr>
<tr>
<td>Barcas</td>
<td>Software for analyzing barcode-seq data</td>
<td><a href="http://medical-genome.kribb.re.kr/barseq/">http://medical-genome.kribb.re.kr/barseq/</a></td>
</tr>
</tbody>
</table>

GWAS, genome-wide association studies.
Namhee Yu: https://orcid.org/0000-0003-3967-0813
Seokjin Ham: https://orcid.org/0000-0002-6950-2848
Insoon Jang: https://orcid.org/0000-0001-9542-6984
Byunghee Kang: https://orcid.org/0000-0002-6398-5762
Sunguk Shin: https://orcid.org/0000-0002-3812-8320
Lian Kim: https://orcid.org/0000-0002-7417-0366
Seung-Won Lee: https://orcid.org/0000-0002-1878-3766
Dougu Nam: https://orcid.org/0000-0003-0239-2899
Namshin Kim: https://orcid.org/0000-0001-6361-274X
Jihyun F. Kim: https://orcid.org/0000-0002-9230-7730
Tae-Young Roh: https://orcid.org/0000-0001-5833-0844
Byungwook Lee: https://orcid.org/0000-0003-2083-8508

Authors’ Contribution

Conceptualization: JFK, NK, SYK, SL, TYR, BL. Data curation: GK, PGK, BL. Formal analysis: JH, NY, SH, IJ, BK, SS, LK, SWL. Funding acquisition: JFK, NK, SYK, SL, TYR, BL. Methodology: YC, SJ, JYK, KHK, HYL, DN. Writing - original draft: DN, JFK, NK, SYK, SL, TYR, BL. Writing - review & editing: BL.

Conflicts of Interest

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

Acknowledgments

This work was supported by the grants from the National Research Foundation of Korea (NRF-2014M3C9A3064552, NRF-2014M3C9A3065221, NRF-2014M3C9A3064548, NRF-2014M3C9A3068554, NRF-2014M3C9A3068822, and NRF-2019M3C9A5069653). A portion of the data used for this study were obtained from the Genome-InfraNet (IDs: 1711041199, 1711057837, 1711031849, 1711042674) at the Korea Bioinformatics Center.

References

1. Bansal V, Boucher C. Sequencing technologies and analyses: where have we been and where are we going? iScience 2019;18:37-41.
Transposable elements (TEs) constitute approximately half of Bovine genome. They can be a powerful species-specific marker without regression mutations by the structure variation (SV) at the time of genomic evolution. In a previous study, we identified the Hanwoo-specific SV that was generated by a TE–association deletion event using traditional PCR method and Sanger sequencing validation. It could be used as a molecular marker to distinguish different cattle breeds (i.e., Hanwoo vs. Holstein). However, PCR is defective with various final copy quantifications from every sample. Thus, we applied to the droplet digital PCR (ddPCR) platform for accurate quantitative detection of the Hanwoo-specific SV. Although samples have low allele frequency variation within Hanwoo population, ddPCR could perform high sensitive detection with absolute quantification. We aimed to use ddPCR for more accurate quantification than PCR. We suggest that the ddPCR platform is applicable for the quantitative evaluation of molecular markers.

Keywords: droplet digital PCR, Hanwoo-specific marker, structure variation

Introduction

Hanwoo (Bos taurus coreanae) is a domesticated mammal that has been used for agriculture and transportation since 5,000 years ago [1]. As the Korea economy developed in 1960, it began to provide as one of food resources [2]. In particular, Hanwoo is consumed more beef than other cattle breeds in Korea [3]. This consumption pattern has led to the emergence of research on the development of molecular makers that distinguish between Hanwoo and other cattle breeds [4-6].

In a recent study, they investigated Hanwoo-specific structural variation (SV) using BreakDancer program (ver 1.1) to distinguish between Hanwoo and Holstein [7]. The SVs typically included insertion, deletion, inversion, translocation, and copy-number variation [8-10]. SVs could affect much greater genomic function and gene expression than single nucleotide variants [11]. In this respect, the previous study focused on transposable element (TE)-mediated deletion events. Thus, Park et al. [7] identified an authentic Hanwoo-specific deletion locus that was confirmed by PCR and Sanger sequencing. It can be utilized to distinguish between Hanwoo and Holstein species. However,
PCR has several defects in detecting DNA amplification. For example, contaminated sample including trace amounts of DNA might lead to misleading outputs [12]. In addition, the specificity of the PCR product could be affected by non-specific binding of the primers to other similar sequences on the template DNA [12]. Complementing these drawbacks, the quantitative PCR (qPCR) could estimate target DNA quantity using either a fluorescent dye (e.g., SYBR Green) that non-specifically intercalates with double-stranded DNA (dsDNA) or TaqMan probe assay. Nevertheless, most qPCR methods rely on the precise number of copies compensated by calibrator, assuming no loss of calibrator molecules during the all experimental steps [15]. However, errors can occur at several levels [14,15]. In addition, the qPCR has the following disadvantages. (1) The accuracy of qPCR depends on proper experimental design based on well-established reference genes. (2) For absolute quantification, you should create a standard curve for data normalization based on changes in the transcription level of the reference gene [16].

The droplet digital PCR (ddPCR) is one of next-generation technologies for absolute quantification of nucleic acids [17]. It counted the fluorescent PCR-positive and PCR-negative droplets to calculate target DNA concentration and thus absolute quantification was directly estimated as the exact number of copies without the aid of calibration curve [15]. Currently, seven commercial digital PCR systems (Thermo Fisher Quantstudio 3D, Fluidigm BioMark qPCR 37K, Formulatrix Constellation, JN Medsys Clarity, Bio-Rad QX200, Raindance Raindrop plus, and Stilla Naica) are available [18]. Among them, the Stilla Naica System for Crystal Digital PCR [19] has a predominant feature of step emulsion generators. It is not necessary to do the flow of oil by developing the Sapphire chip, which development has simplified the operation and reduced potential contamination.

This study uses ddPCR, the Stilla Naica System for Crystal Digital PCR, to overcome the limitations of PCR and to accurately evaluate the Hanwoo-specific SV locus that was identified in the previous study [7]. We suggest that the ddPCR platform can be used as a quantitatively and numerically sensitive method with molecular markers.

**Methods**

The five brown Hanwoo DNAs and five Holstein DNAs were extracted from blood samples using the DNeasy Blood & Tissue kit according to the manufacturer’s instruction (Qiagen, Hilden, Germany). All research protocols and animal experiments in this study were reviewed and approved by the Institutional Animal Care and Use Committee (IACUC) in Gyeongsangbuk-do, Republic of Korea (Gyeongbuk IACUC-87). Next, we confirmed the PCR amplicon pattern of the “Del_96” locus [7] from all samples by PCR. The Hanwoo samples showed a polymorphic pattern of PCR products (680 bp/310 bp) generated by TE-association deletion event. However, Holstein samples contained no the deleted allele, so only PCR products of 680 bp are observed (Fig. 1A).

To more accurately detect the Hanwoo-specific SV, we have applied the “Del_96” locus [7] to the ddPCR platform (Stilla Technologies, Villejuif, France). The FAM primer set and FAM probe (Thermo Fisher Scientific, Waltham, MA, USA) were used for the detection of both Hanwoo and Holstein genomes. The VIC primer set and VIC probe (Thermo Fisher Scientific) were designed at the boundary of Hanwoo-specific deletion (Fig. 1B). Thus, we designed that the FAM primer set and FAM probe were detected in all cattle DNAs (positive control). The VIC primer set and VIC probe were designed to detect fluorescence only in the Hanwoo cattle. We followed the manufacturer’s instructions for experi-

---

**Fig. 1.** Structural variation of the Hanwoo and Holstein genomes. (A) Polymorphic pattern of the Del_96 locus in the Hanwoo and Holstein cattle samples [7]. Gel chromatography showed that five Hanwoo samples (left panel) contained heterozygous alleles (680 bp and 310 bp) but five Holstein samples (right panel) had no the deleted allele (680 bp). (B) To analyze absolute quantification using droplet digital PCR assay, the FAM probe (blue box) was designed to detect all cattle genome (positive control). The VIC probe (green box) was designed in boundary of Hanwoo-specific deletion (Del_96).
menting with the ddPCR platform. Prior to the experiment, we confirmed the quantification of Hanwoo and Holstein DNAs using Qubit 4.0 Fluorometer (Thermo Fisher Scientific) with 1 × dsDNA HS (high-sensitivity) assay kit (Thermo Fisher Scientific) for dsDNA measurement.

The ddPCR reaction mixture (25 µL) contained 12.5 µL of PerfectCt qPCR ToughMix UNG 2× (Quanta Biosciences, Gaithersburg, MD, USA), 2.5 µL of 100 nM of Fluorescein (VWR International, West Chester, PA, USA), 1.25 µL of primer set/VIC probe (final concentration of 900 nM/250 nM, respectively), 50 ng DNA, and nuclease-free water up to 25 µL. The reaction mixtures were loaded into wells of Sapphire chip (Stilla Technologies), respectively. Then, the chips are placed into the Naica Geode equipment and we launched the combined partitioning and thermocycling program. The ddPCR condition was initial denaturation step of 3 min at 95°C, followed by 45 cycles of 95°C for 10 s and 60°C for 15 s, with a release step for 33 min to down temperature and pressure. 20,000 to 30,000 droplets are created from each sample. At the end of template amplification from the separated droplets, the chips were transferred to the Naica Prism3 reader. Finally, extracted fluorescence values for each droplet were analyzed using the Crystal Miner software (Stilla Technologies). Thresholds were set using the automation tools available in the Crystal Miner software.

### Results and Discussion

Hanwoo-specific deletion locus (Del_96 region) was found in a previous study by comparing the cattle genomes with whole-genome sequencing data and proved by PCR and Sanger sequencing methods [7]. It has been reported that the Del_96 region occurred through nonallelic homologous end-joining between LINE (BovB) and unique sequence only in the Hanwoo genome [7]. It can be used as a powerful marker for distinguishing Hanwoo and Holstein (Fig. 1A). Even though validation experiment based on PCR method used in their study are easy to perform at small sample size, the PCR method can be affected by nonspecific binding of primer set to similar sequences on the gDNA [12]. To overcome the shortcomings of the PCR method and apply next-generation technology, we try to verify the Hanwoo-specific deletion region by a ddPCR assay.

To perform the ddPCR assay, we designed two probes (Supplementary Table 1). One designed a positive control probe (FAM dye; blue) to detect all cattle genomes, and the other to a Hanwoo-specific deletion boundary site (VIC dye; green) (Fig. 1B). DNA templates from five Hanwoo and five Holstein blood samples were conducted to the ddPCR assay with designed primer/probe sets. The extracted DNAs should be assessed for accurate quantification using a UV spectrophotometer (NanoDrop, Thermo Fisher Scientific) and an intercalating reagent reaction with the dsDNA (Qubit assay). In particular, it is important to quantify dsDNA because dsDNA of total DNA actually reacts in the ddPCR assay (Table 1) [20].

