A Hypothesis-Directed Approach to the Targeted Development of a Multiplexed Proteomic Biomarker Assay for Cancer

Emily M. Mackay[1,*], Jennifer Koppel[1,*], Pooja Das[1], Joanna Woo[1], David C. Schriemer[2] and Oliver F. Bathe[3,4]

[1]Department of Medical Sciences, [2]Department of Chemistry, [3]Department of Surgery, [4]Department of Oncology, University of Calgary, Calgary, AB, Canada. *These authors contributed equally to this work.

ABSTRACT: In recent years, hundreds of candidate protein biomarkers have been identified using discovery-based proteomics. Despite the large number of candidate biomarkers, few proteins advance to clinical validation. We propose a hypothesis-driven approach to identify candidate biomarkers, previously characterized in the literature, with the highest probability of clinical applicability. A ranking method, called the "hypothesis-directed biomarker ranking" (HDBR) system, was developed to score candidate biomarkers based on seven criteria deemed important in the selection of clinically useful biomarkers. To demonstrate its application, we applied the HDBR system to identify candidate biomarkers for the development of a diagnostic test for the early detection of colorectal cancer. One-hundred and fifty-one candidate biomarkers were identified from the literature and ranked based on the specified criteria. The top-ranked candidates represent a group of biomarkers whose further study and validation would be justified in order to expedite the development of biomarkers that could be used in a clinical setting.

KEYWORDS: biomarker, diagnosis, colorectal cancer, targeted proteomics

CITATION: Mackay et al. A Hypothesis-Directed Approach to the Targeted Development of a Multiplexed Proteomic Biomarker Assay for Cancer. *Cancer Informatics* 2015:14 65–70 doi: 10.4137/CIN.S24388.

RECEIVED: January 28, 2015. **RESUBMITTED:** March 09, 2015. **ACCEPTED FOR PUBLICATION:** March 13, 2015.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Technical Advance

FUNDING: This work was supported by the Canadian Research Chair in Chemical Biology Alberta, and Alberta Innovates Health Solutions as a Heritage Scholar. The authors confirm that the funders had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: bathe@ucalgary.ca

The field of cancer biomarker research for detection, prognosis, and therapeutic response has advanced rapidly in recent years due to improvements in analytical technologies. Potential gene and transcript biomarkers have been identified; however, the abundance of mRNA does not necessarily correlate with the amount of functional protein expressed, due to differences in post-transcriptional processing, translation, and protein degradation.[1,2] Therefore, protein markers may represent a more accurate reflection of the pathophysiological state.

The large dynamic range of protein abundance, multiple isoforms, post-translational modifications, and variations in expression with time and cell type makes characterization of the proteome challenging. With advances in mass spectrometry-based technologies, it is now possible to survey thousands of proteins in a sample, although obtaining a complete quantitative profile of the entire proteome is still not possible.[3] One significant challenge involves the wide range of protein abundance in biological samples, such as human plasma. Lowly abundant proteins (eg, cytokines) may be masked by the presence of extremely abundant proteins (eg, albumin) as the dynamic range of the proteome spans ten of orders of magnitude.[4–6]

Despite technological advancements, no protein biomarkers identified through proteomic discovery experiments have been approved by the US Food and Drug Administration (FDA) for use in routine clinical practice.[7] While hundreds of candidate biomarkers have been identified by discovery-phase studies, subsequent validation of these proteins is lacking due to numerous challenges that must be overcome. These challenges include the high rate of false positive identifications during discovery-phase experiments, the lack of biological relevance for some candidates, and the limited number of quantitative immunoassays available to verify differential abundance of candidate proteins as is needed to develop a clinical test. With the technical challenges and high costs associated with the development of immunoassays, it is not feasible to design novel assays for testing large numbers of candidate biomarkers.[8,9]

Given that a significant obstacle in the biomarker development process is determining which candidate proteins should be taken forward to subsequent validation steps, efforts need to focus on rational selection of proteins with the highest chance of success in clinical applications. Several protein biomarker pipelines have been proposed detailing the processes from discovery through validation.[10,11] In addition, Phan et al.[12]

created omniBiomarker, a knowledge-driven program that uses previously validated genes to guide the selection of the most biologically relevant candidate biomarkers for the disease of interest.

