G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2015;13(3):86-89
http://dx.doi.org/10.5808/GI.2015.13.3.86

APPLICATION NOTE

A Database of Gene Expression Profiles of Korean Cancer Genome

Seon-Kyu Kim[1,2,3], In-Sun Chu[2,3]*

[1]Medical Genomics Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, Korea,
[2]Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, Korea,
[3]Department of Bioinformatics, Korea University of Science and Technology, Daejeon 34131, Korea

Because there are clear molecular differences entailing different treatment effectiveness between Korean and non-Korean cancer patients, identifying distinct molecular characteristics of Korean cancers is profoundly important. Here, we report a web-based data repository, namely Korean Cancer Genome Database (KCGD), for searching gene signatures associated with Korean cancer patients. Currently, a total of 1,403 cancer genomics data were collected, processed and stored in our repository, an ever-growing database. We incorporated most widely used statistical survival analysis methods including the Cox proportional hazard model, log-rank test and Kaplan-Meier plot to provide instant significance estimation for searched molecules. As an initial repository with the aim of Korean-specific marker detection, KCGD would be a promising web application for users without bioinformatics expertise to identify significant factors associated with cancer in Korean.

Keywords: biological markers, database, genomics, Korean, neoplasms, prognosis

Availability: KCGD is freely available at http://www.kcgd.kr.

Introduction

Cancer is a complex disease with heterogeneous clinical behaviors developed by accumulation of multiple genetic or epigenetic alterations. Several global research consortia have made great efforts to improve the understanding of cancer biology and the development of more effective cancer treatments [1, 2]. Most treatment options for Korean cancer patients were established based on such western population investigations. However, there are clear molecular differences showing different treatment effectiveness between Korean and non-Korean cancer patients [3, 4]. Thus, it is very important to identify distinct molecular characteristics of Korean cancer patients.

Currently, numerous databases and analysis toolkits supporting cancer genomics studies have been reported [5-10]. These studies mostly support a database system for searching disease-associated genes or target drugs. Although many researchers have tried to develop platforms to find molecular markers from genomics data, there are few suitable web-based resources that help researchers develop gene signatures associated with Korean cancer patients. Collecting Korean cancer genomics data, comparing with other data obtained from non-Korean and estimating prognostic or predictive value of the genes or gene sets using proper statistical analyses may be a daunting task for many investigators, particularly clinicians and oncologists.

Here, we introduce a web-based initial repository, namely Korean Cancer Genome Database (KCGD), to help investigators in the efforts for searching prognostic signatures in Korean cancer patients. The database contains the gene expression profile with clinical data obtained from more than 1,000 Korean cancer patients. It is designed to be simple to search significant molecules, for which it is available for instant statistical survival analyses. In addition, our database has gradually containing non-Korean datasets so that users can easily compare or validate newly identified molecules independently.

Received July 30, 2015; Revised August 17, 2015; Accepted August 18, 2015

***Corresponding author:** Tel: +82-42-879-8520, Fax: +82-42-879-8519, E-mail: chu@kribb.re.kr

Methods

Collection of datasets

We have collected and generated genomics data obtained from 1,304 Korean cancer patients collaborating with a number of hospitals in Korea, such as Seoul National University Hospital (liver), Chungbuk National University Hospital (bladder), Chonbuk National University Hospital (liver, bile duct, and colon), Yonsei University Severance Hospital (breast), Korea University Medical Center (stomach), and Kosin University Gospel Hospital (stomach). The data were partially deposited to and freely available from the Gene Expression Omnibus of National Center for Biotechnology Information. All datasets stored in the database were normalized using quantile normalization. Detailed repository status was illustrated in Supplementary Table 1.

Implementations

The system architecture consists of various software frameworks for robust activity. Our system was mainly implemented with JAVA-based environment. To provide user friendly and active interfaces, the ICEfaces (version 3.3.0, http://www.icesoft.org/) framework was used. To store and handle the datasets, the MySQL database management system was used (version 5.5.11, http://dev.mysql.com). Data queries on MySQL from JAVA are controlled by MyBatis, an XML-based SQL mapping framework (version 3.1.1, https://code.google.com/p/mybatis). All statistical analysis methods were implemented using R (version 3.0.1, http://www.r-project.org) with Bioconductor plugins (version 2.12, http://www.bioconductor.org). Calling R modules from JAVA is managed by the RCaller framework (version 2.1.1, https://code.google.com/p/rcaller). All services are hosted on an Apache Tomcat web server (version 6.0.26, http://tomcat.apache.org). A schematic diagram of the system architecture is shown in Supplementary Fig. 1.