The Stilla Naica system yields between 20,000 and 30,000 analyzable droplets. In this study, we generate an average approximate 22,392 of droplets using the Stilla Naica system (Table 2). Thus, there are enough droplets to analyze the absolute copy number. As shown in Fig. 2, FAM dye was detected in all cattle genomes and VIC dye showed significant detection only in the Hanwoo samples. It suggests that all Hanwoo genomes contain the specific deletion sequence (Del_96 region). Signals of VIC dye were detected on average 243 Channel concentration (copy/μL) in the Hanwoo samples. However, an average of 0.12 Channel concentration (copy/μL) VIC dye signals, which were very few and insignificant droplets, were also detected in the Holstein samples. In the previ-

### Table 1. Cattle gDNA quality control and dsDNA concentration

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Microvolume spectrometer Concentration (ng/μL)</th>
<th>Qubit fluorescence 4.0 dsDNA concentration (ng/μL)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A260/A280</td>
<td>A260/A230</td>
</tr>
<tr>
<td>Hanwoo_#16</td>
<td>33.2</td>
<td>1.92</td>
</tr>
<tr>
<td>Hanwoo_#23</td>
<td>35.1</td>
<td>1.83</td>
</tr>
<tr>
<td>Hanwoo_#289</td>
<td>35.1</td>
<td>2.01</td>
</tr>
<tr>
<td>Hanwoo_#296</td>
<td>23.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Hanwoo_#303</td>
<td>18.8</td>
<td>1.78</td>
</tr>
<tr>
<td>Holstein_DC2</td>
<td>36.6</td>
<td>1.87</td>
</tr>
<tr>
<td>Holstein_DC5</td>
<td>85.5</td>
<td>1.91</td>
</tr>
<tr>
<td>Holstein_DCM2</td>
<td>34</td>
<td>1.93</td>
</tr>
<tr>
<td>Holstein_DCM3</td>
<td>29.2</td>
<td>1.91</td>
</tr>
<tr>
<td>Holstein_DCM5</td>
<td>25.5</td>
<td>1.63</td>
</tr>
</tbody>
</table>

gDNA, genomic DNA; dsDNA, double-stranded DNA.

https://doi.org/10.5808/GI.2020.18.1.e4
The Del_96 region deleted from the Hanwoo genome was reported to occur in one of the transposable elements, the BovB element region. At present, the cattle reference genome (bosT au9 version) has not well annotated the segmental duplication region and TE positions. Therefore, it is important to consider that VIC probe designed in the TE region can detect non-specific signals on sequences with high similarity. In addition, the signals obtained from these droplets could be recognized as false signals due to the abnormally high fluorescence intensity measured in ddPCR assay [21,22]. Nevertheless, the difference in the average number of VIC dyes detected between the Hanwoo and Holstein samples was statistically sufficient to distinguish them. Our results show that the ddPCR assay is very appropriate to distinguish between Hanwoo and Holstein cattle. On the other hand, the signals of the FAM dye were detected on average 253.5 Channel concentration (copy/μL) in the Hanwoo samples and an average of 516.7 Channel concentration (copy/μL) FAM dye signals were also detected in the Holstein samples (Fig. 3). As shown in Fig. 1B, we designed a FAM probe/primer set for the sequence that exist within Hanwoo-specific deletion region. Thus, the copy numbers that were detected by FAM dye signal were observed two times more in Holstein samples than Hanwoo samples.

Table 2. Statistical result of the ddPCR assay

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Total No. of droplets</th>
<th>Channel concentration (copy/μL)</th>
<th>No. of positive droplets</th>
<th>p-value</th>
<th>Channel concentration (copy/μL)</th>
<th>No. of positive droplets</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTC</td>
<td>23,549</td>
<td>0</td>
<td>0</td>
<td>N/A</td>
<td>0</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>Hanwoo_#16</td>
<td>23,782</td>
<td>266.3</td>
<td>3,436</td>
<td>0.0335</td>
<td>253.2</td>
<td>3,279</td>
<td>0.0343</td>
</tr>
<tr>
<td>Hanwoo_#23</td>
<td>24,123</td>
<td>247</td>
<td>3,250</td>
<td>0.0344</td>
<td>244.8</td>
<td>3,223</td>
<td>0.0346</td>
</tr>
<tr>
<td>Hanwoo_#289</td>
<td>23,601</td>
<td>255</td>
<td>3,275</td>
<td>0.0343</td>
<td>234.2</td>
<td>3,026</td>
<td>0.0357</td>
</tr>
<tr>
<td>Hanwoo_#296</td>
<td>23,568</td>
<td>252.6</td>
<td>3,242</td>
<td>0.0345</td>
<td>245</td>
<td>3,152</td>
<td>0.0349</td>
</tr>
<tr>
<td>Hanwoo_#303</td>
<td>22,839</td>
<td>246.5</td>
<td>3,071</td>
<td>0.0354</td>
<td>236.2</td>
<td>2,952</td>
<td>0.0361</td>
</tr>
<tr>
<td>Holstein_DC2</td>
<td>25,068</td>
<td>524.1</td>
<td>6,628</td>
<td>0.0242</td>
<td>0.07</td>
<td>1</td>
<td>1.96</td>
</tr>
<tr>
<td>Holstein_DC5</td>
<td>20,006</td>
<td>516</td>
<td>5,220</td>
<td>0.0272</td>
<td>0.09</td>
<td>1</td>
<td>1.96</td>
</tr>
<tr>
<td>Holstein_DCM2</td>
<td>19,341</td>
<td>511.3</td>
<td>5,007</td>
<td>0.0278</td>
<td>0.26</td>
<td>3</td>
<td>1.132</td>
</tr>
<tr>
<td>Holstein_DCM3</td>
<td>20,577</td>
<td>527.9</td>
<td>5,474</td>
<td>0.0266</td>
<td>0.08</td>
<td>1</td>
<td>1.96</td>
</tr>
<tr>
<td>Holstein_DCM5</td>
<td>19,863</td>
<td>504.2</td>
<td>5,081</td>
<td>0.0276</td>
<td>0.09</td>
<td>1</td>
<td>1.96</td>
</tr>
</tbody>
</table>

ddPCR, droplet digital PCR; N/A, not available.

Fig. 2. 1D-Dot plot display of mono-color droplet fluorescence intensity. The dots indicate each droplet that was detected by FAM (left plot) and VIC (right plot) dyes using the droplet digital PCR assay. (A) The X- and Y-axis indicate the name of each sample and the number of droplets with positive fluorescence intensity with the FAM probe (blue color), respectively. (B) The X- and Y-axis indicate the name of each sample and the number of droplets with positive fluorescence intensity with the VIC probe (green color).
In ddPCR assay, DNA is divided into numerous wells or droplets, and the concentration of target region is absolute quantified using Poisson statistics [23,24]. The ddPCR assay can be quantified with high accuracy in counting single molecules and analyzing a small number of copies of a particular population [25,26]. However, consumable and equipment cost for ddPCR are still expensive compared to those of qPCR.

For the ddPCR technology, accurate quantification of absolute copy number is a key feature. In the near future, by applying species-identifying makers to ddPCR, it has significant potential as a platform for species identification at large sample sizes. Taken together, we propose that ddPCR is suitable as a platform for verifying species-specific markers.

**ORCID**

Wonseok Shin: https://orcid.org/0000-0001-5964-1425  
Haneul Kim: https://orcid.org/0000-0002-2860-7133  
Dong-Yep Oh: https://orcid.org/0000-0003-4412-7719  
Dong Hee Kim: https://orcid.org/0000-0003-3940-5315  
Kyudong Han: https://orcid.org/0000-0001-6791-2408

**Authors’ Contribution**

Conceptualization: KH. Data curation: KH, WS, DHK. Formal analysis: WS, DYO. Funding acquisition: KH. Methodology: WS, HK. Writing - original draft: KH, WS, HK. Writing - review & editing: KH, WS, DHK.

**Conflict of Interest**

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

**Supplementary Materials**

Supplementary data including one table can be found with this article at https://doi.org/10.5808/GL2020.18.1.e4.

**References**


Detecting outliers in segmented genomes of flu virus using an alignment-free approach

Mosaab Daoud*
Independent Research Scientist, Toronto, ON M1S1G2, Canada

In this paper, we propose a new approach to detecting outliers in a set of segmented genomes of the flu virus, a data set with a heterogeneous set of sequences. The approach has the following computational phases: feature extraction, which is a mapping into feature space, alignment-free distance measure to measure the distance between any two segmented genomes, and a mapping into distance space to analyze a quantum of distance values. The approach is implemented using supervised and unsupervised learning modes. The experiments show robustness in detecting outliers of the segmented genome of the flu virus.

Keywords: composite data point, distance space, flu virus, Mosaab-metric space, outliers, statistical learning

Introduction

Recent years have witnessed a dramatic increase in the amount of genome data that is submitted to on-line databases. Analyzing sequence-based datasets is the aim of sequence analysis and biodata mining research fields. The engineering solutions have not been achieved to analyze data sets with heterogeneous feature. In other words, the datasets under consideration are sets of sequences with different biological functions and different base-composition distributions. The problem under consideration has several computational challenges. The first challenge is the representation of the inner information structure of a segmented genome of flu virus in feature spaces. Another challenge is to define a metric and metric space to measure the distance between any two information structures that are embedded in a well-defined feature space or composite feature space, and the third challenge is to analyze a quantum of distance values in distance space. The approach that we propose in this paper is alignment-free approach, which is different from classical alignment approaches in terms of time complexity, selectivity, and sensitivity analysis.

At this point, the structure of this paper can be summarized as follows. In next subsection, we shall present a review of the existing approaches to tackle related research problems. In section (Methods), we shall present the approaches of detecting outliers in segmented genomes of the flu virus. The experiments and results are presented in section (Results). Finally, conclusions and future work will be presented in section (Discussion).

The related work

The Influenza virus is a highly mutated virus. It has a negative impact on the human population. Consequently, it has a negative impact on public health and the economy. The virus has a segmented genome that can be encoded to 10–11 proteins. The virus is classi-
fied into types and subtypes. The variation in the base composition of the surface proteins haemagglutinin (HA) and neuraminidase (NA) indicates the type and the subtype of the influenza virus [1].

The influenza virus is a negative stranded RNA-virus. It is classified under the family Orthomyxoviridae. The virus has three types A, B, and C. The most variable type is the influenza virus A compared to other types [2]. The accumulation of point mutation in the HA and NA surface proteins causes an antigenic drift. The evolution process is a continuous or discrete in real time. It takes place on the genetic information of the virus. Consequently, it produces new distinct strains.