Here we describe a method by which candidate cancer biomarkers can be identified through an initial literature search and subsequently ranked using a scoring system we called the "hypothesis-directed biomarker ranking" (HDBR) system (Fig. 1). By applying specific selection criteria, one can prioritize candidates with the most probable value as a biomarker for the disease and application of interest. With this method, repetition of large-scale discovery experiments can be avoided. To demonstrate the utility of the HDBR system, the criteria were employed to identify protein biomarkers for a serum-based diagnostic test for the early detection of colorectal cancer (CRC).

Initially, a comprehensive literature search is conducted on PubMed and/or EMBASE to identify candidate proteins relevant to the cancer and application of interest. For this example, the search terms used in the PubMed database included various combinations of "colorectal", "colon", "rectum", "adenoma", "carcinoma", "cancer", "biomarker", "diagnosis", "microarray", "protein", and "serum".

Proteins identified in our initial literature review were then scored using the seven criteria outlined in Table 1. First, the candidate biomarker was assessed using the Oxford Center for Evidence-based Medicine (OCEBM) Levels of Evidence,[13] which scores the quality of evidence available on a particular test based on the study design, validation of results in multiple centers, and use of appropriate reference standards.

Next, specimen source and biomolecule type were evaluated. As a serum biomarker was desired for this specific diagnostic test, more points were assigned to biomarkers identified in the blood over those identified in tissue or cell lines. Similarly, as protein biomarkers were preferred in this instance, two points were given to molecules such as caspase 3, which is demonstrated to be associated with CRC at the protein level, while only one point was given to molecules such as nitric oxide synthase, which is identified to be associated with CRC only at the RNA level.

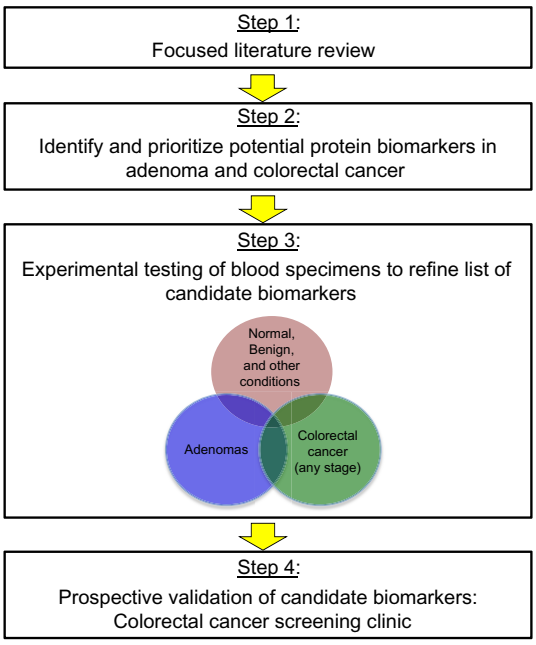

Figure 1. General description of the HDBR system approach to develop a multiplexed protein assay for the early diagnosis of colorectal cancer.

Biomarkers were then scored based on the throughput format in which they were identified to be associated with CRC. False positives from multiple hypothesis testing and erroneous statistical inferences are common at the discovery phase of research using high-throughput assays due to the large number of variables measured and the statistical tests applied to each of these variables.[14,15] Therefore, biomarkers assessed by high-throughput assays and then confirmed with low-throughput assays were given a higher score than biomolecules evaluated using solely high-throughput techniques.

The relevance of the candidate biomarkers in the context of colon adenomas and carcinomas was then scored. The highest score was given to biomarkers present in adenomas and potentially in carcinomas, as this test was aimed at early diagnosis of CRC at the adenoma stage. A lower score was assigned to biomarkers present in colon carcinomas only, and a score of zero was given to biomarkers that were identified to be present in colon adenomas and carcinomas while also present in other conditions, as these biomarkers would not be specific for the diagnosis of CRC.

Next, the functions of the biomarker candidates were assessed. As the HDBR system was designed to identify biomarkers relevant to cancer, proteins with known functions related to the hallmarks of cancer[16,17] are of greatest interest. Therefore, proteins with a known pro-neoplastic function are given a higher score in the ranking system.

Finally, the number of publications related to the biomarker candidates was examined. Points are assigned based on this criterion because it is assumed that the number of publications is a reflection of the degree of interest in that biomarker by the scientific community.

