Development of an Online Tool and Bladder Cancer Gene Expression Database for Biomarker Evaluation

Abstract

Bladder cancer is the fourth most common cancer in males in the United States and one of the most expensive cancers to treat. Personalized treatment of bladder cancer, not yet realized, would involve using biomarkers that are associated with disease outcome or clinical subtype to guide treatment decisions. The objective of this proposal is to develop a gene expression database for biomarker evaluation in bladder cancer patients. Such a database will be a valuable resource for translational biologists studying bladder cancer and its completion will establish a new research program in bioinformatics at Eastern Connecticut State University.

1. Background and Significance

Significance: Cancer is a genetic disease whose biology is driven by mutated and abnormally expressed genes. Gene expression profiling, which simultaneously measures the expression levels of thousands of genes, is consequently a powerful tool for investigating cancer. In the past ~15 years, the identification of diagnostic and prognostic biomarkers from gene expression data has increased our understanding of cancer biology and has led to advances in the personalized treatment of cancer. A *diagnostic* biomarker is a molecule that is indicative of cancer diagnosis or existence, such as the stage, grade, and clinical subtype of a tumor; a *prognostic* biomarker is indicative of disease outcome. Although gene expression profiles of bladder cancer patients are publicly available, their analysis is time-consuming and requires computational resources and bioinformatics expertise often not available to biologists or clinician-researchers. The objective of this proposal is to develop an easy-to-use gene expression database that allows a user to evaluate whether or not a gene of interest is a diagnostic or prognostic biomarker in bladder

cancer patients. The database will be a valuable resource for translational biologists studying bladder cancer worldwide and completion of the funded project will establish a new research program in bioinformatics at Eastern Connecticut State University.

Background: Cancer is a genetic disease (Stratton et al., 2009). A cancer cell inherits or acquires mutations that enable it to grow efficiently, replicate indefinitely, support angiogenesis, avoid apoptosis, and in some cases metastasize (Hanahan and Weinberg, 2011). In the past ~15 years, gene expression profiling of human cancers has revolutionized our understanding of cancer as a genetic disease and has expedited the identification of driver mutations and biomarkers for personalized treatment. Examples of prognostic biomarkers in routine clinical use include the Onco*type*Dx and MammaPrint gene panels, which both predict the likelihood of disease recurrence in breast cancer and provide patients and clinicians with relevant information regarding the potential benefit of chemotherapy (Knauer et al., 2009; Markopoulos et al., 2011).

In the United States, bladder cancer is the fourth most common cancer in males, the eighth most common cancer in females (Siegel et al., 2012), and one of the most expensive cancers to treat (Botteman et al., 2003). At diagnosis, approximately 20-30% of bladder cancer patients harbor muscle-invasive (MI) tumors (Jacobs et al., 2010) and these patients have a five-year survival rate of approximately 43% (Stein et al., 2001). In patients harboring non-muscle invasive (NMI) tumors, progression to MI disease occurs in ~ 20% of all patients, and in ~50% of high risk patients with high grade, recurrent tumors (Cookson et al., 1997). Despite the importance of this disease there are no prognostic or predictive biomarkers in clinical use.

Related work and necessity of a bladder cancer gene expression database: Gene expression datasets are typically deposited into public databases such as the Gene Expression Omnibus (GEO; Barrett and Edgar, 2006) and ArrayExpress (Rustici et al., 2013). These databases

function primarily as data repositories where data can be downloaded and analyzed using in-house bioinformatics tools. However, despite the availability of these databases and others, there are currently no bladder cancer gene expression databases that allow for an automated and comprehensive evaluation of diagnostic and prognostic biomarkers in patients. In particular, GEO includes some curated datasets with pre-computed sets of differentially expressed genes, but does not include curated datasets for bladder cancer patients. Other databases include the KM plotter (Gyorffy et al., 2010), which generates Kaplan-Meier survival curves for lung, breast, and ovarian cancer patients only; and PrognoScan (Mizuno et al., 2009) and SurvExpress (Aguirre-Gamboa et al., 2013), which generate Kaplan-Meier curves for only a small subset of the available bladder cancer datasets, do not evaluate diagnostic biomarkers, and do not allow for patient subset selection (e.g., of patients with MI tumors only, an important clinical subgroup). On the other hand, the proposed database will allow, for example, a user to ask: *is the gene TP53 a biomarker of disease outcome in bladder cancer patients with MI tumors?* Currently, an answer to this clinically important type of question cannot be answered directly using any of the existing databases.

2. **Methodology**

A timeline for the work plan is provided in **Table 1**. Gene expression profiles of bladder cancer patients will be downloaded from GEO and Array Express and stored on the requested data server. Also requested is a server that will run an Apache HTTP (web) server which will host the website for the bladder cancer gene expression database. A website interface will be created allowing a user to specify the following: a gene of interest, a patient subgroup (e.g., based on gender, tumor stage, grade, etc) to analyze, and the clinical endpoint or variables of interest (e.g.,

tumor grade, disease-free survival, etc) to assess. Students interested in carrying out research in my Bioinformatics Laboratory will be encouraged to contribute to the design and development of the website.