An alignment-free sequence comparison analysis is a new developing research direction. It has the potential of solving the sequence proximity problem with less time complexity compared to the alignment-based analysis [3]. There are several strengths behind this fact: we can project those sequences into several feature spaces to detect the information structure in various ways. This approach helps research in bioinformatics and biotechnology fields to gather more information about sequences or genomes. Mapping those sequences into feature spaces in a format of data-vectors allows the computational research community to implement a wide range of techniques in data mining, machine learning, and statistical learning in feature spaces, which are behind the capacity of alignment-based techniques [4]. No prior biological assumptions about sequences are required to implement alignment-free techniques, while the alignment-based techniques have to be implemented with pre-assumptions about the inheritance of sequences. In this context, two concepts are arising: homogeneous and heterogeneous sequences under consideration. Moreover, the alignment-free techniques can be implemented when the alignment-based techniques are inapplicable.

Any biosequence is linear in time. Therefore, the sequential relation is the most promising feature in biosequences [5]. Biosequences are drawn from finite alphabets. Any biosequence can be mapped into a feature space using n-grams technique as feature extraction technique. The computational mechanism of this technique can be implemented in different ways. Without loss of generality, assume that we have a sliding window of length $W$, moving a sliding window from one end to another to estimate the relative frequency of the occurrences of n-grams. The sliding window can be shifted by a shift distance $a$. Local statistical information about biosequences can be extracted in this way. The distance between any two sequences can be measured in a feature space by measuring the distance between the frequency distributions (i.e., data vectors) of the two sequences. There are several similarity/distance measures that can be used to measure the distance between data-vectors that are extracted using n-grams. The extraction can be achieved either using frequency distribution or relative frequency distribution. A distance function $D()$ is a mapping from a well defined domain to measure the proximity between two entities (e.g., two vectors or sequences) into the interval $[0, \infty]$. $D()$ is a metric if it is satisfying the following conditions: positivity, symmetry, and triangular inequality [5]. The similarity measure $S()$ is a mapping into the interval $[0, 1]$, where the value 0 represents the lowest similarity and the value 1 represents the highest similarity.

There are a number of distance measures that can be implemented in measuring the proximity between any two sequences without using alignment. Those measures are either similarity measures or distance measures.

One of the distance measures used in multivariate analysis is the angle cosine between two data-vectors [6]. Each data-vector represents a sequence, and the proximity of two sequences is measured by the angle cosine. The measure detects the differences between two data-vectors, where each data vector represents the relative occurrences of selected n-grams. The measure is not sensitive to repetition of motifs. In information theory, the Kullback-Leibler discrepancy is a well-known measure and it measures the divergence between two probability distributions, where each probability distribution represents a sequence, and defined as the occurrences of selected n-grams in a sequence.

Han et al. [7] proposed an alignment-free sequence comparison method to detect the dissimilarity between any two sequences. The defined distance is based on two factors: the relative frequency distribution of n-grams as a data-vector and the position information as a normalized average data-vector. The distance measure is defined as a weighted distance measure, and the weights are defined in terms of variations of those two factors in a selected genome set. The computational mechanism used in this method is a window-based mechanism. Finally, the phylogenetic tree is composed based on the distance values of the proposed distance measure.

Daoud [8] proposed an alignment-free sequence comparison technique to analyze sequences in feature space. The stochastic membership values of a query sequence with respect to different classes of sequences are estimated using Minkowski measure. The working mechanism proposed in this research is window-based mechanism. The membership value is estimated based on the following question: Is a query sequence probably approximately belongs to a specific class of sequences? In this case, the quantum of distance values composes an empirical distance distribution and the membership value is estimated from the empirical distance distribution.

Daoud [9] proposed a visualization approach to visualize composite data points in feature spaces using the variation theory. The implementation of this computational approach is directed to seg-
mented genomes. It is based on window-based mechanism. The robustness of this approach is implicitly depending on its implementation to flu virus. It is the first attempt to graphically show the serious of difference among the segmented genome of the flu virus.

The other measure used in measuring the distance between two sequences without using alignment is the Euclidean distance \[5\].

The first step is to extract frequency data-vectors for each sequence using \(n\)-grams feature extraction technique, then the Euclidean distance is applied to measure the distance between two frequency data-vectors, which reflects the identicalness between two sequences.

As we mentioned in this paper, alignment-free is new developing research direction. There are more than 45 alignment-free tools available with different applications in the area of sequence-analysis. A summary of comparisons between alignment-free and alignment-based algorithms are given in Table 1. As an expectation, the next generation of computational pipelines will use those computational algorithms and tools to achieve fast and reliable computations in sequence analysis.

In this paper, we are focusing on detecting outliers in composite data points (e.g., segmented genome of flu virus). An outlier is a data point that diverge from the majority of other data points in terms of its measured features \[12\]. In addition, finding patterns of data points that do not confirm to the expected feature measurements are a research challenge. There are many applications to the outlier’s detection, for example, detecting tumors in magnetic resonance imaging, finding frauds in health care insurance, or detecting biodiversity in viruses.

In this section, we presented the most popular alignment-free techniques, and a brief introduction about outlier detection. In the next section, we shall present an outlier detection approach for identifying anomalies in segmented genomes of the flu virus.

### Table 1. Comparison between alignment-free and alignment-based techniques \[3,5,10,11\]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Alignment-based</th>
<th>Alignment-free</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input data</td>
<td>Sequences</td>
<td>Data-vectors</td>
</tr>
<tr>
<td>Assumptions about data</td>
<td>Required</td>
<td>Not-required</td>
</tr>
<tr>
<td>Computational scheme</td>
<td>Dynamic programming</td>
<td>Distance-measures</td>
</tr>
<tr>
<td>Time complexity</td>
<td>Quadratic</td>
<td>Linear</td>
</tr>
<tr>
<td>Applications</td>
<td>Sequence comparison</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td></td>
<td>Phylogenetic tree</td>
<td>Phylogenetic tree</td>
</tr>
<tr>
<td></td>
<td>Function prediction</td>
<td>General mapper</td>
</tr>
<tr>
<td></td>
<td>Genome assembly</td>
<td>Genome assembly</td>
</tr>
<tr>
<td></td>
<td>Reads correcting errors</td>
<td>Reads error correction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transcript quantification</td>
</tr>
<tr>
<td></td>
<td>Metagenomics</td>
<td>Metagenomics</td>
</tr>
</tbody>
</table>

### Methods

In this section, we shall present an approach to identify outliers in a dataset of composite data points. A composite data point is a dataset (e.g., set of data-vectors or set of sequences). We shall present two directions: (1) supervised and (2) unsupervised learning modes. In case of the existing training data, the approach can be designed using sequential computational phases. The first computational phase is to map each composite data point into a feature space by defining \((p \times 1)\) feature vector. Each composite data point can be mapped into a set of data vectors. Those data vectors are extracted from heterogeneous sequences; therefore, the base composition of nucleotide distribution is expected to be heterogeneous. In this context, the next phase is to build an information structure for each composite data point. One of the most popular information structures is the variance-covariance structure. Measuring the distance between any two information structures can be achieved by defining a distance measure or metric. The metric space is defined as a metric and a class of matrices, where each matrix represents an information structure of a set of data vectors with unknown distribution.

Daoud \[13\] proposed a solution for the composite data points proximity problem. The solution defined a new metric space \((\Psi, D_\gamma)\), where \(\Psi\) is a class of composite data points, and \(D_\gamma\) is a metric. \(D_\gamma\) is defined as follows:

\[
D_\gamma = \gamma' (\Sigma^x - \Sigma^y) \gamma = |\lambda_i| > 0
\]

where \(\lambda_i\) is the largest generalized eigenvalue (associated with the generalized eigenvector \(\gamma_i\)) of the matrix \((\Sigma^x - \Sigma^y)\), where \(x_n\) is random vector that measures the occurrences of \(n\)-grams in two composite data points \(i\) and \(j\), such that each one represents an in-
stance of a segmented genome of a flu virus.

Measuring the distance between any two variance-covariance matrices $\Sigma^{(i)}$ and $\Sigma^{(j)}$ of the same random vector will result in distance values. Those distance values represent a random variable. In case of considering more than one feature mapping or feature vector (i.e., projected data into more than one feature space), in this case those distance values represent a random vector. The random vector is a random distance vector and it has a distribution with statistical characteristics, and in this context, we define the concept of the distance distribution paradigm (for more details, see Daoud and Kremer [14]).

Now consider the analysis of distance values as another phase to integrate the computational process to detect the outliers in distance space. In case of multivariate distance-random vector, there are different multivariate techniques that can be implemented in detecting outliers in distance space. By considering those distance-data vectors as observations of a random vector, we can implement one of the multivariate outlier-detection techniques, which is based on Mahalanobis distance. In this context, the outlier can be defined as a data-vector with largest squared Mahalanobis distance. The empirical distribution function of the ordered square distances and theoretical distribution function, in this case $\chi^2$-distribution, can be compared to identify outliers with a specific threshold value (quantile) [15,16]. The computational phases are illustrated in Fig. 1, which are in certain way identical to deep learning approach in the sense of using composite feature spaces [17]. Finally, the sketch of the proposed computational techniques (supervised and unsupervised modes) are shown in Figs. 2 and 3, respectively.

The validity of the proposed computational approaches

The proposed computational approaches are bio-data mining approaches and it is build upon using data-vectors extracted from bio-sequences based on $n$-grams features. Those features are numerical features. The numerical features represent the biological features. The main contribution of this paper is to propose a new computational framework to detect outliers in composite data points using distance space. The extracted distance vectors in the composite feature space are multivariate random vectors. Then implementing the existing multivariate outliers’ techniques on those distance data-vectors is a validated computational process. In fact, there is no-need to validate those existing multivariate statistical outlier detection techniques.

---

**Fig. 1.** The sketch of the proposed computational model. The upper part represents the abstract of mapping, while the lower part represents the problem under analysis and the expected output.
Algorithm 1: Detecting outliers in sets of biosequences using supervised approach.

<table>
<thead>
<tr>
<th>Input: Given a collection of unlabeled sets of biosequences $Q = S_Q^{(1)}, S_Q^{(2)}, ..., S_Q^{(k)}$ and labeled sets of biosequences $A = {S_A^{(1)}, S_A^{(2)}, ..., S_A^{(k)}}$.</th>
<th>Output: Detecting outliers in $S_Q^{(1)}, S_Q^{(2)}, ..., S_Q^{(k)}$. Moreover, the algorithm produces: (i) Data-vectors, (ii) Distance-Values, and (iii) Detected-Outliers.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Compute the data vectors for each unlabeled biosequence-set in $Q$ and for each labeled biosequence-set in $A$. The computation process includes mapping biosequence-sets in $Q$ and $A$ into feature space $R^p \times R^p$, using n-grams, and variance-covariance structure.</td>
<td>2 Find the distance between each biosequence-set in $Q$ and biosequence-set in $A$ using a well defined distance measure $D_{ij}(f_i)$. The outcome of this process distance-vectors.</td>
</tr>
<tr>
<td>3 Analyze the distance-vectors as observations of a random vector embedded in distance space and detect outliers based on the computed distance values. In case of multivariate, there are number of existing algorithms that can be implemented to detect outliers based on the distance vectors.</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. Algorithm 1: Detecting outliers in sets of biosequences using supervised approach.