Using the results from the primary literature review, 151 candidate biomarkers were identified to be potentially useful for the early diagnosis of CRC using a serum-based diagnostic test. The distribution of the total scores is depicted in Figure 2A. The average score was 9 ± 2 (range 2–15). Figures 2B–H show the distribution of scores as a function of the various criteria utilized to derive a final HDBR score. Table 1 (Supporting Information) depicts the ranking of the 151 candidate biomarkers.

The top-ranked biomarkers included carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1), matrix metalloproteinase 9 (MMP9), and insulin-like growth factor II (IGF2). The top-ranked candidate biomarkers shared common features contributing to a higher rank: they were proteins that were detectable in the blood, had a close association

Table 1. Proteomic biomarker ranking system used to score protein biomarkers for the early diagnosis of colorectal cancer.

KEY CRITERIA	CRITERIA FOR PROTEIN ASSAY DEVELOPMENT	SCORE				
		0	1	2	3	4
Oxford levels of evidence*	Level 4		X			
	Level 3a and 3b			X		
	Level 2a, 2b, and 2c				X	
	Level 1a, 1b, and 1c					X
Specimen source**	Cell lines		X			
	Tissue			X		
	Blood				X	
Biomolecule evaluated	RNA		X			
	Protein			X		
Assay throughput format	High-throughput (multiplexed) assay with conflicting outcome on low-throughput confirmatory test	X				
	High-throughput (multiplexed) assay		X			
	High-throughput (multiplexed) assay with low-throughput confirmatory test			X		
Relevance to target disease entity	Present in adenoma or carcinoma, but possibly also other conditions	X				
	Present in carcinoma only		X			
	Present in adenoma and possibly also in carcinoma			X		
Relevance to biological function	Hallmark of cancer unknown or absent	X				
	Hallmark of cancer identified		X			
Number of NCBI PubMed citations	<100	X				
	≥ 100		X			

Notes: *Level 5 evidence excluded. **Human.

to colorectal adenoma or CRC, and were measurable using low-throughput assays. Studies demonstrating a high level of evidence for the utility of these proteins for diagnostic purposes had been conducted. Additionally, these studies involved the use of good reference standards as well as comparisons to independent normal and nonmalignant controls.

The HDBR model may be criticized for its tendency to penalize biomarker candidates that are relatively novel. A novel protein typically has a poorly defined or unknown function and is not cited frequently in the literature. In addition, novel proteins are not likely to have high levels of evidence of utility in a diagnostic test. In order to avoid the exclusion of novel proteins with unknown functions or limited citations from the final list of candidate biomarkers, only one point was assigned to these two criteria.

Other ranking methods have been described for transcriptomic and proteomic research. Chan et al.[18] performed a meta-analysis on 25 independent gene expression studies on CRC. This ranking system was based on the number of studies reporting on the differential gene expression, the number of samples used in the study, and the average fold change seen. The direction of differential gene expression was also considered. Similar methods have been used to rank candidate microRNA biomarkers for CRC detection based on the number of microRNA expression studies, sample size, and direction of differential expression.[19] Sagynaliev et al.[20] proposed the creation of a "data warehouse" containing all of the currently available information from transcriptomic and proteomic studies with regard to CRC. This resource would, in turn, serve as a starting point for further investigation of genes repeatedly observed to be differentially expressed in multiple independent studies, confirmed both at the transcript and protein levels.

The HDBR model is complementary to previously described ranking methods. It is inclusive of high-throughput techniques but has the ability to preferentially select candidate proteins that have gone through a validation step using low-throughput assays. It expands beyond cell lines and tissue to include other specimens (eg, blood), and has the ability to prioritize protein over gene biomarkers. The HDBR system could make use of previously compiled resources such as Sagynaliev et al's "data warehouse" of expression studies,[20] the extensive list of candidate cancer protein biomarkers compiled by Polanski and Anderson,[21] and the thorough review of cancer-associated proteins and markers identified in Jimenez and coworker's meta-analysis of CRC studies.[22] Candidate biomarkers from these previously assembled resources would be subject to the ranking criteria of the HDBR system to decide on the most promising biomarkers for further study, potentially increasing the efficiency of identifying clinically relevant bio-