Appropriate graphical and statistical summaries will be displayed on the webpage indicating whether or not the selected gene is a statistically significant biomarker in the selected patient

Table 1. Timeline for work plan. *, see **Outcomes and Reporting** for more information

<table>
<tr><th>July-August</th><th>September-December</th><th>January - April</th><th>May – June</th></tr>
<tr><td rowspan="2">Data server and Apache HTTP server set-up, data download and processing</td><td colspan="2">Development of website</td><td rowspan="3">Manuscript preparation and submission to a peer-reviewed journal such as BMC Urology*</td></tr>
<tr><td>Submission of abstract to AACR Annual Meeting*</td><td>Data download of newly available datasets; website testing and refinement</td></tr>
<tr><td colspan="3">Development of R scripts for statistical analysis</td></tr>
</table>

group, across available patient cohorts. Specifically, the *R* programming language (http://cran.us.r-project.org) will be used for all statistical analyses and results communicated back to the web server using the Common Gateway Interface (CGI). The non-parametric Wilcoxon-Rank Sum test (*wilcox.test* function in *R*) will be used to test for differential expression (whether expression levels are significantly different between two categories) between tumor and normal bladder samples, low and high grade tumors, NMI and MI tumors, and node-positive and node-negative tumors. Cox proportional regression (*coxph* function) will be used for survival analyses and Kaplan-Meier curves will be generated for patients with low and high expression levels of the specified gene for a chosen endpoint including disease-free survival, recurrence-free survival, and overall survival.

3. Statement of Qualification. Although this is my first CSU-AAUP Research grant proposal, I have a demonstrated publication history in gene expression analysis. I have previously developed a classification method and software for selecting cancer cell lines based on their molecular similarity with human tumors (with respect to tissue of origin, stage, grade, and prognosis), based on their gene expression profiles (Dancik et al., 2011); analyzed bladder and lung tumor gene expression profiles and found that prognostic signatures are dependent on genes related to cell-cycle proliferation (Dancik and Theodorescu, 2014); and identified gene expression signatures of common bladder cancer mutations and evaluated their ability to predict cancer stage, progression, and survival (Dancik et al., 2013).

4. Expected Outcomes

At the completion of this project, the bladder cancer research community will have an easy-to-use online resource for evaluation of bladder cancer diagnostic and prognostic biomarkers from gene expression data. Such a resource promises to increase our understanding of bladder cancer biology and to aid in the identification of clinically relevant biomarkers. In addition, I will prepare presentations for the CSUS Research Conference and the American Association for Cancer Research (AACR) Annual Meeting in 2015; and prepare a manuscript for publication in a peer-reviewed journal such as *BMC Urology*. The AACR annual meeting is the largest cancer research meeting in the world, which regularly has over 18,000 attendees. *BMC Urology* is an open-access peer reviewed journal that covers all aspects of urological disorders such as bladder cancer.

Bibliography

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

Botteman MF, Pashos CL, Redaelli A, et al. (2003) The health economics of bladder cancer: a comprehensive review of the published literature. *Pharmacoeconomics* 21: 1315-1330.

Cookson MS, Herr HW, Zhang ZF, et al. (1997) The treated natural history of high risk superficial bladder cancer: 15-year outcome. *J Urol* 158: 62-67.

Dancik G and Theodorescu D. (2014) Robust prognostic gene expression signatures in bladder cancer and lung adenocarcinoma depend on cell cycle related genes. *PLoS ONE* 9(1): e85249

Dancik GM, Owens CR, Iczkowski KA, et al. (2013) A cell of origin gene signature indicates human bladder cancer has distinct cellular progenitors. *Stem Cells*. Accepted.

Dancik GM, Ru Y, Owens CR, et al. (2011) A framework to select clinically relevant cancer cell lines for investigation by establishing their molecular similarity with primary human cancers. *Cancer Res* 71: 7398-7409.

Gyorffy B, Lanczky A, Eklund AC, et al. (2010) An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat* 123: 725-731.

Hanahan D and Weinberg RA. (2011) Hallmarks of cancer: the next generation. *Cell* 144: 646-674.

Jacobs BL, Lee CT and Montie JE. (2010) Bladder cancer in 2010: how far have we come? *CA Cancer J Clin* 60: 244-272.

Knauer M, Straver M, Rutgers E, et al. (2009) The 70-gene MammaPrint signature is predictive for chemotherapy benefit in early breast cancer. *Breast* 18: S36-S37.

Markopoulos C, Xepapadakis G, Venizelos V, et al. (2011) Clinical Use of OncotypeDX Recurrence Score as an Adjuvant-Treatment Decision Tool in Early Breast Cancer Patients. *European Journal of Cancer* 47: S379-S379.

Mizuno H, Kitada K, Nakai K, et al. (2009) PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med Genomics* 2: 18.

Rustici G, Kolesnikov N, Brandizi M, et al. (2013) ArrayExpress update--trends in database growth and links to data analysis tools. *Nucleic Acids Res* 41: D987-990.

Siegel R, Naishadham D and Jemal A. (2012) Cancer statistics, 2012. *CA Cancer J Clin* 62: 10-29.

Stein JP, Lieskovsky G, Cote R, et al. (2001) Radical cystectomy in the treatment of invasive bladder cancer: long-term results in 1,054 patients. *J Clin Oncol* 19: 666-675.

Stratton MR, Campbell PJ and Futreal PA. (2009) The cancer genome. *Nature* 458: 719-724.