Algorithm 2: Detecting outliers in sets of biosequences using unsupervised approach.

<table>
<thead>
<tr>
<th>Input: Given a collection of sets of biosequences $F_Q^{(1)}, F_Q^{(2)}, ..., F_Q^{(k)}$</th>
<th>Output: Detecting outliers in $F_Q^{(1)}, F_Q^{(2)}, ..., F_Q^{(k)}$. Moreover, the algorithm produce: (i) Data-vectors, (ii) Variance-Covariance Matrix, (iii) Distance-Values, and (iv) Detecting Outliers.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Compute the variance-covariance matrix for each set of biosequences $F_Q^{(1)}, F_Q^{(2)}, ..., F_Q^{(k)}$. The computation process includes mapping each biosequence in $F_Q^{(i)}$ into feature space $R^p$, and then compute the variance-covariance matrix $\Sigma_i, (i = 1, 2, 3, ..., k)$</td>
<td>2 Find the distance between each pair of sets of biosequence in $F_Q^{(1)}, F_Q^{(2)}, ..., F_Q^{(k)}$ using the distance measure $D_{ij}(y_i)$. The outcome of this process distance-vectors.</td>
</tr>
<tr>
<td>3 Analyze the distance vectors as observations of a random vector embedded in distance space and detect outliers based on the computed distance values. In case of multivariate, there are number of existing algorithms that can be implemented to detect outliers based on the distance vectors.</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Algorithm 2: Detecting outliers in sets of biosequences using unsupervised approach.
Results

In this section, we shall present the robustness of the proposed approaches by showing two experiments. In the first experiment, we downloaded 47 and 46 segmented genomes of the flu virus A and B respectively, from NCBI website [18]. Those segmented genomes are collected between May and December 2016. The first computational phase, the segmented genomes mapped into three feature spaces. Those feature spaces are (1) 1-grams (bases), (2) 2-grams (dimers), and (3) 3-grams (codons). The dimensionality of the considered feature spaces is: 4, 16, and 64. As a second computational phase, the sets of extracted data-vectors are mapped

![Image](https://doi.org/10.5808/GI.2020.18.1.e2)

**Fig. 4.** The output from implementing the proposed supervised outlier detection approach (flu virus A). (A, B) The sub-graphs represent the scatter diagram and the distance distribution of composite data points respectively. (C, D) The sub-diagrams represent outlier detections using different quintiles. The figures are generated by using R-package: mvoutlier. We use the function aq.plot to process the distance data-vectors. In addition, left-upper subfigures showing the data projected into two-dimensional space using the first and second principal components.
Table 2. The output of supervised and unsupervised learning in detecting the outliers of segmented genomes

<table>
<thead>
<tr>
<th>Learning approach</th>
<th>Outliers of outliers of flu virus A</th>
<th>Outliers of outliers of flu virus B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>25, 29</td>
<td>28, 6, 26</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>33, 41, 46, 47</td>
<td>46, 4, 45</td>
</tr>
<tr>
<td></td>
<td>45, 44, 43, 42</td>
<td>43, 44, 42</td>
</tr>
</tbody>
</table>

Fig. 5. The output from implementing the proposed unsupervised outlier detection approach (flu virus A). (A, B) The sub-graphs represent the scatter diagram and the distance distribution of composite data points respectively. (C, D) The sub-diagrams represent outlier detections using different quintiles. The figures are generated by using R-package: mvoutlier. We use the function aq.plot to process the distance data-vectors. In addition, panel A showing the data projected into two-dimensional space using the first and second principal components.
Fig. 6. The output from implementing the proposed supervised outlier detection approach (flu virus B). (A, B) The sub-graphs represent the scatter diagram and the distance distribution of composite data points respectively. (C, D) The sub-diagrams represent outlier detections using different quintiles. The figures are generated by using R-package: mvoutlier. We use the function aq.plot to process the distance data-vectors. In addition, panel A showing the data projected into two-dimensional space using the first and second principal components.
Fig. 7. The output from implementing the proposed unsupervised outlier detection approach (flu virus B). (A, B) The sub-graphs represent the scatter diagram and the distance distribution of composite data points respectively. (C, D) The sub-diagrams represent outlier detections using different quintiles. The figures are generated by using R-package: mvoutlier. We use the function aq.plot to process the distance data-vectors. In addition, panel A showing the data projected into two-dimensional space using the first and second principal components.
into the form of variance-covariance matrices. By implementing the metric given in Eq. (1). We measured the proximity of any two variance-covariance matrices using the largest generalized eigenvalue. In the third phase, we analyzed the distance values as observations of (3 × 1) random vector. Based on the squared Mahalanobis distance of the distance data-vectors, the segmented genomes with largest squared Mahalanobis distance are given in Table 2. In the second experiment, we implemented the unsupervised approach to detect outliers in the same dataset of segmented genomes, and the segmented genomes with largest squared Mahalanobis distance are shown in Table 2.

The output of the proposed approach is illustrated in Figs. 4–7. The outliers of segmented genomes are identified efficiently using supervised learning approach compared with unsupervised learning approach (Table 2). This conclusion is inferred based on the number of identified outliers.

Finally, in this section, we presented the results of implementing the proposed outlier detection approach. In next section, we shall present conclusions and future work.

Discussion

In this paper, we proposed a new approach to detect outliers in segmented genomes of the flu virus. The flu virus has eight segments that can be encoded into 10–11 proteins, where each protein has different biological function and consequently has different nucleotide composition. Those segmented genomes are heterogeneous by nature. The computational challenges are solved in systematic approach, as feature mapping into the feature space, composite feature representation as variance-covariance matrices, defining a metric space to measure the distance between any two variance-covariance matrices, and finally analyzing those distance-values in the feature space. To evaluate the approach, we implemented it using two datasets: (1) 47 segmented genome of the flu virus A and, (2) 46 segmented genomes of the flu virus B. The output of the proposed approach shows the difference between supervised learning and unsupervised learning, and we identified the weaknesses and strengths of each learning mode.

ORCID

Mosaab Daoud: https://orcid.org/0000-0002-5829-641X

Conflicts of Interest

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

Acknowledgments

I revised the final version of this paper in Lebach, Saarland State, Germany in 2019.

References

13. Daoud M. A new variance-covariance structure-based statistical pattern recognition system for solving the sequence-set proximity problem under the homology-free assumption [dissertation].


A note on the distance distribution paradigm for Mosaab-metric to process segmented genomes of influenza virus

Mosaab Daoud

Independent Research Scientist, Toronto, ON M1S1B2, Canada

In this paper, we present few technical notes about the distance distribution paradigm for Mosaab-metric using 1, 2, and 3 grams feature extraction techniques to analyze composite data points in high dimensional feature spaces. This technical analysis will help the specialist in bioinformatics and biotechnology to deeply explore the biodiversity of influenza virus genome as a composite data point. Various technical examples are presented in this paper, in addition, the integrated statistical learning pipeline to process segmented genomes of influenza virus is illustrated as sequential-parallel computational pipeline.

Keywords: composite data point, distance distribution paradigm, Mosaab-metric space, segmented genome of influenza virus

Introduction

One of the main advances in bioinformatics, computational biology, and biotechnology is the sequence-set analysis. It is a new research direction parallel to sequence analysis. The main idea behind it is to analyze composite data points in data space, feature space or distance space. A composite data points is a dataset, for example set of feature vectors, set of sequences. This generalized concept proposed in [1,2]. Now, dealing with segmented genomes of influenza virus as composite data points has different aspects: biodiversity, bio-intelligent system, genomic variation, and vaccine efficiency.

The influenza viruses have a negative impact on public health and still creating threats for different life aspects. The early pandemic of H1N1 flu virus started in 1918 [3]. Recent advances in bioinformatics and biotechnology have extended and expanding the insights of analyzing the segmented genome of the flu virus and exploring the influenza biology [2]. Influenza virus has the following features: (1) it is a single RNA-stranded enveloped virus, (2) its genome is segmented, and it has eight segments, each segment can be encoded to one or two proteins, (3) it is a negative-sense virus, and (4) it can be rated as highly-mutated genome [4]. The virus can infect various hosts, and it has different types and subtypes. The subtypes can be identified according to its surface proteins, haemagglutinin (HA) and neuraminidase (NA) [2,3]. Now, there are 18 HA and 11 NA distinct surface proteins [3,4]. The source of genetic variation is two processes: (1) antigenic drift, or (2) antigenic shift.

As defined by Daoud’s study [1,2], a segmented genome of influenza virus is a composite data point. A composite data point is a dataset from unknow or a well know probability distribution. In machine learning and data mining there are many algorithms that
they can be used to analyze, visualize, classify and cluster data points. Usually regular data points, for example, data vectors, univariate data points, and sequences. Processing composite data points is another complicated computational task for existing computational pipelines. Building a statistical learning computational pipeline has several computational challenges [5]. As defined in James et al. [6], statistical learning is a set of unsupervised and supervised computational algorithms that can be used in processing datapoints to extract knowledge and deep understanding about the relationship and structure of data. In other words, statistical learning focuses on learning the relationship and structure from data vectors (i.e., observations of a feature vector). In deep statistical learning, we learn about the relationship and structure of data from distance data vectors after mapping datapoints into different feature spaces using the extension principle of data life cycle [1]. Developing a statistical learning computational pipeline for analyzing the segmented genomes of flu virus is a completed task. One of the computational aspects in statistical learning is to analyze the distance distribution paradigm for the datapoints under consideration [6-8]. A distance distribution paradigm is defined as the probability distribution of a distance measure or metric [6]. In other words, the distance measure or metric is defined as a random variable or random vector [6-8]. In the next section, we shall present a note on the distance distribution paradigm for Mosaab-metric space.