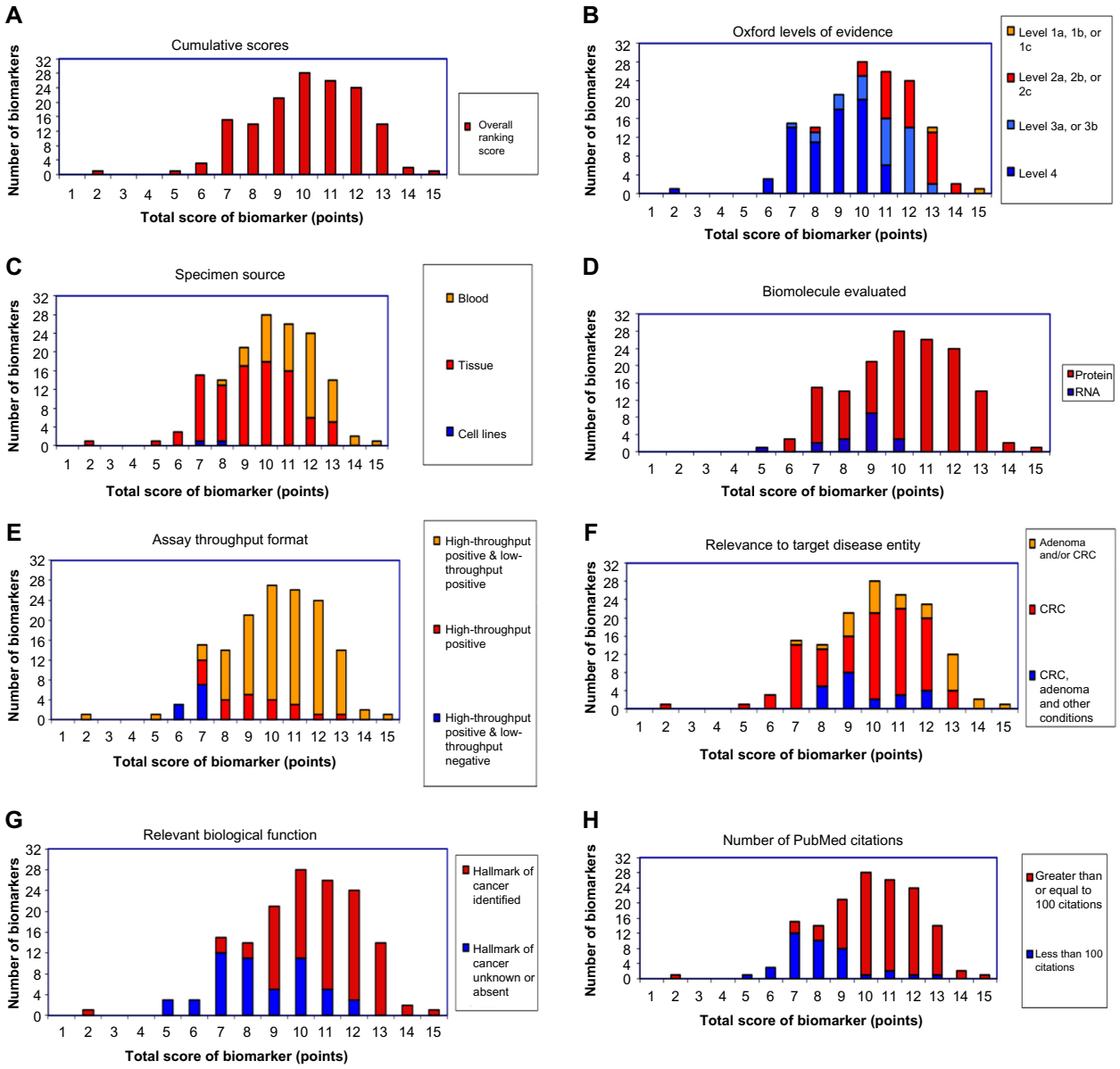

Figure 2. (**A**) Distribution of ranked colorectal cancer biomarker scores. The average score was 9 ± 2 with a range 2–15. Contributions of each criterion to the ranking are summarized in subsequent panels. (**B**) Oxford levels of evidence. (**C**) Specimen source. (**D**) Biomolecule evaluated. (**E**) Assay throughput format. (**F**) Relevance to target disease entity. (**G**) Relevant biological function. (**H**) Number of PubMed citations.

markers for translational purposes. A summary of the ranking parameters used in several of the previously noted candidate cancer biomarker investigations is presented to highlight some of the commonly used ranking measures (Table 2).

