A PROPOSED CURATION PROTOCOL FOR DISCOVERY CANCER POTENTIAL BIOMARKER CANDIDATES

Suharoschi Ramona [1*], Iuga Cristina[2*], N Crisan[3], D Pamfil[1], O Balacescu[4], IL Muntean[5]

[1] *Molecular Nutrition & Proteomics Lab, Food Science and Technology Department, University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, ramona.suaroschi@usamvcluj.ro;*
[2]*Drug Analysis Department, University of Medicine and Pharmacy "Iuliu Hatieganu" Cluj-Napoca, Cluj-Napoca, Romania,iugac@umfcluj.ro*
[3] *Municipal Clinical Hospital, University of Medicine and Pharmacy "Iuliu Hatieganu" Cluj-Napoca, Cluj-Napoca, Romania*
[4]*Oncology Institute "Dr Ion Chiricuta" Cluj-Napoca, Cluj-Napoca, Romania*
[5] *Computer Science Department, Technical University of Cluj-Napoca, Cluj-Napoca, Romania*

Abstract. *Omics technologies generate a large amount of data of various types, and their huge diversity in postgenomic era imposes the need to identify a useful functional flow connected to other resources already existed to enable more accurate discovery and selection of cancer potential biomarker candidates selection. On the other hand the biomarker discovery and validation (BMDV) studies presumed long-term and cost experiments. In order to reduce expenses due to exploration of the overwhelming research steps and to optimize the BMDV experimental design, we proposed a curation protocol applicable to the discovery of potential cancer biomarker candidates.*

Keywords: meta-data analysis, genomics, transcriptomics and proteomics data, data integration; BMDV: Biomarker discovery and validation; prostate cancer

INTRODUCTION

Prostate cancer (PCa) is a leading cause of men death in the Western world, in the meantime PCa is the most commonly diagnosed malignancy in males. Nevertheless, it is still unknown why one man gets aggressive PCa, another gets indolent PCa, and a third one is at no risk. Novel biomarkers are strongly needed to enable more accurate detection and differential diagnosis of PCa, to improve prediction of tumor aggressiveness and to facilitate the discovery of new therapeutic targets for personalized medicine. Biomarkers Definitions Working Group define 'biological marker (biomarker-BM) as a measurable parameter that characterizes an organism's state of health or disease, or a pharmacological responses to a therapeutic intervention (BDVG, 2001). Biological markers could be divided into diagnostic, prognostic, predictive, and therapeutic response markers, and are represented by different genes expression, altered or mutated genes, miRNA, transcription factors, RNA, proteins, lipids, carbohydrates, small metabolites molecules and modified expression of those molecules that can be correlated with a biological aspects or a clinical outcome (Phan, 2009). The biomarker discovery through mining a wide range of repositories is inherently and has a high degree of parallel and distributed processing. The biomarker discovery and validation flow could be based on experimental data and laboratory process or a preliminary computational (in silico) process of potential candidates' biomarker discovery and validation (BMDV). The technology employed in BMDV process could be exhaustive, based on high-throughput technology or classical, robust molecular technology, which generate a high variety of data types. Our study is based on a computational (in silico/synthetic) discovery methodology followed by an in

silico validation. Linking expressional data gained through genomic, transcriptomic, and proteomic studies to biological pathways of interest underlay with a comprehensive understanding of prostate cancer system biology.

MATERIAL AND METHOD

Our methodological approach includes the development of the BioGenProtOMICS services provided through the integrative open software platform (Figure 1). The proposed curation protocol for the discovery of cancer potential biomarker candidates developed in this study has the following main steps: 1) identification of the medical/clinical problem; 1a) selection of gene expression datasets; 2) generation of a set of candidate genes (ListG); 3) identification of differential expression genes; 3a) ranking of genes (using statistical tools for filtering significantly differentially expressed data; 3b) removal of non-significant data; 4) data overlay (overlap- potential biomarker candidates); 5) statistics filtering; 6) determination of the list of meaningful genes; 7) biological knowledge extraction (link discovery); 8) filtering service (based on PubMed literature data); 9) in silico dry BM validation by cross-validation method; 9a) split data (training set and testing set) (Figure 2).