**Technical Implementation**

In this section, we shall present technical analysis of the deep distance distribution for Mosaab-metric to process segmented genomes of flu virus as composite datapoints, and by using the following three feature spaces: 1-grams, 2-grams, and 3-grams. Mapping each composite data point into various feature spaces by using n-grams technique (in this case n = 1, 2, and 3) has the following outcomes: data-vectors are embedded into feature spaces. The feature spaces are high dimensional spaces. Each composite data point is represented by a dataset, and each dataset is a set of data-vectors. Transforming each set of data-vectors to variance-covariance structure is another information structure, and the outcomes are matrices. Finding the distance between each matrix in the testing dataset and each matrix in the training dataset has the following outcomes: distance values. By using the extension principle of the data life cycle, and in this case by consider three feature spaces (deep statistical learning), the combined outcomes are \((3 \times 1)\) distance-data vectors. The distance-data vectors represent a random vector. The random vector has a probability distribution, and since the extracted information is a combination of three feature spaces, then the probability distribution is called the deep distance distribution (or the deep distance paradigm). Now, we shall consider three technical cases about this implementation. We have downloaded 30 segmented genome of influenza virus A, 30 segmented genome of influenza virus B from NCBI-Influenza Virus Database as training datasets [9]. In addition, we have downloaded 108 segmented genome of influenza virus A and B from NCBI-Influenza Virus Database as testing dataset [9]. In case 1, the sizes of training datasets are: 30 segmented genomes for flu A virus, and 30 segmented genomes for flu B virus respectively. The size of testing dataset is 108 segmented genome of influenza virus A and B. Fig. 1 illustrates the analytics of deep statistical learning approach in dealing with composite data points. The first subfigure has two aspects (Fig. 1A): combining two feature spaces (1 and 2 grams feature vectors) to produce two distance values (or \((2 \times 1)\) distance vector) with respect to a training dataset using the extension principle. In other words, the concept of deep statistical learning is based on extension of the data life cycle. The second subfigure has the same pervious aspects, and by combining three feature spaces, therefore, the result is a \((3 \times 1)\) distance vector (Fig. 1B). It should be noted that a distance vector is a random vector and it has observations, and those observations are called distance-data vectors. For each feature space, the distance vector as a random vector has a probability distribution, and in this case, it is called the distance distribution paradigm. The distance distribution paradigm for 1-grams, 2-grams, 3-grams feature spaces are illustrated in Fig. 1C, 1D, and 1E, respectively. From these subfigures we can conclude the following: each subfigure has two peaks, each peak represents a class, influenza A virus and influenza B virus. One bell-shaped density curve skewed to the right and another curve skewed to the left. One class has more dispersion than the other, which is in this case influenza A virus. Now consider a training dataset with lack of diversity. Suppose we have two training datasets that represent only one class (in this case influenza A virus), one has 30 composite datapoints and another one has 10 composite datapoints, hence, Figs. 2 and 3 represent the outcomes from these two experiments respectively. Based on the subfigures of Figs. 2 and 3, we have different dispersion maps, two classes, and two peaks. This note has effective conclusions about the impact of size and diversity of datasets on classification results using the distance distribution paradigm.

In this section we presented the technical notes about the distance distribution paradigm for Mosaab-metric using 1, 2, and 3 grams feature extraction techniques to analyze composite data points in high dimensional feature spaces. In the next section we shall present the conclusions.
Fig. 1. The distance distribution paradigm for Mosaab-metric using 1, 2, and 3 grams feature extraction techniques to analyze composite data points in high dimensional feature spaces (case: 60 composite data points represent two classes of influenza virus, class A and class B). (A) Scatter-plot of composite data points in 2-dimensional space. (B) Scatter-plot of composite data points in 3-dimensional space. (C) The distance distribution paradigm for 1-grams feature space. (D) The distance distribution paradigm for 2-grams feature space. (E) The distance distribution paradigm for 3-grams feature space.
Fig. 2. The distance distribution paradigm for Mosaab-metric using 1, 2, and 3 grams feature extraction techniques to analyze composite data points in high dimensional feature spaces (case: 30 composite data points represent one class of influenza virus, class A). (A) Scatter-plot of composite data points in 2-dimensional space. (B) Scatter-plot of composite data points in 3-dimensional space. (C) The distance distribution paradigm for 1-grams feature space. (D) The distance distribution paradigm for 2-grams feature space. (E) The distance distribution paradigm for 3-grams feature space.
Fig. 3. The distance distribution paradigm for Mosaab-metric using 1, 2, and 3 grams feature extraction techniques to analyze composite data points in high dimensional feature spaces (case: 10 composite data points represent one class of influenza virus, class A). (A) Scatter-plot of composite data points in 2-dimensional space. (B) Scatter-plot of composite data points in 3-dimensional space. (C) The distance distribution paradigm for 1-grams feature space. (D) The distance distribution paradigm for 2-grams feature space. (E) The distance distribution paradigm for 3-grams feature space.
Conclusions

In this paper we presented the distance distribution paradigm for Mosaab-metric using three feature spaces: 1-grams, 2-grams, and 3-grams. We technically showed the impact of the size and diversity of training dataset on the classification results. We successfully analyzed the distance distribution of Mosaab-metric space as the most recent metric space in statistical learning research field. This part of analytics (as analytical techniques) about the distance distribution and the dispersion maps is expected to be in a integrated statistical learning computational pipeline for processing and analyzing composite data points (in this case segmented genome of influenza virus, see Fig. 4). The pipeline is sequentially partitioned into components. The first component is to map the segmented genomes into feature spaces (parallel computational mode can be applied), the second component can be executed in parallel mode, and it has different tools (algorithms/techniques). These tools can be summarized as: classification, clustering, outlier detection, and visualization. In the future work, we shall discuss, and present other computational algorithms and/or tools that will be included in this integrated pipeline.

References

7. Daoud M, Kremer SC. A new distance distribution paradigm to detect the variability of the influenza-A virus in high dimensional

Conflicts of Interest

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

ORCID

Mosaab Daoud: https://orcid.org/0000-0002-5829-641X


PAIVS: prediction of avian influenza virus subtype

Hyeon-Chun Park1,2,3, Juyoun Shin2, Sung-Min Cho3, Shinseok Kang4, Yeun-Jun Chung2,3*, Seung-Hyun Jung5,6**

1Department of Biomedicine & Health Sciences, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea
2Department of Microbiology, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea
3Integrated Research Center for Genome Polymorphism, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea
4Chungbuk Veterinary Service Laboratory, Chungju 27336, Korea
5Cancer Evolution Research Center, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea
6Department of Biochemistry, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea

Highly pathogenic avian influenza (HPAI) viruses have caused severe respiratory disease and death in poultry and human beings. Although most of the avian influenza viruses (AIVs) are of low pathogenicity and cause mild infections in birds, some subtypes including hemagglutinin H5 and H7 subtype cause HPAI. Therefore, sensitive and accurate subtyping of AIV is important to prepare and prevent for the spread of HPAI. Next-generation sequencing (NGS) can analyze the full-length sequence information of entire AIV genome at once, so this technology is becoming a more common in detecting AIVs and predicting subtypes. However, an analysis pipeline of NGS-based AIV sequencing data, including AIV subtyping, has not yet been established. Here, in order to support the pre-processing of NGS data and its interpretation, we developed a user-friendly tool, named prediction of avian influenza virus subtype (PAIVS). PAIVS has multiple functions that support the pre-processing of NGS data, reference-guided AIV subtyping, de novo assembly, variant calling and identifying the closest full-length sequences by BLAST, and provide the graphical summary to the end users.

Keywords: AIV subtypes, avian influenza virus, next-generation sequencing, viral genome

Availability: PAIVS is available at http://ircgp.com/paivs.
2018 in South Korea. The outbreaks of the 2016–2017 and 2017–2018 winter seasons, caused by novel reassortant clade 2.3.4.4 H5N6 viruses, resulted in loss of one billion birds in 440 farms in South Korea (https://www.kahis.go.kr/). Therefore, accurate and rapid subtyping of AIV is important to prepare and prevent for the spread of HPAI.

There are several methods to detect the AIVs such as rapid influenza diagnostic test (RIDT), nucleic acid-based tests (NATs), and next-generation sequencing (NGS) [5]. RIDTs, the most widely used method, can detect AIVs rapidly but do not provide information on AIV subtype or specific virus strain information. NATs, which include reverse transcriptase-PCR and loop-mediated isothermal amplification-based assay, detect virus-specific DNA or RNA sequences. Although most NATs are more sensitive and specific than RIDTs, they are laborious and time-consuming tests to get subtype information. In contrast to these methods, NGS can analyze the full-length sequence information of all eight AIV segments at once, so it can identify subtypes sensitively and accurately. Furthermore, since single nucleotide variants and evolutionary analyses are possible with the complete genome sequencing data, there are many advantages to identify the origin and pathogenicity of AIVs. However, an analysis pipeline of NGS-based AIV sequencing data, including AIV subtyping, has not yet been established.

In this study, we developed a user-friendly tool, named prediction of avian influenza virus subtype (PAIVS), to support the pre-processing of NGS data and its interpretation. PAIVS analyzes the NGS data for pre-processing, reference-guided AIV subtyping, de novo assembly and variant calling and provides graphical summary for subtype identification. In addition, PAIVS supports the BLAST (Basic Local Alignment Search Tool) function to identify the closest full-length sequences that can be used as genetic resources for downstream analysis, such as phylogenetic analysis.

Overview Pipeline

PAIVS is an automated pipeline that analyzes AIV NGS data and consists of five steps; pre-processing, reference-guided alignment or de novo assembly, subtyping, variant calling, and identifying the closest sequences. We implemented PAIVS using python and its workflow is illustrated in Fig. 1. First, PAIVS takes FASTQ files from paired-end viral genome sequencing data as input. In the pre-processing step, PAIVS trims the sequence reads to remove the sequences with adaptor or low base quality. Trimmed reads are then aligned to the host reference genome, such as avian or human, to remove the sequences from the host. In the next step, unmapped reads are aligned to the AIV reference genome and subsequently identify subtypes and variants based on the sequence composition and coverage. Users can also select de novo assembly option to get the closest full-length sequences and use them for downstream analysis. All results for each step is stored in its own directory.

Data Pre-processing

In the pre-processing step, TRIMMOMATIC [6] tool trims the sequence reads to remove the sequences with adaptor or low base quality. Host reads subtraction by read mapping is performed by using the HISAT2 [7] or BWA [8] aligner against host organism genome. SAMtools software package [9] ‘view’ option is used for extracting unmapped read. Unmapped viral reads are then mapped to the AIV reference genome using the HISAT2 or BWA aligner. When we compared the mapping rate, memory usage, and run time between two aligners, BWA mem algorithm showed reasonable run time and mapping rate to the reference AIV genome (Supplementary Table 1). HISAT2 was faster and used fewer resources than BWA, but more than 90% of sequencing reads were not mapped to the AIV genome. Therefore, we recommend to use BWA aligner unless user has ultra high depth sequencing data (> 10,000 ×).

Prediction of AIV Subtype

In order to predict the AIV subtypes, PAIVS calculates the segment mean coverages as a coverage value for each 18 different HA and 11 different NA segments by using SAMtools. Coverage value means the sum of the coverages is then normalized by dividing by the total length of segment.

$$\text{Segment Mean Coverage} = \frac{\sum_{k=1}^{2C} D_{k/p}}{p \cdot \text{Seq. length}}$$

where $Dp$ is depth of each genomic position, $p$ is segment length, $k = 1, \ldots, p$ is each genomic position.

The coverage values are displayed on the linux terminal console and saved a comma separate text file (Fig. 2A). In addition, coverage depth for each HA and NA segments is saved as a image file in png format (Fig. 2B and 2C).

