BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 20 2010, pages 2649–2650
doi:10.1093/bioinformatics/btq487

Databases and ontologies

Advance access publication August 24, 2010

ICPS: an integrative cancer profiler system

Xin-yu Zhang[1,*], Lin Shi[2], Yan Liu[2], Feng Tian[2], Hai-tao Zhao[3], Xiao-ping Miao[4], Ming-lie Huang[1] and Xiao-yan Zhu[1,*]

[1]State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, [2]Ministry of Education Key Laboratory of Bioinformatics, School of Biomedicine, Tsinghua University, Beijing 100084, [3]Department of Liver Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing and [4]Department of Epidemiology and Biostatistics, Tongji Medical College of Huazhong University of Science and Technology, Wuhan, China

Associate Editor: Alex Bateman

ABSTRACT

Founded upon the database of 570 public signatures, ICPS is a web-based application to obtain biomarker profiles among 11 common cancers by integrating genomic alterations with transcription signatures on the basis of a previously developed integrative pipeline. ICPS supports both public data and user's in-house data, and performs meta-analysis at a cancer subtype level by combining heterogeneous datasets. Finally, ICPS returns the robust gene signature containing potential cancer biomarkers that may be useful to carcinogenesis study and clinical cancer diagnosis.

Availability: http://server.bioicps.org

Contact: zhxy@mail.tsinghua.edu.cn; zxy-dcs@mail.tsinghua.edu.cn

Received on March 24, 2010; revised on August 15, 2010; accepted on August 17, 2010

1 INTRODUCTION

DNA microarrays and serial analysis of gene expression (SAGE) have been applied widely to the measurements of gene expression profiles in tumor tissues versus corresponding normal tissues over the last several years. A large amount of resources of global quantitative gene expression profiles in cancer are now publicly available, such as Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo), Stanford Microarray Database (SMD, http://genome-www5.stanford.edu) and Cancer Genome Anatomy Project (CGAP, http://cgap.nci.nih.gov/). However, any single high-throughput technology experiment is merely able to capture a snapshot of gene expression profile in a specific tumor sample. Furthermore, each high-throughput technique has its own limitations. For example, the results of microarray are hardly reproducible, and the SAGE and EST library techniques are subject to the sizes and reliabilities of the libraries. Therefore, it is essential and still challenging to develop computational integration analysis tools that can discover the robust gene signature (RGS) by combining all pieces into a most complete picture of cancer tissues (Rhodes and Chinnaiyan, 2005).

It has been discovered that many genome alterations, such as loss of heterozygosity (LOH) and chromosomal gains and losses, affect mRNA expressions (Spanakis *et al.*, 1999). ICPS applies a previously developed integrative pipeline to combine genomic variations with gene expressions, aiming to discover the RGS in different types or subtypes of cancer (Zhang *et al.*, 2007).

*To whom correspondence should be addressed.

2 IMPLEMENTATION AND VISUALIZATION

ICPS was implemented by using PERL language (Ver. 5.8.8), the statistical software R (Ver. 2.10.0) and MySQL database (Ver. 5.1.30). It was composed of a web interface, a MySQL database managing public datasets together with gene annotation information and an integrative analysis system. Gene id and unigene id were used to cross-link the entries in the selected signatures, and the former was preferable. All the datasets were collected from GEO, CGAP and SMD. The experiments were grouped by types and subtypes of cancer and then analyses were conducted. The SAM method (Tusher *et al.*, 2001) and Fisher's exact test (Beissbarth *et al.*, 2004) were applied to the mRNA expression microarray datasets and the SAGE libraries, respectively, with a significant threshold of false discovery rate (FDR) <0.1. The array CGH (aCGH) datasets were first processed by BioHMM segmentation (Marioni *et al.*, 2006), and then filtered with a threshold of $|\text{logRatio}| > 0.1$. The current version of ICPS contains 64 datasets, 570 subcancer-type signatures and 23 375 experiments, and covers 11 major types of cancers, i.e. lung cancer, liver cancer, breast cancer, bladder cancer, cervical cancer, colon cancer, kidney cancer, ovarian cancer, pancreatic cancer, prostate cancer and stomach cancer. The Cancer LOH database was constructed by reviewing previous publications obtained by searching PubMed with key words: (LOH OR 'loss of heterozygosity') AND 'cancer'. The LOH database of lung cancer have been manually reviewed and released in this version, and it will be released later for the other cancer types. The ICPS database will be updated quarterly each year.

After applying penalized *t*-statistics (Efron and Tibshirani, 2002) to the log ratio values of the parallel experiments in each selected signature, a weighted voting algorithm and the SAM method were used to perform the integrative analysis. The detailed pipeline of genomics variation data processing and the integrative algorithm were described in our previous article (Zhang *et al.*, 2007).

In addition to a genome-based heatmap view, ICPS also provided an integrative visualization for both genomic alterations and gene expression profiles of the RGS results.

3 USING ICPS

ICPS provides a user-friendly interface and detailed, step-by-step instructions. Users can select one or more interested entries among 11 cancer types and five present dataset types on ICPS homepage. All the signatures belonging to the selected cancer types and data types will be shown on the next page. The information, including the data source, cancer type, cancer subtype and platform, will be displayed by positioning the mouse pointer over any signature item, which may help users concentrate on specific signatures according to the objects of interest. In addition to the public datasets stored in ICPS database, users can also input their in-house gene expression data to expand the integration. SAM method is provided as an alternative allowing users to screen differentially expressed genes (DEGs) from the in-house data before integration.