There are also unique features to the HDBR system that may better inform the biomarker developer. Foremost is the inclusion of a score derived from OCEBM levels of evidence. This is widely used to grade the evidence supporting clinical practice guidelines. Since the ultimate use for any biomarker is in the clinic, a particularly high score was assigned based on a high level of evidence supporting its clinical applicability. The use of one bibliometric measure (number of citations), which reflects the scientific community's interest in the biomarker, is also unique. Impact factor as another bibliometric criterion was deliberately excluded because it was considered to potentially interact with the OCEBM level of evidence. Finally, the HDBR system leverages the use of transcriptional data for proteomic biomarker research while at the same time limiting the contribution of candidate biomarkers that have no accompanying protein information.

Validation is required to determine the ability of the HDBR system to accurately predict biomarker candidates that have utility in the purpose for which the scoring criteria were designed. In addition, the list of top-ranked biomarkers generated by the HDBR system will need to be assessed experimentally. This might include targeted mass spectrometry using a selected reaction monitoring (SRM)/multiple reaction monitoring (MRM) workflow, as this is a sensitive

Table 2. Ranking parameters for candidate cancer markers used in other ranking methods.

CHAN et al.[18]	MA et al.[19]	POLANSKI et al.[21]
Number of studies reporting differential expression of candidate gene	Number of studies in agreement about miRNA expression	Number of total citations; Number of recent (>2004) citations
Total number of samples used in studies	Total study size	Known plasma concentration
Average fold change	Direction of differential expression (increased or decreased expression in cancer)	Marker in current clinical use
		Commercially available antibody

and accurate method for the quantification of target molecules.[23] Workflows for designing MRM transitions have been described for targeted proteomic investigations[24] and could be utilized to design experimental tests for the biomarker candidates identified through the HDBR system. In addition, Kim et al.[25] have described a detailed framework for choosing a biomarker validation platform using SRM/MRM assays, immunoassays, or immuno-mass spectrometry.

Going forward, we intend to test the performance of the HDBR system in sera from patients with CRC and disease-free controls. To do this, representative tryptic peptides derived from candidate proteins will be synthesized and labeled with stable isotopes. Peptides from proteins that are highly ranked in the HDBR system will be synthesized, as well as peptides from proteins that have a lower ranking. Using the MRM workflow, MS will be used to quantify each protein to determine the differential abundance of each protein in the disease state and in disease-free controls. The validity of the ranking system to target useful proteins will be demonstrated if higher ranked proteins have a greater capability to identify the disease state.

If the performance of the HDBR system is adequate, then it may be possible to automate the analysis, using standalone software or by leveraging software that is currently available. For example, literature search software to query text-based data can be used to replace a manual search. One example is the Agilent Literature Search software, a meta-search tool for automatically querying multiple text-based search engines.[26] In addition, the multi-attribute rankings output can be enhanced using an application such as LineUp,[27] which is a scalable visualization tool that uses bar charts to depict the relative contribution of each ranking criterion. This would enable the user to explore the effects of changes and refinements in the parameters used to rank biomarker candidates.

The HDBR system represents a method for the identification and ranking of candidate biomarkers that could potentially expedite the biomarker development process by focusing resources on biomarkers already identified to have disease-specific relevance and clinical utility. The system has some unique attributes compared to other knowledge-driven biomarker candidate ranking approaches. The HDBR system has the benefit of being widely applicable to various purposes from detection and prognosis to predictive biomarkers, and can be tailored specifically to the disease of interest.

Author Contributions

Conceived and designed experiments: OB, JK, DS. Analyzed the data: EM, PD, JK, JW. Wrote the first draft of the manuscript: EM, OB. Contributed to the writing of the manuscript: EM, OB, DS, JK, JW, PD. Agree with manuscript results and conclusions: EM, OB, DS, JK, JW, PD. Jointly developed the structure and arguments for the paper: EM, OB, DS, JK, JW, PD. Made critical revisions and approved the final manuscript: EM, OB, DS, JK, JW, PD.