Variants Calling and De Novo Assembly

SAMtools and BCFtools [10] are implemented in PAIVS to detect the single nucleotide variant, which are known to be faster and use less memory than other variant calling methods such as Varscan2 [11] and HaplotypeCaller [12]. The result is saved as a text file in variant call format, an example is shown in Fig. 2. Regarding de
novo assembly, Iterative Virus Assembler (IVA) designed to assemble virus genomes [13] is implemented in PAIVS. When the viral FASTA file is generated as an output of the IVA, the closest full-length sequences are then identified by BLAST [14] (Fig. 2E). Both variants and BLAST results can be used as genetic resources for further downstream analysis, such as phylogenetic analysis and clade classification.

**Conclusion**

In this paper, we introduce the PAIVS which is a user-friendly tool with multiple functions that support the pre-processing of NGS data, reference-guided AIV subtyping, de novo assembly, variant calling, and identifying the closest full-length sequences by BLAST and provide the graphical summary to the end users. In addition, PAIVS can be applied to other viral genomes for viral genome detection. By replacing the reference genome in the configuration file, users can easily perform read mapping, variant calling, de novo assembly, and identifying the closest full-length sequences with BLAST. However, the viral subtyping and graphical summary functions of PAIVS were specifically designed for AIVs, so there is a limit to applying these functions to other viral genome. Considering that NGS technology is becoming a more common for detecting AIVs and predicting subtypes, PAIVS can be helpful for beginners who are not familiar with NGS-based AIV data processing. PAIVS is freely available at http://ircgp.com/paivs.
Fig. 2. Examples of prediction of avian influenza virus subtype (PAIVS) output. To evaluate the PAIVS, we analyzed a H5N6 avian influenza virus (AIV) confirmed by PCR and Sanger sequencing test. For the reference-guided read mapping, AIV genome sequences were obtained from Influenza Virus Resource Database (https://www.ncbi.nlm.nih.gov/genomes/FLU) and customized to analyze 16 different hemagglutinin (HA) and 9 different neuraminidase (NA) segments (H1 through H16 and N1 through N9). (A) PAIVS generates the coverage value for each HA and NA segments and presents the predicted subtype. As expected, analyzed sample was predicted as H5N6 subtype. (B, C) PAIVS also generates image files for coverage depth of HA (B) and NA (C) segments. Coverage depth of H5 and N6 segments was much higher than other segments. (D) Detected single nucleotide variants are reported as a variant call format. (E) The sequences closest to the analyzed segment are reported with BLAST score and E-value.

**Authors’ Contribution**

Conceptualization: YJC, SHJ. Data curation: HCP, JS, SMC, SK. Formal analysis: HCP, SHJ. Writing - original draft: HCP, JS, YJC, SHJ. Writing - review & editing: HCP, JS, YJC, SHJ.

**Conflicts of Interest**

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

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) through Animal Disease Management Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (318046-2).

Supplementary Materials

Supplementary data including one table can be found with this article online at https://doi.org/10.5808/GI.2020.18.1.e4.

References

3. Claes F, Morzaria SP, Donis RO. Emergence and dissemination of clade 2.3.4.4 H5Nx influenza viruses: how is the Asian HPAI H5 lineage maintained. Curr Opin Virol 2016;16:158-163.
HisCoM–PCA: software for hierarchical structural component analysis for pathway analysis based using principal component analysis

Nan Jiang\(^1\), Sungyoung Lee\(^2\), Taesung Park\(^{1,3*}\)

\(^1\)Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Korea  
\(^2\)Center for Precision Medicine, Seoul National University Hospital, Seoul 08826, Korea  
\(^3\)Department of Statistics, Seoul National University, Seoul 08826, Korea

Introduction

In genome-wide association studies (GWAS), researchers have identified many single-nucleotide polymorphisms (SNPs) associated with the traits of interest (phenotypes) [1]. However, these SNPs sometimes may sometimes suffer from a lack of biological interpretation [2]. To enhance interpretation of SNP association results, many gene-based analysis and pathway-based analysis have been widely performed in GWAS. For examples, PHARAOH was developed for pathway analysis of rare variants, and hierarchical structural component analysis of gene-gene interactions (HisCoM-GGI) was proposed for gene-gene interaction analysis of common variants [3,4].

Recently, we presented a hierarchical structural component model (HisCoM) for pathway analysis of common variants (HisCoM-PCA) to identify pathways associated with traits [5]. HisCoM-PCA is based on principal component analysis (PCA) for dimensional reduction of single nucleotide polymorphisms in each gene, and the HisCoM for pathway analysis. In this study, we developed a HisCoM-PCA software for the hierarchical pathway analysis of common variants. HisCoM-PCA software has several features. Various principle component scores selection criteria in PCA step can be specified by users who want to summarize common variants at each gene-level by different threshold values. In addition, multiple public pathway databases and customized pathway information can be used to perform pathway analysis. We expect that HisCoM-PCA software will be useful for users to perform powerful pathway analysis.

Keywords: genome-wide association study, hierarchical structural component model, pathway analysis, principal component analysis

Availability: HisCoM-PCA is available on the website (http://statgen.snu.ac.kr/software/HisCoM-PCA/index.html).
al reduction step of HisCoM-PCA, various principle component scores (PC) selection criteria may be used. The criterion may be defined by a threshold of cumulative proportion of variances. It can also be defined by using only the first PC for each gene. In the pathway analysis step, multiple published pathway databases and specific combination of pathways can be used to identify pathways associated with the traits of interest.

In our previous study, we used only pathway information of the Kyoto Encyclopedia of Genes and Genome pathway database [6] and adopted two criteria to select PC: (1) using only the first PC and (2) using the PCs whose cumulative proportion of variances are more than 30%. To enable researchers to perform various pathway analysis, we developed HisCoM-PCA software for allowing the users to set the PC selection criterion and pathway information flexibly.

Implementation

The workflow of the HisCoM-PCA software is shown in Fig. 1. The HisCoM-PCA method has been proposed for pathway analysis of common variants by constructing a hierarchical model using SNP-gene-pathway information. The HisCoM-PCA method consists of two steps: (1) dimensional reduction of SNPs by PCA and (2) pathway analysis with a hierarchical component model. In the first dimensional reduction step of HisCoM-PCA software, the user can define the number of PCs for each gene by one of the following two options: (1) the threshold of cumulative proportion of variances and (2) only the first PC. Using the selected PCs, pathway analysis can be performed based on the user-entered pathway datasets. In the second pathway analysis step, HisCoM-PCA utilizes ridge-type penalization and performs a permutation test to estimate the gene and pathway effects on the phenotypes.

Input file

The HisCoM-PCA software takes four inputs: (1) a SNP dataset in csv file or PLINK format files, (2) a txt file with phenotype and covariate(s), (3) a set file that consists of two columns for gene name and SNP id, and (4) a set file that consists of two columns for pathway name and gene name. Furthermore, the program also accepts the published pathway databases in MsigDB [7,8].

Output file

The HisCoM-PCA program generates the following output files: (1) a ’[prefix].gene-pca_summary.csv’ file that contains the number of SNPs and PCs for each gene, (2) a ’[prefix].gene.ressum.csv’ file that contains pathway name, gene name, number of permutation, weight of gene for the pathway, gene coefficient, and permutation p-value of gene, and (3) a ’[prefix].pathway.ressum.csv’ file that contains pathway name, number of permutation, number of genes in each pathway, pathway coefficient, and permutation p-value of pathway.

---

**Fig. 1.** The workflow of the HisCoM-PCA software. PCA, principal component analysis; SNP, single nucleotide polymorphism; HisCoM-PCA, hierarchical structural component model for pathway analysis of common variants.
Conclusion

We introduced our HisCoM-PCA software for pathway analysis of common variants in GWAS. HisCoM-PCA software supports pathway analyses using multiple candidate pathways and customized PC selection criterion. We expect that our HisCoM-PCA software is useful for users to perform various pathway analyses. 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).

ORCID

Nan Jiang: https://orcid.org/0000-0003-0705-6173
Sungyoung Lee: https://orcid.org/0000-0003-3458-1440
Taesung Park: https://orcid.org/0000-0002-8294-590X

Authors’ Contribution

Conceptualization: TP. Data curation: NJ. Formal analysis: NJ. Funding acquisition: TP. Methodology: NJ, SL, TP. Writing – original draft: NJ. Writing – review & editing: TP.

Conflicts of Interest

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

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant number: HI16C2037) and the Bio-Synergy Research Project of the Ministry of Science, ICT and Future Planning through the National Research Foundation (grant number: 2013M3A9C4078158).

References

Accelerating next generation sequencing data analysis: an evaluation of optimized best practices for Genome Analysis Toolkit algorithms

Karl R. Franke*, Erin L. Crowgey

Department of Pediatrics, Nemours Alfred I duPont Hospital for Children, Wilmington, DE 19803, USA

Advancements in next generation sequencing (NGS) technologies have significantly increased the translational use of genomics data in the medical field as well as the demand for computational infrastructure capable processing that data. To enhance the current understanding of software and hardware used to compute large scale human genomic datasets (NGS), the performance and accuracy of optimized versions of GATK algorithms, including Parabricks and Sentieon, were compared to the results of the original application (GATK V4.1.0, Intel x86 CPUs). Parabricks was able to process a 50x whole-genome sequencing library in under 3 h and Sentieon finished in under 8 h, whereas GATK v4.1.0 needed nearly 24 h. These results were achieved while maintaining greater than 99% accuracy and precision compared to stock GATK. Sentieon’s somatic pipeline achieved similar results greater than 99%. Additionally, the IBM POWER9 CPU performed well on bioinformatic workloads when tested with 10 different tools for alignment/mapping.

Keywords: clinical genomics, Genome Analysis Toolkit, GPUs, next generation sequencing, variant detection

Introduction

As the price of next generation sequencing (NGS) decreases and the data footprint increases, compute power is a major limitation. The bioinformatics processing of NGS data is routine in many translational research and precision medicine efforts with the Genome Analysis Toolkit (GATK) from the Broad Institute and their “Best Practices” workflows [1-3] being widely accepted as standards for germline and somatic short variant discovery.

The majority of computational algorithms specific for NGS were developed for use in a CPU environment with standard threading techniques. However, as NGS becomes more routine, the need to offer faster processing is essential. Despite the fact that x86_64 (henceforth x86) CPUs from Intel represent the most common processing architecture in both the server space as well as desktop workstations, alternative architectures exist which offer some advantages such as POWER9 from IBM and GPUs from NVIDIA.

This article compares the speed, accuracy, and scalability of the official GATK algorithms running on Intel x86 CPUs, a POWER9 optimized version provided by IBM, a GPU optimized version provided by Parabricks, and an x86 optimized version provided by Sentieon [4]. Additionally, the ability of POWER9 CPUs to handle bioinformatic workloads is assessed by comparing the performance of a number of other bioinformatic tools compiled to run on both x86 and POWER9 systems.
Methods

Hardware configuration
All x86 jobs were run on an HPC cluster node (Thinkmate HDX XB24-5260V4) powered by two 8-core Intel Xeon E5-2620 v4 CPUs and 256GB ECC memory. All POWER9 jobs were run on an HPC cluster node (IBM AC922) powered by two 16-core POWER9 CPUs (Part #02CY417), 512GB ECC memory, and four Nvidia TESLA V100 GPUs each with 32GB HBM2 memory. Disk I/O was handled via a DDN storage appliance connected to each node via InfiniBand.

