G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2018;16(4):e16
https://doi.org/10.5808/GI.2018.16.4.e16

EDITORIAL

Editor's Introduction to This Issue (G&I 16:4, 2018)

Taesung Park*

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

In this issue, there are 24 articles: five Review Articles, 13 Original Articles, one Clinical Genomics, four Application Notes, and one Opinion article.

The first review is about single-cell sequencing in cancer. Single-cell sequencing technologies already have led to many discoveries in the field of cancer immunology and proven it can provide further insight into understanding of intratumor heterogeneity, immunoediting, and immunogenicity that are difficult to resolve from bulk sequencing. The comprehensive review by Sierant and Choi (Yale University, New Haven, CT, USA) presents reviews on the perspectives of single-cell genomic analysis in cancer immunology, the potential pitfalls of different single-cell analysis technology, and the therapeutic opportunity for cancer treatment as the hallmarks of precision medicine in the future.

The second review article of Do and Kim (Korea National University of Education) discussed the roles long non-coding RNAs (lncRNAs) in biological processes and describe the lncRNA-mediated signaling pathways involved in cancers, with a focus on oncogenic lncRNAs. Oncogenic lncRNAs have the potential to become promising biomarkers and might be potent prognostic targets in cancer therapy. This review summarizes studies on the regulatory and functional roles of oncogenic lncRNAs in the development and progression of various types of cancer.

The third review by Seunghee Cha (University of Florida, Gainesville, FL, USA) and colleagues provide an excellent review on microRNAs (miRNAs) in Autoimmune Sjögren's Syndrome (SjS). This review compiles and highlights differentially-expressed miRNAs in various samples collected from SjS patients and their potential implications in the pathogenesis of SjS. This review suggests potential clinical implications of miRNAs to select disease-specific diagnostic and prognostic biomarkers by utilizing different types of tissues or biological specimens of SjS patients.

The fourth review is about genetic hearing loss and gene therapy. Carpena and Lee (Dankook University, Cheonan, Korea) consolidated the genes that are currently identified to be associated with hearing loss. They provided reviews on the recent advances in elucidating the genomics of genetic hearing loss and technologies aimed at developing a gene therapy that may become a treatment option for in the near future.

The final review by Jeon and Galvao (Long Island University, Brookville, NY, USA) focused on some recent studies based on high-throughput sequencing claiming that metritis is a microbiota-associated disease. They provided the reviews some recent studies that used metagenomic approaches to explore uterine bacterial community involved in metritis and discuss uterine bacteria that are known and newly identified in metagenomic sequencing. They proposed that metritis is associated with uterine microbiota with high abundance of *Bacteroides*, *Porphyromonas*, and *Fusobacterium*.

This issue contains 13 Original Articles articles. First, there are three articles on metagenome. With the development of genome sequence analysis technology, there is an increasing demand for virus taxonomy to be extended from *in vivo* and *in vitro* to *in silico*. Kang and Kim (Chungbuk National University, Cheongju, Korea) verified the consistency of the current International Classification of Viruses (ICTV) taxonomy using an *in silico* approach, aiming to identify the specific sequence for each virus. They applied this approach to *norovirus* in *Caliciviridae*, which causes 90% of gastro-enteritis cases worldwide. In their second article, Kang and Kim constructed a web-based system and extension of an associated database, based on ICTV taxonomy. Their virus taxonomy web system was specifically designed to extend the virus taxonomy up to strain and isolation, which was then connected with the NCBI database to facilitate searching for specific viral genes; there are also links to journals provided by the EMBL RESTful API that improve accessibility for academic groups. Thus, the web-based virus taxonomy could be augmented the quality by adding virus classification, which is derived by virus metagenomics

***Corresponding author:** Tel: +82-2-880-8924, Fax: +82-2-883-6144, E-mail: tspark@stats.snu.ac.kr

analysis.

Taxonomy identification is fundamental to in metagenomics. Park and Won (Institute of Health and Environment, Seoul National University, Seoul, Korea) evaluated the accuracy of three widely used 16S databases, Greengenes, Silva, and EzBioCloud, and suggested basic guidelines for selecting reference databases. Using public mock community data, they showed that EzBioCloud performs well compared to other existing databases.