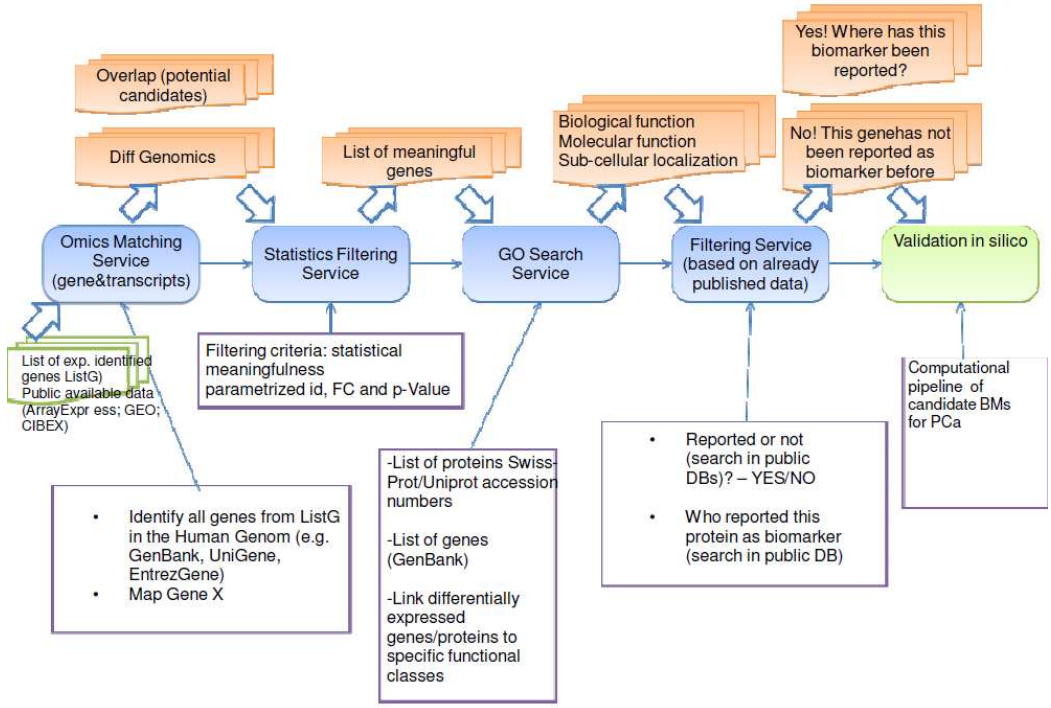

Fig. 1 Biomarker discovery and validation services provided through the intergative open software platform BioGenProtOMICS (developed by IL Muntean, C Iuga and R Suharoschi)

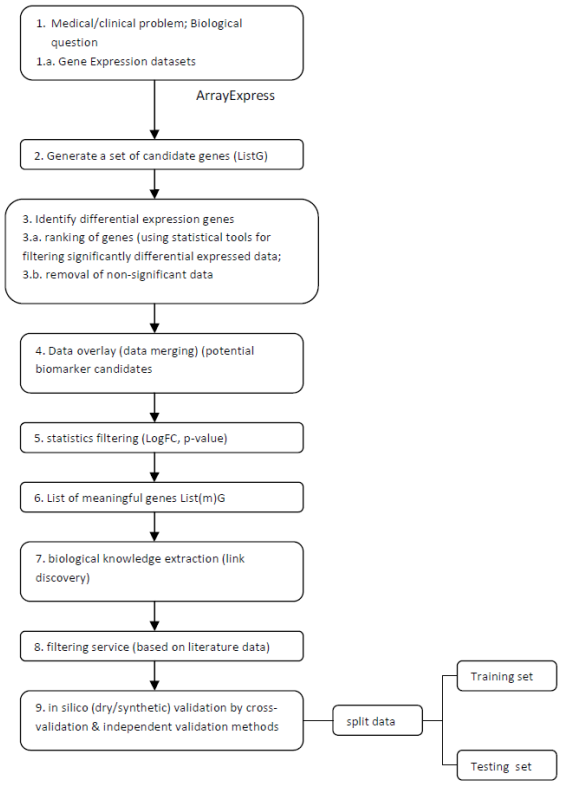

Fig. 2. The proposed curation protocol for discovery cancer potential biomarker candidates. The curation protocol correlate two approaches – literature search (PubMed) and data submitted into public repositories (ArrayExpress), that overlap statistics tools, and computational tools with system biology (in silico / synthesis / computational BM discovery and in silico / synthesis / computational BM validation proposed protocol).