Supported analysis methods

Our system currently contains statistical survival analysis methods for identifying a signature associated with cancer outcome and estimating its predictive value, i.e., the Cox proportional hazard model, log-rank test and Kaplan-Meier curves. Detailed methodologies are available in the previous methodology paper [11]. In addition, the system determines the significance of a molecule using bar plots illustrating the landscape of intensities among patients and two or more group box plots with a p-value obtained from a two sample t test or ANOVA methods.

Results

Web-based bioinformatics tool

A platform to simply search and estimate statistical significance of molecules across various cancer types accessible to investigators without bioinformatics or statistics expertise are available at the KCGD website (Fig. 1). It supports dataset search, in which a user explores cancer genomics or epigenomics datasets in Korean stored in the database, and gene search, which determines statistical difference between cancer known subtypes or prognostic

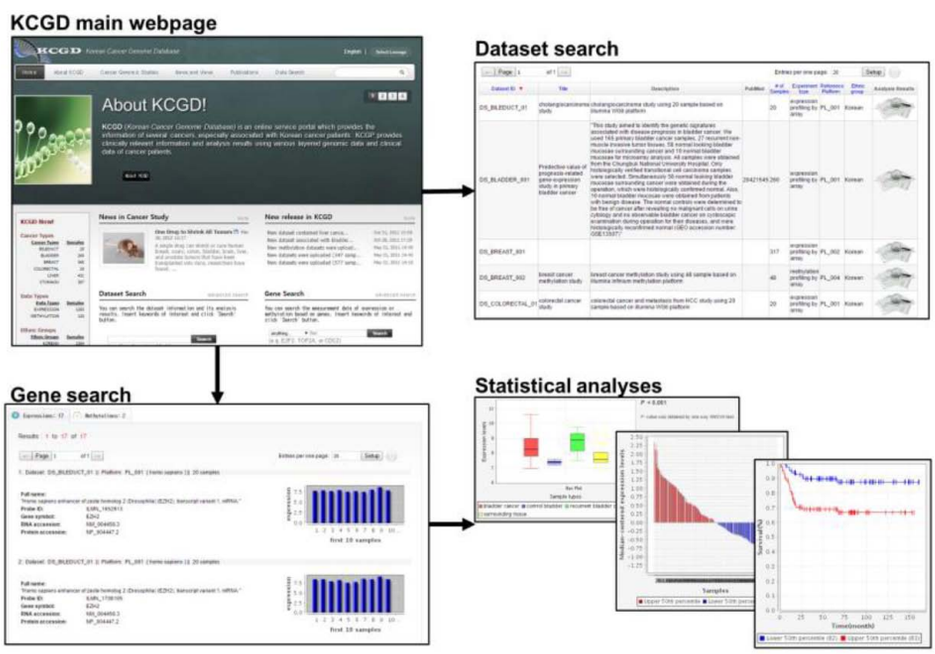

Fig. 1. Snapshot of the Korean Cancer Genome Database (KCGD) webpage. The system supports dataset search, in which a user explores cancer genomics or epigenomics datasets in Korean stored in the database (right upper panel), and gene search, which determines statistical difference between cancer known subtypes or prognostic value of a molecule (right lower panel).

value of a molecule in which users are interested.

When accessing the gene search module, a number of accession identifiers including gene symbol, RefSeq (NM_* or XM_*), Entrez and protein accession number (NP_* or XP_*) can be selected, then the user input keyword(s) for search and click the search button. Although the search criteria were arranged using small number of identifiers, most known molecule identifiers were incorporated in the database by full text indexing method, indicating any molecule identifiers were allowed to search by clicking "anything…" category. Clicking the search button, the measurement data of expression or methylation are sought by the keyword and matched list with molecule information and thumb nail image of measurement intensities across cancers are displayed. Next, by clicking a thumb nail image, the system try to estimate the significance of a molecule using the log-rank test with Kaplan-Meier curves and display bar plots to illustrate the landscape of intensities among patients and box plots with a p-value obtained from a two sample t test or ANOVA test (Fig. 1). With p-values by log-rank test or significance test with t test or ANOVA, the prognostic value of a molecule or its difference between cancer subtypes may be determined, respectively. The entire analysis procedures of the system were described in Supplementary Fig. 2.