Supplementary Material

Supplementary Table 1. Proteomic biomarker ranking system applied to literature review of candidate biomarkers for the early detection of colorectal cancer for a serum based proteomic test. One-hundred and fifty-one candidate biomarkers were ranked based on the seven criteria outlined in the HDBR system.

REFERENCES

1. de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C. Global signatures of protein and mRNA expression levels. *Mol Biosyst*. 2009;5:1512–26.
2. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*. 2012;13:227–32.
3. Nilsson T, Mann M, Aebersold R, Yates JR III, Bairoch A, Bergeron JJ. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods*. 2010;7:681–5.
4. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics*. 2002;1:845–67.
5. Anderson L, Hunter CL. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol Cell Proteomics*. 2006;5:573–88.
6. Mallick P, Kuster B. Proteomics: a pragmatic perspective. *Nat Biotechnol*. 2010;28:695–709.
7. Anderson NL. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. *Clin Chem*. 2010;56:177–85.
8. Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol*. 2006;24:971–83.
9. Paulovich AG, Whiteaker JR, Hoofnagle AN, Wang P. The interface between biomarker discovery and clinical validation: the tar pit of the protein biomarker pipeline. *Proteomics Clin Appl*. 2008;2:1386–402.
10. Whiteaker JR, Lin C, Kennedy J, et al. A targeted proteomics – based pipeline for verification of biomarkers in plasma. *Nat Biotechnol*. 2011;29:625–34.
11. Freue G, Meredith A, Smith D, Bergman A. Computational biomarker pipeline from discovery to clinical implementation: plasma proteomic biomarkers for cardiac transplantation. *PLoS Comput*. 2013;9(4):e1002963.
12. Phan J, Young A, Wang M. omniBiomarker: a web based application for knowledge-driven biomarker identification. *IEEE Trans Biomed Eng*. 2012;60(12):3364–67.
13. Philips B, Ball C, Sackett D, Badenoch D. Philips: Oxford centre for evidence-based medicine. Google Scholar; 2001.
14. Pounds SB. Estimation and control of multiple testing error rates for microarray studies. *Brief Bioinform*. 2006;7:25–36.

15. Tinker AV, Boussioutas A, Bowtell DDL. The challenges of gene expression microarrays for the study of human cancer. *Cancer Cell*. 2006;9:333–9.
16. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. 2000;100:57–70.
17. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144:646–74.
18. Chan SK, Griffith OL, Tai IT, Jones SJM. Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers. *Cancer Epidemiol Biomarkers Prev*. 2008;17:543–52.
19. Ma Y, Zhang P, Yang J, Liu Z, Yang Z, Qin H. Candidate microRNA biomarkers in human colorectal cancer: systematic review profiling studies and experimental validation. *Int J Cancer*. 2011;130(9):2077–87.
20. Sagynaliev E, Steinert R, Nestler G, Lippert H, Knoch M, Reymond MA. Web-based data warehouse on gene expression in human colorectal cancer. *Proteomics*. 2005;5:3066–78.
21. Polanski M, Anderson NL. A list of candidate cancer biomarkers for targeted proteomics. *Biomark Insights*. 2007;1:1–48.
22. Jimenez CR, Knol JC, Meijer GA, Fijneman RJA. Proteomics of colorectal cancer: overview of discovery studies and identification of commonly identified cancer-associated proteins and candidate CRC serum markers. *J Proteomics*. 2010;73:1873–95.
23. Picotti P, Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Methods*. 2012;9:555–66.
24. Walsh GM, Lin S, Evans DM, Khosrovi-Eghbal A, Beavis RC, Kast J. Implementation of a data repository-driven approach for targeted proteomics experiments by multiple reaction monitoring. *J Proteomics*. 2009;72:838–52.
25. Kim JW, You J. Protein target quantification decision tree. *Int J Proteomics*. 2013;2013:1–8.
26. Available from: http://www.agilent.com/labs/research/litsearch.html.
27. Gratzl S, Lex A, Gehlenborg N, Pfister H, and Streit M. LineUp: visual analysis of multi-attribute rankings – Best Paper Award. In: IEEE Transactions on Visualization and Computer Graphics (InfoVis '13), Volume 19, Issue 12, 2277–86; 2013.