BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 22 2014, pages 3284–3286
doi:10.1093/bioinformatics/btu521

Gene expression

Advance Access publication August 4, 2014

APPEX: analysis platform for the identification of prognostic gene expression signatures in cancer

Seon-Kyu Kim[1,2], Jong Hwan Kim[1,3], Seok-Joong Yun[4], Wun-Jae Kim[4] and Seon-Young Kim[1,3,*]

[1]Medical Genomics Research Center, [2]Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, [3]Department of Functional Genomics, University of Science and Technology, Daejeon 305-806, Korea and [4]Department of Urology, Chungbuk National University College of Medicine, Cheongju 360-100, Korea

Associate Editor: Janet Kelso

ABSTRACT

Summary: Because cancer has heterogeneous clinical behaviors due to the progressive accumulation of multiple genetic and epigenetic alterations, the identification of robust molecular signatures for predicting cancer outcome is profoundly important. Here, we introduce the APPEX Web-based analysis platform as a versatile tool for identifying prognostic molecular signatures that predict cancer diversity. We incorporated most of statistical methods for survival analysis and implemented seven survival analysis workflows, including *CoxSingle*, *CoxMulti*, *IntransSingle*, *IntransMulti*, *SuperPC*, *TimeRoc* and *multivariate*. A total of 236 publicly available datasets were collected, processed and stored to support easy independent validation of prognostic signatures. Two case studies including disease recurrence and bladder cancer progression were described using different combinations of the seven workflows.

Availability and implementation: APPEX is freely available at http://www.appex.kr.

Contact: kimsy@kribb.re.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 20, 2014; revised on June 29, 2014; accepted on July 24, 2014

1 INTRODUCTION

The identification of robust molecular signatures that predict cancer patient outcome is profoundly important because cancers have heterogeneous clinical courses even if they have similar clinicopathological characteristics. By using prognostic molecular signatures, cancer patients may be treated more effectively. For example, the Onco*type* DX breast cancer assay is now used in the clinic to predict the clinical behavior of breast cancer patients (Paik *et al.*, 2004).

Various software tools supporting cancer genomics studies have been reported. Most of these studies provide a database platform for searching disease-associated genes and target drugs (Aguirre-Gamboa *et al.*, 2013; Gao *et al.*, 2013; Madden *et al.*, 2013; Reinhold *et al.*, 2012; Ringner *et al.*, 2011), whereas only one system provides a software environment for handling a user's own cancer genomics data with clinical information to determine a significant prognostic signature (Corradi *et al.*, 2009). Beyond software toolkits for cancer research, many investigators use commercial programs or script language, such as SPSS, Matlab or R, for advanced statistical analysis. However, there are few suitable Web-based analysis tools that help researchers develop gene signatures. For many oncology investigators, doing proper statistical analyses using publicly available tools can be a daunting task. In addition, most genome-wide analysis tools are not equipped with tools for identifying prognostic signatures by survival analysis.

Here, we constructed APPEX, a Web-based software platform to help researchers in the efforts to identify prognostic signatures from genomics data. APPEX is designed to be easy to use and flexible, and it is freely available for advanced statistical survival analyses. A user-friendly graphical interface similar to a desktop application is provided so that users can easily handle their own data on APPEX even if they are not familiar with statistical analysis packages. In addition, APPEX contains >200 publicly available datasets directly applicable on the system so that users can easily validate newly identified signatures in independent patient cohorts.

2 METHODS

2.1 Analysis methods incorporated in APPEX

The APPEX system currently contains four independent statistical approaches for identifying and estimating a signature associated with cancer outcome, i.e. the Cox proportional hazard model (Cox, 1972), an *in-trans* correlation approach (Lee *et al.*, 2010), SuperPC (Bair and Tibshirani, 2004) and time-dependent ROC curves (Heagerty *et al.*, 2000). Details for their methodologies are available in the Supplementary Methods.

2.2 Collection and processing of public datasets

We have collected 236 public datasets containing continuous numeric intensity data (i.e. mRNA expression, methylation, genomic variation and non-coding RNA profiles) and patient follow-up information from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) and incorporated them into the APPEX. The public datasets were normalized using quantile normalization. To handle data generated from heterogeneous experimental platforms, we considered a combination of unique probe IDs and gene symbols as gene-identifying criteria. Therefore, APPEX processes intensity values with probe IDs and determines gene annotations with gene symbols.

*To whom correspondence should be addressed.

2.3 Implementation and system architecture

Basically, APPEX was implemented with JAVA. To provide user-friendly and active interfaces, the Google Web toolkit (GWT, ver. 2.5.1) and GWT extended (GXT, ver. 3.0.1) frameworks were used. Data exchange between clients and the APPEX server is controlled by a GWT remote procedure call. All statistical analysis methods in APPEX were implemented using R (ver. 3.0.1) with Bioconductor plugins (ver. 2.12). To handle multiple time-consuming jobs concurrently, the Quartz framework was used (ver. 2.1.6). To handle public datasets from the NCBI GEO and user analysis history, the MySQL database server was used (ver. 5.5.11) (Supplementary Fig. S1).