Statistics of datasets

Currently, we have created a database containing a total of 1,403 patient samples, among which 1,304 samples are Korean and the remained 99 samples are non-Korean patients for comparison. Most data in the current database were created by gene expression profiling (n = 1,283, 91%) and the remained (n = 120, 9%) were by methylation profiling methods. The vast majority of cancer tissues (n = 1,363, 97%) consists of liver (n = 431, 31%), breast (n = 365, 26%), stomach (n = 307, 22%), and bladder (n = 260, 19%), among which highly ranked three cancer types (i.e., liver, breast, and stomach) were known to be most frequently occurred in Korean. In each cancer type, we also designated previously known sample subtypes to support assessment of difference of a molecule between cancer subtypes. In addition, survival data of liver (n = 99) and bladder cancer patients (n = 165) were involved in the database to instantly estimate the prognostic value of a molecule in which users are interested (n = 264, 19%). Supplementary Table 1 elucidated detailed statistics in KCGD database.

Discussion

We have constructed the KCGD system, a web-based search platform, to assist investigators in estimating predictive value of interested genes and in identifying molecular signatures from Korean cancer genomics data. As an initial repository with the aim of Korean-specific marker detection, KCGD always open to collaborate with any research investigators or clinicians in Korea and is ready to share Korean cancer genomics data with clinical information to solve the critical problems in the biomedical fields. KCGD is a growing database: although current database mainly contains gene expression with clinical data, the database is ready to handle the RNA-seq data, another gene expression profiling method, or any continuous numeric intensity data (i.e., methylation or genomic variation). In addition, for direct comparative genomic analysis with non-Korean data, we are trying to collect other genomic or epigenomic data from several public database including the Cancer Genome Atlas consortium and comparative browsing would be implemented in near future. Therefore, we suggest that KCGD may be one of the best choice as a co-work partner when users try to discover significant novel factors associated with genomic studies of Korean cancer patients.

Supplementary materials

Supplementary data including one table and two figures can be found with this article online at http://www.genominfo.org/src/sm/gni-13-86-s001.pdf.

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) grant (2011-0019745) funded by the Korea government (MEST) and a grant from the KRIBB Research Initiative Program.

References

1. Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium. Pathway and network analysis of cancer genomes. *Nat Methods* 2015;12:615-621.
2. Cancer Genome Atlas Research Network, Weinstein JN, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 2013;45:1113-1120.
3. Lee SY, Kim DW, Shin YK, Ihn MH, Lee SM, Oh HK, *et al*. Validation of prediction models for mismatch repair gene mutations in Koreans. *Cancer Res Treat* 2015 Jun 5 [Epub]. http://dx.doi.org/10.4143/crt.2014.288.
4. Mok Y, Son DK, Yun YD, Jee SH, Samet JM. Gamma-glutamyltransferase and cancer risk: the Korean cancer prevention study. *Int J Cancer* 2015 Jun 25 [Epub]. http://dx.doi.org/10.1002/ijc.29659.

5. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, *et al*. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;6:pl1.
6. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, *et al*. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res* 2012;72:3499-3511.
7. Madhavan S, Gusev Y, Harris M, Tanenbaum DM, Gauba R, Bhuvaneshwar K, *et al*. G-DOC: a systems medicine platform for personalized oncology. *Neoplasia* 2011;13:771-783.
8. Feichtinger J, McFarlane RJ, Larcombe LD. CancerMA: a web-based tool for automatic meta-analysis of public cancer microarray data. *Database* (Oxford) 2012;2012:bas055.
9. Kurian AW, Munoz DF, Rust P, Schackmann EA, Smith M, Clarke L, *et al*. Online tool to guide decisions for BRCA1/2 mutation carriers. *J Clin Oncol* 2012;30:497-506.
10. Kim SK, Kim JH, Yun SJ, Kim WJ, Kim SY. APPEX: analysis platform for the identification of prognostic gene expression signatures in cancer. *Bioinformatics* 2014;30:3284-3286.
11. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Methodol* 1972;34:187-220.