Software configuration
Job scheduling on the HPC cluster was handled by SLURM. GATK software was run with parameters according to their best practices workflows with performance optimizations found by Heldenbrand et al. [5] with any deviations indicated as such. The Sentieon and Parabricks pipelines were run according to their included documentation. Runtimes were calculated using the GNU time utility.

Sensitivity and precision calculations
GATK’s Concordance tool was utilized to compare the tested Variant Call Format (VCF) outputs to different reference VCFs depending on the comparison being made. The reference VCF is treated as a truthset, and the number of true-positive, false-positive, and false-negative variants in the query VCF are counted. Sensitivity is calculated as true-positives/(true-positives + false-negatives). Precision is calculated as true-positives/(true-positives + false-positives).

POWER9 testing workloads
For the RNA sequencing (RNAseq) benchmarks with STAR v2.7.0c [6], Tophat v2.1.1 [7], HISAT2 v2.1.0 [8], and BBMap v38.34 [9], 15 paired end RNAseq libraries were downloaded from the ENA database (Study PRJEB23554). FASTQ files trimmed via BBduk before being mapped to GRCh38 using each tool’s default settings. For BBMap and Tophat only one library was used for benchmarking, while HISAT2 and STAR performed mapping using all 15. For the splice aware aligners, STAR, HISAT2, and Tophat, the Gencode v29 annotation GTF file was used.

For the DNA short read mapping benchmarks with BWA MEM v0.7.17-r1188 [10], Bowtie v1.2.2 [11], and Bowtie2 v2.3.5.1 [12], the NA12878 50x WGS library was used. The FASTQ files were mapped to GRCh38 using each tool’s default settings.

For blastx v2.8.1 [13], all *Brachypodium distachyon* transcript sequences from the BD21 v2.1 annotation were queried against the *Oryza sativa* v7 MSU protein database. For blastn v2.8.1 [13], the same transcript sequences were queried against a local copy of NCBI’s nt database. For pblat v36x2 [14], all human transcript sequences from the Gencode v29 release were used as a query against GRCh38.

Results

Germline variant detection pipeline speed and scalability
The pipelines for germline variant discovery were tested using a 50× WGS library from the NA12878 sample of Illumina’s Platinum Genomes project (ENA study PRJEB3381) [15]. The baseline run of GATK4 took over 100 h on the x86 system mainly due to the HaplotypeCaller step (Table 1). The most significant speed improvements (Fig. 1A) were demonstrated with single threaded PairHMM and splitting the hg38 genome into 32 scattered intervals (two per x86 core); this also allowed for parallelization of the BaseRecalibrator and ApplyBQSR steps. This manual form of multithreading brought the nearly 107-h runtime down to just under 24 h (Table 1). On the POWER9 with SMT4, the most efficient use of resources was observed when 2 intervals were run per core with each using 2 PairHMM threads (Fig. 1A) which yielded a ~10 h decrease in runtime (~42.4%) compared to x86 with lower times in the BWA, MarkDuplicates, and HaplotypeCaller steps.

The Sentience DNAseq/DNAscope pipelines reduced runtimes even further to under 8 h with much faster BWA, MarkDuplicates, and HaplotypeCaller steps. Sentience’s version of BWA finished ~3 h earlier (~38.2%) than stock on the same x86 node. Marking duplicate reads only took 39 minutes whereas previously mentioned pipelines measured that time in hours. Finally, the HaplotypeCaller step took just over 1 h for Sentience’s pipeline. Nearly identical runtimes were observed for DNAseq and DNAscope pipelines with the latter being Sentience’s proprietary algorithm and not a reimplementation of HaplotypeCaller. The fastest analysis performed was Parabricks which yielded a genomic VCF (GVCF) file in 2 h 21 min. When set to output VCF directly, with no GVCF conversion, a faster runtime of 2 h 3 min was observed.

Stock GATK did not scale linearly due to having multiple single threaded steps (Fig. 1B); this is in contrast to Sentience’s implementation which was almost completely multithreaded allowing for much better scalability. In GPU scalability testing, Parabricks hit diminishing returns when adding a fourth GPU which only reduced runtime by ~23 min (~13.7%) whereas going from two to three GPUs reduced the runtime by about an hour (~25.8%).

Germline variant detection pipeline accuracy
To determine the accuracy of the various analyses, GATK’s concor-
dance tool was utilized to compare the final VCF outputs to GATK4's baseline VCF and the NA12878 Platinum Genomes truthset VCF (Table 2). Parabricks most closely matched the GATK baseline run with 99.8% accuracy and precision for single nucleotide polymorphisms (SNPs) and 99.1% accuracy and precision for INDELs. When the output was compared with the version of GATK Parabricks was designed to match (v4.0.4), accuracy and precision were both 99.99% (data not shown). Sentieon’s DNAseq pipeline scored 99.7% for accuracy and precision in calling SNPs, but had around one percentage point decrease for INDEL calling.

In the truthset comparison, Sentieon’s DNAscope yielded the highest SNP sensitivity by calling ~7,000 SNPs that the other callers missed; however, this increase in sensitivity comes at the cost of a loss in precision due to an additional ~4,000 false-positive SNPs. Despite Sentieon’s DNAseq pipeline having the lowest sensitivity and precision for INDELs in the previous comparison, here it ranks highest on both fronts due to it calling ~200–300 more true-positive INDELs that the other pipelines missed and having ~500 fewer false-positives.

### Somatic variant detection pipeline speed

To test the performance of the somatic variant calling pipelines, the cancer like mixture of two Genome in a Bottle samples, NA12878 and NA24385, constructed by the Hartwig Medical Foundation and Utrecht Medical Center [16], was utilized. The baseline Mutect2 run took the longest at over 118 h due to the minimally parallelized Mutect2 step. Manually parallelizing this step using scattered intervals allowed for a reduction in total runtime to ~23.5 h on one of our x86 nodes and ~18.5 h on the POWER9 node. Sentieon’s TNseq pipeline reduced this even further to ~7.5 h and their TNscope pipeline was the fastest at ~3.5 h.

Of the steps unique to the Mutect2 "Best Practices" pipeline, Mutect2 and GetPileupSummaries account for over 99% of the total time. While GATK v4.1.0 allows for Mutect2 to be manually parallelized via scattered intervals and a subsequent MergeVCF step, there is no built-in method to merge the output of multiple GetPileupSummaries jobs. This fact hurt the scalability of the pipelines on both x86 and POWER9 systems with Sentieon’s pipeline reduced this even further to ~7.5 h and their TNscope pipeline was the fastest at ~3.5 h.

Of the steps unique to the Mutect2 "Best Practices" pipeline, Mutect2 and GetPileupSummaries account for over 99% of the total time. While GATK v4.1.0 allows for Mutect2 to be manually parallelized via scattered intervals and a subsequent MergeVCF step, there is no built-in method to merge the output of multiple GetPileupSummaries jobs. This fact hurt the scalability of the pipelines on both x86 and POWER9 systems with Sentieon’s pipeline reduced this even further to ~7.5 h and their TNscope pipeline was the fastest at ~3.5 h.

### Somatic variant detection pipeline accuracy

The accuracy of the Mutect2 based pipelines was analyzed in a similar fashion to the assessment of the HaplotypeCaller based

---

[Table 1. Variant calling pipelines’ speed]

<table>
<thead>
<tr>
<th>Germine pipeline</th>
<th>Germline HaplotypeCaller</th>
<th>Somatic pipeline</th>
<th>Parabricks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>x86 baseline</td>
<td>x86 32-interval</td>
<td>Power9 64-interval</td>
</tr>
<tr>
<td><strong>Times</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MarkDupes</td>
<td>6:35:29</td>
<td>5:35:05</td>
<td>3:05:17</td>
</tr>
<tr>
<td>Samtools Index</td>
<td>1:51:02</td>
<td>1:33:05</td>
<td>1:16:05</td>
</tr>
<tr>
<td>BaseRecalibrator</td>
<td>10:10:20</td>
<td>0:38:20</td>
<td>0:18:01</td>
</tr>
<tr>
<td>ApplyBQSR</td>
<td>7:59:57</td>
<td>0:43:08</td>
<td>0:23:55</td>
</tr>
<tr>
<td>HaplotypeCaller</td>
<td>71:02:51</td>
<td>5:30:32</td>
<td>2:28:12</td>
</tr>
<tr>
<td>CombineGVCVF</td>
<td>0:51:42</td>
<td>0:59:12</td>
<td></td>
</tr>
<tr>
<td>GenotypeGVCFs</td>
<td>0:40:38</td>
<td>0:29:11</td>
<td>0:30:50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Somatic pipeline</strong></th>
<th>Mutect2</th>
<th>TNseq</th>
<th>TNscope</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Times</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MergeVCF</td>
<td>-</td>
<td>0:00:30</td>
<td>0:00:44</td>
<td></td>
</tr>
<tr>
<td>GetPileupSummaries</td>
<td>8:47:05</td>
<td>8:40:35</td>
<td>9:17:00</td>
<td></td>
</tr>
<tr>
<td>CalculateContamination</td>
<td>0:00:11</td>
<td>0:00:11</td>
<td>0:00:13</td>
<td>0:04:52</td>
</tr>
<tr>
<td>FilterMutectCalls</td>
<td>0:02:22</td>
<td>0:03:26</td>
<td>0:02:51</td>
<td></td>
</tr>
</tbody>
</table>

N/A, not available.
Fig. 1. Variant calling pipelines’ scalability. (A) To determine the most efficient way of parallelizing GATK’s HaplotypeCaller, different combinations of scattered intervals and PairHMM OpenMP threads were tested on x86 and Power9 systems. The recorded times also include the CombineGVCFs step. (B) The scalability of GATK’s HaplotypeCaller pipeline on x86 and Power9 was tested with varying amounts of compute resources alongside Sentieon’s DNAseq and Parabricks. (C) Scalability of GATK’s Mutect2 pipeline on x86 and Power9 was tested with varying amounts of compute resources alongside Sentieon’s TNseq and TNScope. At the time of this analysis, Parabricks had not yet ported their somatic pipeline to Power9 therefore it could not be tested (Continued to the next page).
Variant calling pipelines' scalability. (A) To determine the most efficient way of parallelizing GATK’s HaplotypeCaller, different combinations of scattered intervals and PairHMM OpenMP threads were tested on x86 and Power9 systems. The recorded times also include the CombineGVCFs step. (B) The scalability of GATK’s HaplotypeCaller pipeline on x86 and Power9 was tested with varying amounts of compute resources alongside Sentieon’s DNAseq and Parabricks. (C) Scalability of GATK’s Mutect2 pipeline on x86 and Power9 was tested with varying amounts of compute resources alongside Sentieon’s TNseq and TNscope. At the time of this analysis, Parabricks had not yet ported their somatic pipeline to Power9 therefore it could not be tested.

https://doi.org/10.5808/GI.2020.18.1.e10
pipelines via GATK’s Concordance tool (Table 2). TNseq had a higher amount of variation when compared to Mutect2 v4.1.0 than DNaseq had to HaplotypeCaller v4.1.0; however, when TNseq was compared against the specific version of Mutect2 it was designed to mimic (v4.0.2.1), the results were more similar with accuracy and sensitivity both over 99.5% (data not shown).