Two integrative analysis parameters, ranking score and FDR, can be applied to filter the integration results in ICPS.

After users have confirmed the selected items and started the analysis, ICPS will refer to the background computation to arrange the information and return it to a web page containing progress messages. The whole process may last from 1 min to several hours, depending on the number and size of present signatures. Users can retrieve the results on the same page later.

4 DISCUSSION

Our integrative method had been applied to identify RGS genes in squamous cell lung cancer (SCC). As a result, 109 RGS genes were obtained and validated by performing lab experiments with reverse transcription polymerase chain reaction (RT–PCR) method as well as reviewing previous publications (Zhang *et al.*, 2007).

ICPS generates two types of human genome-based visualization, i.e. a heatmap view and an integrative map. The global profiling provides great convenience for comparing the elements of genomic alterations with expression signatures. The mean copy number of all selected aCGH datasets is displayed as a virtual sample named 'summaryCGH' in order to assist users to obtain the summary statistics of copy number variation and compare it with the expression profiles. The visualization can be constructed in PDF format and used as a 'database' which is convenient for searching genes of interest. Figure 1 exemplifies the maps at chromosome 3 in SCC, in which DNA amplifications and up-regulations of gene expression at 3q (red) and deletions and down-regulations (blue) are indicated.

As we know, Oncomine (https://www.oncomine.org) is a powerful web-based program processing cancer-specific data. However, it performs normalization and presents the processed data for further integration without any easy means of downloading either raw or processed data. In contrast, ICPS starts from the 'log ratio value' which reflects the gene expression variations in cancer tissues versus normal tissues. This strategy avoids cross-platform normalization of the signal intensity, which is a proverbial difficulty and may easily induce biases, although it is not able to fully eliminate the batch effects of such normalization. On the other hand, ICPS will no longer need login information to access the site. One of the most powerful features of ICPS is its capacity to cover not only transcriptome profiles but also genomic alterations. Next generation datasets, such as RNAseq, will be included in future versions of ICPS. In our opinion, a global integration of such datasets, at various levels, will be required to expand our understanding of oncology. ICPS provides a useful approach for this integration.

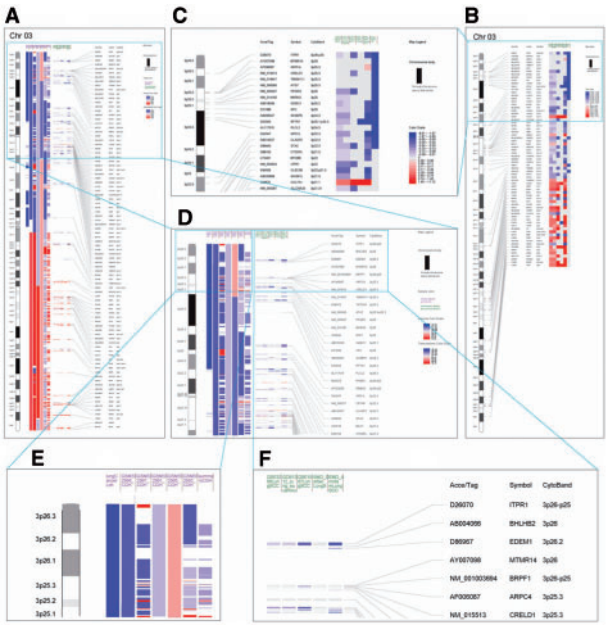

Fig. 1. The heatmap and integrative visualization of RGS genes at chromosome 3 in SCC. (**A**) Whole view of the integrative visualization. (**B**) Whole heatmap view of the RGS genes. (**C**) Heatmap view at 3p in amplified size. The RGS genes are ordered by their chromosome location. Red color corresponds to up-regulation, and blue down-regulation. (**D**) Integrative visualization at 3p in amplified size. Red color reflects amplification in CNV and up-regulation in expression, whereas blue reflects deletion and down-regulation. (**E**) Genome part in the integrative visualization shown as vertical bars, including array CGH and LOH profiles. (**F**) Transcriptome part in the integrative visualization shown as horizontal bars, including expression microarray and SAGE profiles.

ACKNOWLEDGEMENTS

The authors thank Beijing Cofly Bioinformatics Co., LTD (http://www.co-fly.net) for the technique support.

Conflict of Interest: none declared.

REFERENCES

Beissbarth,T. *et al.* (2004) Statistical modeling of sequencing errors in SAGE libraries. *Bioinformatics*, **20** (Suppl. 1), I31–I39.

Efron,B. and Tibshirani,R. (2002) Empirical bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.*, **23**, 70–86.

Marioni,J.C. *et al.* (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22**, 1144–1146.

Rhodes,D.R. and Chinnaiyan,A.M. (2005) Integrative analysis of the cancer transcriptome. *Nat. Genet.*, **37** (Suppl.), S31–S37.

Spanakis,N.E. *et al.* (1999) Aberrant p16 expression is correlated with hemizygous deletions at the 9p21-22 chromosome region in non-small cell lung carcinomas. *Anticancer Res.*, **19**, 1893–1899.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Zhang,X.Y. *et al.* (2007) Integrative analysis and validation of robust gene signature in lung cancer. *Biochem. Biophys. Res. Commun.*, **358**, 710–715.