In this issue, there is one on transcriptomics. Niloofar Avazpour (Shahid Chamran University of Ahvaz, Ahvaz, Iran) and colleagues showed that *HOTAIR*, one of the most cited lncRNAs with a critical role in initiation and progression of the gene expression regulation, can be a potential biomarkers for coronary artery disease (CAD). They compared the expression of *HOTAIR* lncRNA in the blood samples of patients with CAD and control samples. The expression level was examined by semi-quantitative reverse transcriptase polymerase chain reaction technique. They showed that expression of *HOTAIR* is up-regulated in blood samples of patients with CAD.

Md. Amran Gazi (International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh) and his colleagues extensively analyzed the genome of *Shigella flexneri* and found 4,362 proteins among which the functions of 674 proteins, termed as hypothetical proteins (HPs) had not been previously elucidated. These HPs were found to belong to various classes of proteins such as enzymes, binding proteins, signal transducers, lipoprotein, transporters, virulence, and other proteins. Their comprehensive analysis will help to gain greater understanding for the development of many novel potential therapeutic interventions to defeat *Shigella* infection.

In this issue, there are also several articles on Informatics. Hyun-Seok Park (Ewha Womans University, Korea) investigated the effect of the primary DNA sequence on the epigenomic landscape across a 200-base pair of genomic units by integrating 127 publicly available ChromHMM BED files from the Roadmap Genomics project. He analyzed nucleotide frequency profiles of 127 chromatin annotations stratified by chromatin variability and built integrative hidden Markov models to detect Markov properties of chromatin regions.

Jonghwan Yoon (The Catholic University of Korea, Seoul, Korea) and his colleagues studied the tRNA adaptation index which is a measure of translational efficiency for a gene and is calculated based on the abundance of intracellular tRNA and the binding strength between a codon and a tRNA. The index has been widely used in various fields of molecular evolution, genetics, and pharmacology. For an improved version of the index, named specific tRNA adaptation index (stAI), they calculated stAI values for whole coding sequences in 148 species. In addition, they constructed a novel web database, STADIUM (Species-specific tRNA adaptive index compendium) to enable easy access to improved version of this index.

Young-Joon Kim's group (Yonsei University, Seoul, Korea) built a hybrid capture-based targeted sequencing analysis pipeline for global application to many studies of targeted sequencing by using genomic DNA extracted from tumor tissues of colorectal cancer patients. This pipeline was specialized for unique molecular index data. Through this pipeline, they were able to estimate the even on-target rates and filtered consensus reads for more accurate variant calling. These results suggest the potential of our analysis pipeline in the precise examination of quality and efficiency of conducted experiments.

In response to the urgent need for effective and cost-efficient data, DNA synthesis and sequencing has been used as a new tool for storing digital information. Most studies have focused on making use of 100–150 bp of short read sizse in both synthesis and sequencing. TaeJin Ahn's group (Handong University, Pohang, Korea) suggested novel data encoding / decoding scheme which makes use of long read DNA (~1,000 bp). Longer DNA read can store larger amount of digital information within a single molecule. Thus, their approach is more scalable than short DNA methods and there is no need to wait for denovo synthesis of DNA.

In this issue, there are several articles on data mining and statistical application to omics data.

The first article is about metabolic syndrome (MS) in the nonobese population. While the prevalence of MS in the nonobese population is not low, the identification and risk mitigation of MS are not easy. Seung Ho Cho's group (Seoul National University Hospital) developed an MS prediction model using genetic and clinical factors of nonobese Koreans with five machine learning algorithms, including naïve Bayes classification (NB). The analysis was performed in two stages (training and test sets) with evaluation measures such as sensitivity, specificity, area under the receiver operating characteristic curve. While the performance of NB was best, its prediction performance was not as good as expected.

The second article is about building prognostic prediction model of ovarian cancer. Seokho Jeong (Seoul National University, Seoul, Korea) and his colleagues presented an efficient strategy to build a prognostic prediction model of ovarian cancer by integrating the high dimensional RNA sequencing data with their clinical data through the following steps: (1) gene filtration, (2) pre-screening, (3) gene marker selection, and (4) integration of selected gene

markers and prediction model building. This strategy is so general that it can be applied to other types of cancer besides ovarian cancer.