3 RESULTS

3.1 Web-based bioinformatics tool

Various analysis methods designed to be easily accessible to investigators without bioinformatics or statistics expertise are available at the APPEX Web site (Fig. 1). APPEX consists of two parts: the APPEX analyzer, which determines signatures associated with cancer outcome, and the public dataset explorer, in which a user explores previously published cancer patient cohorts and directly applies them to the APPEX analyzer (Supplementary Fig. S2A).

APPEX currently provides seven independent workflows for searching a signature (Fig. 1). For user convenience, we defined a short name for each analysis method: *CoxSingle*, *CoxMulti*, *SuperPC*, *IntransSingle*, *IntransMulti*, *TimeRoc* and *Multivariate*. Although each workflow has its own statistical model, they all have identical execution flows (Supplementary Fig. S3). When accessing the APPEX analyzer, users select one of the seven tools by clicking a button (Supplementary Fig. S2B). Next, a simple copy and paste action or file upload is performed to submit a user's own data to the APPEX server. If there is a history of previous analyses, a user may choose one of them. Parameters of each analysis are then configured, and when the analysis is completed, information about how to access the result is delivered to a user's e-mail address. Details of seven analysis workflows and APPEX operating policy are available in the Supplementary Results.

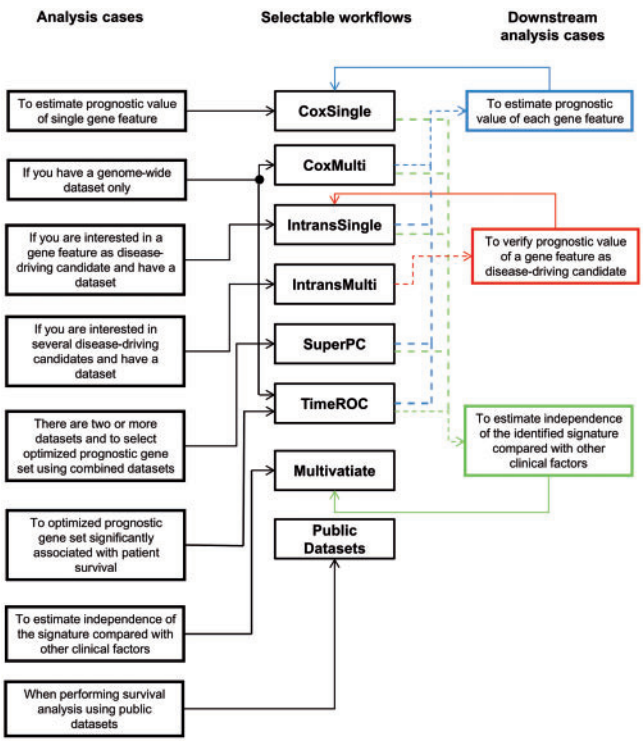

Fig. 1. Typical analysis cases for selecting APPEX workflows

3.2 Public patient cohorts

Currently, we have collected 236 cohorts from the NCBI GEO and constructed a database for users to explore. When a user chooses one of the datasets in the APPEX dataset explorer and clicks an analysis method, the APPEX analyzer will apply the data to an analysis method selected by the user (Supplementary Fig. S4). The generated file is automatically saved in the user's storage area and is accessible at a later date.

3.3 Prognostic signatures identified by APPEX

To validate the APPEX utility, we demonstrate practical examples of the APPEX application including the novel finding of a prognostic and predictive signature in bladder cancer. We describe two important cases of cancer prognosis: disease progression and the recurrence of non-muscle invasive bladder cancer. Detailed explanations of examples are provided in the Supplementary Results.

4 CONCLUSIONS

Because two or more combined approaches are frequently needed to obtain a practical signature consisting of a small number of genes, we suggest practical guidelines for selecting APPEX workflows in Figure 1. APPEX may be the best choice when users try to discover significant novel factors to predict diverse behaviors of cancer.

Funding: This work was supported by grants from the stem cell (2012M3A9B4027954) and genomics (2012M3A9D1054670) program of the National Research Foundation of Korea funded by the Ministry of Science, ICT and Future Planning and a KRIBB Research Initiative grant.

Conflicts of interest: none declared.

REFERENCES

Aguirre-Gamboa,R. *et al.* (2013) SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One*, **8**, e74250.

Bair,E. and Tibshirani,R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.*, **2**, E108.

Corradi,L. *et al.* (2009) Survival Online: a web-based service for the analysis of correlations between gene expression and clinical and follow-up data. *BMC Bioinformatics*, **10(Suppl. 12)**, S10.

Cox,D.R. (1972) Regression models and life-tables. *J. Roy. Stat. Soc. Ser. B Methodol.*, **34**, 187–220.

Gao,J. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**, pl1.

Heagerty,P.J. *et al.* (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, **56**, 337–344.

Lee,J.S. *et al.* (2010) Expression signature of E2F1 and its associated genes predict superficial to invasive progression of bladder tumors. *J. Clin. Oncol.*, **28**, 2660–2667.

Madden,S.F. *et al.* (2013) BreastMark: an integrated approach to mining publicly available transcriptomic datasets relating to breast cancer outcome. *Breast Cancer Res.*, **15**, R52.

Paik,S. *et al.* (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.*, **351**, 2817–2826.

Reinhold,W.C. *et al.* (2012) 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.*, **72**, 3499–3511.

Ringner,M. *et al.* (2011) GOBO: gene expression-based outcome for breast cancer online. *PLoS One*, **6**, e17911.