RESULTS AND DISCUSSION

Currently, there is a large number of studies in published literature and publicly available databases (a keyword search for "prostate" gets approx. 116974 papers (we make a set of query with the following keywords: prostate cancer/prostate adenocarcinoma (PCa), high-grade prostatic intraepithelial neoplasia (HG-PIN)/ prostatic intraepithelial neoplasia (PIN), atypical small acinar proliferation (ASAP)/atypical adenomatous hyperplasia (AAH)), as well as benign prostatic hyperplasia (BPH), with the results presented in Table 1 (first interrogation). For the next set of interrogation we used the following keyword: differentially expressed genes (DEG) associated with the previous keywords (second interrogation) (Table 1). Our approach was to first identify the relevant literature and datasets (e.g. microarray data deposited in repositories publicly available such as ArrayExpress (Parkinson, 2011, www.ncbi.nlm.nih.gov/pubmed). After generating the list of genes (ListG), specific searches were carried out to identify the presence of these genes in different stages. Finally, a set of queries were carried out to determine the status of these differentially expressed genes in BPH, PIN and ASAP, which are important considerations in the differential diagnosis of PCa.

Table 1

Summary of literature interrogation by "keyword" (between brackets is the keyword used to interrogate PubMed) (http://www.ncbi.nlm.nih.gov/pubmed/)

First interrogation: keyword search	PubMed no of articles (07/02/2012)	Second interrogation: keyword search	PubMed no of articles (07/02/2012)
"Prostate"	116974	"Differentially expressed genes" + "Prostate"	437
"Prostate cancer"	102018	"Differentially expressed genes" + "Prostate cancer"	385
"Prostate adenocarcinoma"	15134	"Differentially expressed genes" + "Prostate adenocarcinoma"	58
"High-grade prostatic intra-epithelial neoplasia"	1010	"Differentially expressed genes" + "High-grade prostatic intraepithelial neoplasia"	4
"Benign prostatic hyperplasia"	20195	"Differentially expressed genes" + "Benign prostatic hyperplasia"	39

CONCLUSIONS

Our efforts focused on proposing this curation protocol for discovery cancer potential biomarker candidates represents the first step in undertaking biomarkers for prostate cancer in a comprehensive and systematic approach. As is inherited the data often requires subsequent validation by in silico methods. It must be pointed out that the protocol is referring to validation by cross-validation and independent validation model. Further, several high-throughput studies carried out to identify genes that are differentially expressed in prostate cancer have used tissues that do not use laser capture microdissection (LCM) to separate cancer cells from stroma. Thus, it is undecipherable in many cases if the observed difference in the expression of a special gene arises in the stroma or in cancer cells. This further underscores the importance of validating these observations using alternative methods (in silico/ dry methods).

Acknowledgements: This work was supported by the USAMV institutional grant no. 1215/6/6.02.2012.

REFERENCES

1. Phan JH, Moffitt RA, Stokes TH, Liu J, Young AN, Nie S, Wang MD., 2009, Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment. s.l.: Trends in Biotechnology, Trends in Biotechnology, Vol. 27, pp. 350-358.

2. Biomarkers Definitions Working Group, Biomarkers and surrogate endpoints: preferred definitions and conceptual framework.. 69, 2001, pp. 89-95. Clin. Pharmacol. Ther.89-95.

3. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, A. Brazma, 2007, ArrayExpress-A public database of microarray experiments and gene expression profiles, Nucleic Acids Res 35: D747-D750

4. Parkinson, H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N, Brazma A., 2011, ArrayExpress update-un archive of microarray and high-throughput sequencing-based functional genomics experiments., Nucleic Acids Research 39, D1002-D1004

5. http://www.ncbi.nlm.nih.gov/pubmed/