Finally, the third article is about multi-block analysis of genomic data. In general, multi-block data is used for enhancing analysis of different block's relationship. Mira Park's group (Eulji University, Daejeon, Korea) performed multi-block analysis of genomic data. Using generalized canonical correlation analysis, they could identify relationships between SNP block, phenotype block, and disease block data from Korean Association Resource (KARE) project.

There is one interesting article on a Korean medicinal herb. Neha Samir Roy (Kangwon National University, Chuncheon, Korea) and colleagues reported the first *de novo* assembly of the transcriptome of *Cirsium japonicum* var. *spinossimum* obtained from the Korean Peninsula. They identified the expression of genes related to the synthesis of silymarin in *C. japonicum* in three different tissues, flowers, leaves and roots, through RNA sequencing. Their study provided resources for comparative transcriptomics, especially in the field of the biochemical and molecular biosynthesis pathways of silymarin.

In Clinical Genomics section, there is one article. Kyong-Ah Yoon (Konkuk University, Seoul, Korea) and her colleagues reported the genetic characteristics of biliary tract cancer (BTC). They analyzed whole exome sequencing data and identified somatic mutations from the seven BTC patients with cholangiocarcinoma who were underwent surgical resection. They discovered inactivating mutations of tumor suppressor genes including *APC*, *TP53*, and *ARID1A* genes in three patients. Activating mutations of *KRAS* and *NRAS* were also identified.

In the Application Note section, four programs are introduced. For the integrative analysis of the RNA sequencing (RNA-Seq) data, Sang Cheol Kim (National Institute of Health, Cheongju, Korea) and his colleagues developed a web-based application using Shiny, COEX-seq (Convert a Variety of Measurements of Gene Expression in RNA-Seq) that easily converts data in a variety of measurement formats of gene expression used in most bioinformatics tools for RNA-Seq. It provides a workflow that includes loading data set, selecting measurement formats of gene expression, and identifying gene names. COEX-seq is freely available for academic purposes and can be run on Windows, Mac OS, and Linux operating systems. Source code, sample data sets, and supplementary documentation are available as well.

It is well known that gene-gene interaction is a key factor to explain the missing heritability in genome-wide association studies (GWAS). Many methods have been proposed to identify gene-gene interactions. Multi-factor dimensionality reduction (MDR) is a well-known method for gene-gene interaction detection by reduction from genotypes of SNP combination to a binary variable with a value of high risk or low risk. Sangseob Leem (Seoul National University, Seoul, Korea) and his colleague proposed empirical fuzzy MDR (EFMDR) by combining maximum likelihood estimation and fuzzy set theory into MDR and implemented EFMDR using RCPP (c++ package) for faster executions.

Sungkyoung Choi (Yonsei University College of Medicine, Seoul, Korea) also developed the hierarchical structural component model for gene-gene interaction analysis for a continuous phenotype. The software HisCoM-GGI is for performing this gene-level and SNP-level interaction analysis. HisCoM-GGI handles various type of genomic data, and supports a data management and multithreading to improve the efficiency of GWAS data analysis. HisCoM-GGI is expected to provide advanced accessibility to the researchers on the genetic interaction studies and a more effective way to understand biological mechanisms of complex diseases.

Sungyoung Lee (Seoul National University Hospital, Seoul, Korea) and his colleagues proposed a novel statistical method that enables the pathway-based association study of large-scale genetic dataset. The software PHARAOH is for performing this pathway analysis. PHARAOH program not only supports various *de facto* standard genetic data formats but also provides many quality control measures and filters based on those measures. PHARAOH can provide advanced accessibility of the pathway-level analysis of large-scale genetic datasets to researchers.

Finally, in the new Opinion section, Hyun-Seok Park (Ewha Womans University, Seoul, Korea) presented his opinion on Strategy of Semi-Automatically Annotating Full Text Corpus of *Genomics & Informatics* (*GNI*). He listed issues associated with upgrading annotations, and gave opinion on methodology to develop next version of GNI corpus based on a semi-automatic strategy for more linguistically rich corpus annotation.