Benchmarking other bioinformatic tools: POWER9 for bioinformatics

To assess how well POWER9 handles bioinformatic workloads the performance of ten commonly used alignment/mapping tools were compared when run on our Intel x86 nodes and the IBM POWER9 node: STAR, BBMap, Tophat, HISAT2, BWA MEM, Bowtie, Bowtie2, blastx, blastn, and pblat a parallelized version of BLAT. Each tool was run using default settings with the same workload on both x86 and POWER9. This is not a comparison of the tools themselves; some similar tools were given different workloads, such as STAR and Tophat, to allow for meaningful runtimes. Each tool was run with the following resource configurations to allow for thread-vs-thread and core-vs-core comparisons: x86-16t/8c, x86-32t/8c, POWER9-16t/4c, and POWER9-32t/8c.

For thread-vs-thread comparisons, the results are somewhat split between similar runtimes and a slight edge to x86 threads (Fig. 2). STAR, HISAT2, and Bowtie, achieved nearly identical runtimes on both platforms. BBMap, Tophat, BWA MEM, Bowtie2, and blastx performed better with an equal number of x86 threads, while blastn ran nearly twice as fast with an equal number of POWER9 threads and pBLAT had a slight edge. In the core-vs-core performance comparison, the POWER9 architecture resulted in lower runtimes in every instance.

### Table 2. Variant calling pipelines’ accuracy

<table>
<thead>
<tr>
<th></th>
<th>Germline</th>
<th>Somatic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GATK</td>
<td>Sentieon</td>
</tr>
<tr>
<td>VS Baseline VCF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True-positive</td>
<td>-</td>
<td>3,827,008</td>
</tr>
<tr>
<td>False-positive</td>
<td>-</td>
<td>9,703</td>
</tr>
<tr>
<td>False-negative</td>
<td>-</td>
<td>11,202</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>-</td>
<td>0.99708</td>
</tr>
<tr>
<td>Precision</td>
<td>-</td>
<td>0.99747</td>
</tr>
<tr>
<td>INDEL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True-positive</td>
<td>-</td>
<td>815,205</td>
</tr>
<tr>
<td>False-positive</td>
<td>-</td>
<td>11,314</td>
</tr>
<tr>
<td>False-negative</td>
<td>-</td>
<td>10,756</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>-</td>
<td>0.98698</td>
</tr>
<tr>
<td>Precision</td>
<td>-</td>
<td>0.98631</td>
</tr>
<tr>
<td>VS Truthset VCF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False-positive</td>
<td>2,345</td>
<td>2,344</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.96968</td>
<td>0.96976</td>
</tr>
<tr>
<td>Precision</td>
<td>0.99933</td>
<td>0.99933</td>
</tr>
<tr>
<td>INDEL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True-positive</td>
<td>548,276</td>
<td>548,574</td>
</tr>
<tr>
<td>False-positive</td>
<td>9,496</td>
<td>8,987</td>
</tr>
<tr>
<td>False-negative</td>
<td>24,451</td>
<td>24,153</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.95731</td>
<td>0.95783</td>
</tr>
<tr>
<td>Precision</td>
<td>0.98298</td>
<td>0.98388</td>
</tr>
</tbody>
</table>

GATK, Genome Analysis Toolkit; VCF, Variant Call Format; SNP, single nucleotide polymorphism.

https://doi.org/10.5808/GI.2020.18.1.e10
Fig. 2. x86/POWER9 performance comparison of aligners/mappers. The performance of 10 different tools for alignment/mapping was compared between POWER9 and x86 systems. Jobs were run in triplicate across different days; averaged results are shown in the graphs with the error bars representing standard deviation.
Discussion

Both Parabricks and Sentieon offer highly accelerated GATK algorithms that greatly reduce variant calling processing time. Parabricks’s utilization of the higher number of cores available on a GPU compared to a CPU allows for a much faster processing time of a single sample compared to Sentieon. However, the legacy of algorithm development in a CPU environment compared to a GPU for NGS pipelines has impacted the relative number of CPUs available compared to GPUs within an institution. The value of being able to process a single sample in ~2 h should not be discounted, however, especially in the age of precision medicine. Delivering “fast” genome sequencing to help diagnosis a newborn in a neonatal intensive care unit or to identify a somatic variant that can be used in treatment selections is becoming a reality for many institutions.

Our testing of Sentieon’s ability to reproduce GATK’s output yielded similar results to a previous assessment done by Kendig et al. [18]; however, the output from Parabricks was essentially identical to the targeted version which may make it more attractive to those within the research community. If GATK compliance is less of a concern, Sentieon does offer their own proprietary algorithms for variant calling which did perform better when compared to the VCF truthsets used in this study.

While the CPUs used in this comparison are all based on 14nm microarchitectures, the Intel CPUs in our x86 nodes are a few years old, and to do a true x86 versus POWER9 comparison a newer Intel CPU would be needed. Despite this, we felt that these benchmarking comparisons would shed light on POWER9’s ability to handle bioinformatic workloads and prove useful to researchers in designing their compute infrastructure.

The POWER9 architecture performed well in the core-vs-core comparisons. Part of this advantage comes from IBM’s implementation of simultaneous multithreading, which is superior to Intel’s Hyper-threading by allowing for twice as many threads to run simultaneously on the same core. This analysis demonstrated that even in instances where the tools were originally coded to take advantage of Intel specific instruction sets such as SSE2, the total runtime was lower on POWER9 than x86 with identical core counts. However, this does allude to an issue that can arise when using a POWER9 system: not all bioinformatic programs that were originally written for x86 systems can easily compile and run on the POWER9 architecture.

Acknowledgments

We thank both Parabricks and Sentieon for trial licenses of their software suites. We also thank Ruzhu Chen from IBM for POWER9 versions of BWA, Bowtie2, and Bowtie.

This work was supported by the Lisa Dean Moseley Foundation (PI Crowgey); and the Nemours Alfred I duPont Hospital for Children.

References


Instructions for authors

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

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

Research and publication ethics

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

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

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

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

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

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

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

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

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

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

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

Preparation of manuscripts

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

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

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

**Clinical genomics**

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

**Genome archives**

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

**Letters to the editor**

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

**Review articles**

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

**Opinions / Commentaries**

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

**Minireviews**

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

**Research communications**

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

**Manuscript Format**

**Title**

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

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

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

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

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

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

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

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

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

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

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

Examples of references are given below:

Journal article

Books

Book sections

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

Conference paper

Dissertation/Thesis

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

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

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

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

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

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

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

- Quick Time files (.mov)
Selection and Description of Participants

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

Authors should define how they determined race or ethnicity and justify their relevance.

Submission of Manuscript

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

Peer review and revision of manuscripts

Peer review

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

Manuscript revision

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

Copyrights, open access policy and open data policy

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

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

Archiving policy
It is accessible without barrier from Korea Citation Index (https://kci.go.kr), National Library of Korea (http://nl.go.kr), or PubMed Central (https://www.ncbi.nlm.nih.gov/pmc/journals/1928/) in the event a journal is no longer published.

Deposit policy (Self-archiving policy) according to Sherpa/Romeo (http://www.sherpa.ac.uk/): Author can not archive pre-print (i.e., pre-refereeing). Author can archive post-print (i.e., final draft post-refereeing).

Author can archive publisher’s version/PDF.

Open data policy
Data sharing is recommended. If the data are already public, the URL site or sources should be disclosed. If data can not be publicized, it can be negotiated with the editor. If there are any inquiries on depositing data, author’s should contact the editorial office.

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

Detailed Description of Use of Articles of Genomics & Informatics

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

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

Copyrights
Publisher applies the Creative Commons Attribution Non-Commercial license to works it publishes and allows free immediate access to, and unrestricted reuse of, original works of all types.
Table 1. Examples of data sharing statements that fulfill ICMJE requirements

<table>
<thead>
<tr>
<th>Element</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
<th>Example 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will individual participant data be available (including data dictionaries)?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>What data in particular will be shared?</td>
<td>All of the individual participant data collected during the trial, after deidentification.</td>
<td>Individual participant data that underlie the results reported in this article, after deidentification (text, tables, figures, and appendices).</td>
<td>Individual participant data that underlie the results reported in this article, after deidentification (text, tables, figures, and appendices).</td>
<td>Not available</td>
</tr>
<tr>
<td>What other documents will be available?</td>
<td>Study protocol, statistical analysis plan, informed consent form, clinical study report, analytic code</td>
<td>Study protocol, statistical analysis plan, analytic code</td>
<td>Study protocol</td>
<td>Not available</td>
</tr>
<tr>
<td>When will data be available (start and end dates)?</td>
<td>Immediately following publication. No end date.</td>
<td>Beginning 3 months and ending 5 years following article publication.</td>
<td>Beginning 9 months and ending 36 months following article publication.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>With whom?</td>
<td>Anyone who wishes to access the data.</td>
<td>Researchers who provide a methodologically sound proposal.</td>
<td>Investigators whose proposed use of the data has been approved by an independent review committee (&quot;learned intermediary&quot;) identified for this purpose.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>For what types of analyses?</td>
<td>Any purpose</td>
<td>To achieve aims in the approved proposal.</td>
<td>For individual participant data meta-analysis.</td>
<td>Not applicable</td>
</tr>
<tr>
<td>By what mechanism will data be made available?</td>
<td>Data are available indefinitely at (link to be included).</td>
<td>Proposals should be directed to xxx@yyy. To gain access, data requestors will need to sign a data access agreement.</td>
<td>Proposals may be submitted up to 36 months following article publication. After 36 months, the data will be available in our University's data warehouse but without investigator support other than deposited metadata. Information regarding submitting proposals and accessing data may be found at (link to be provided).</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Data are available for 5 years at a third-party website (link to be included).</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

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

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

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

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

**Contact address**

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

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

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

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

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

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

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

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

Statement of Informed Consent

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

Statement of Human and Animal Rights

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

Authorship

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

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

Corresponding author and first author

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

Correction of authorship after publication

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

Conflict of Interest Statement

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

Originality and Duplicate Publication

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

Secondary Publication

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

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

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


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

☐ 3. Abstract in unstructured format within 250 words.

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

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

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

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

☐ 8. Figures as 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.