Published in final edited form as:
Cancer Epidemiol Biomarkers Prev. 2017 May ; 26(5): 711–718. doi:10.1158/1055-9965.EPI-16-0732.

Proteomic profiling of serial pre-diagnostic serum samples for early detection of colon cancer in the U.S. military

Stephanie Shao[1,2], **Benjamin A Neely**[3], **Tzu-Cheg Kao**[1], **Janet Eckhaus**[1], **Jolie Bourgeois**[1], **Jasmin Brooks**[3], **Elizabeth E. Jones**[3], **Richard R. Drake**[3], and **Kangmin Zhu**[1,2]

[1]Division of Epidemiology and Biostatistics, Department of Preventive Medicine and Biostatistics, Uniformed Services University of the Health Sciences, Bethesda, MD

[2]John P. Murtha Cancer Center, Walter Reed National Military Medical Center, Bethesda, MD

[3]Department of Cell and Molecular Pharmacology and Experimental Therapeutics and MUSC Proteomics Center, Medical University of South Carolina, Charleston, SC

Abstract

Background—Serum proteomic biomarkers offer a promising approach for early detection of cancer. In this study, we aimed to identify proteomic profiles that could distinguish colon cancer cases from controls using serial pre-diagnostic serum samples.

Methods—This was a nested case-control study of active duty military members. Cases consisted of 264 patients diagnosed with colon cancer between 2001 and 2009. Controls were matched to cases on age, gender, race, serum sample count, and collection date. We identified peaks that discriminated cases from controls using random forest data analysis with a 2/3 training and 1/3 validation data set. We then included epidemiologic data to see if further improvement of model performance was obtainable. Proteins that corresponded to discriminatory peaks were identified.

Results—Peaks with *m/z* values of 3119.32, 2886.67, 2939.23, and 5078.81 were found to discriminate cases from controls with a sensitivity of 69% and a specificity of 67% in the year before diagnosis. When smoking status was included, sensitivity increased to 76% while histories of other cancer and tonsillectomy raised specificity to 76%. Peaks at 2886.67 and 3119.32 *m/z* were identified as histone acetyltransferases while 2939.24 *m/z* was a transporting ATPase subunit.

Conclusion—Proteomic profiles in the year before cancer diagnosis have the potential to discriminate colon cancer patients from controls and the addition of epidemiologic information may increase the sensitivity and specificity of discrimination.

Impact—Our findings indicate the potential value of using serum pre-diagnostic proteomic biomarkers in combination with epidemiologic data for early detection of colon cancer.

Corresponding author: Kangmin Zhu, Division of Epidemiology and Biostatistics, Department of Preventive Medicine and Biostatistics, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814. Phone: 301-816-4786, Fax: 301-881-7197, Kangmin.zhu@usuhs.edu.

Conflict of interest: The authors declare that they have no conflicts of interest.

Disclaimer

The opinions and assertions expressed in this article represent the private views of the authors and do not reflect the official views of the U.S. Departments of the Army and Navy, the Uniformed Services University of the Health Sciences, the Department of Defense, National Cancer Institute, or U.S. Government. Nothing in the presentation implies any Federal/Department of Defense endorsement.

Introduction

Proteomic profiling of prediagnostic serum samples offers a promising approach for early detection and prediction of cancer that has the potential to minimize expense and invasiveness of current cancer screening methods. Since cancer develops from aberrant DNA mutations that change protein expression patterns through modification of protein structure and functions, proteomic profiling of serum may reflect the pathological status of organs [1].

Many studies have been conducted to profile proteomic patterns for early detection of colorectal cancer using serum samples. In one case-control study, two protein peaks (13732.4 and 13912.3 *m/z*) classified colorectal cases from healthy controls [2]. Another case-control study reported peak *m/z* values of 1208, 1467, 1505, 1618, and 4215 that differentiated cases from controls with an accuracy close to 100% [3]. Furthermore, one study using MALDI-TOF MS protein fractionation methods reported a sensitivity of 87% and a specificity of 85% using 4 peaks (2870.7, 3084.0, 9180.5, and 13748.8 *m/z*) [4], while another study reported a sensitivity of 94.4% and a specificity of 75.5% using three peaks (1778.97, 1866.16, 1934.65, and 2022.46 *m/z*) [5].

Beyond proteomic peaks, specific blood-derived proteins have also been identified as potential biomarkers in predicting colorectal cancer [6–11]. Increased plasma levels of Apo Al (AUC=0.621) and decreased levels of C9 (AUC=0.730) were found in colorectal cancer patients compared to healthy ones [7]. Another case-control study reported a decrease in fragments of complement C3f among serum samples of colorectal patients compared to healthy controls [5]. Two immunoreactive antigens in serum, MAPKAPK3 and ACVR2B, successfully discriminated cases from controls with a sensitivity and specificity of 83.3% and 73.9%, respectively [12]. Additionally, a unique serum peptide and protein biomarker signature were reported in a large study of 126 colorectal cancer and 277 control serum samples. Results were validated among 50 cases and 82 controls (AUC=0.93) [8]. Although most, but not all [13], studies identified proteomic peaks or specific proteins that distinguished colorectal cancer patients from controls, identified biomarkers were largely different among studies.

Inconsistencies in biomarker detection may be due to the timing of sample collection. Blood, serum, and plasma samples collected at clinical diagnosis or later may identify patterns that result from anxiety or stress hormones related to the cancer diagnosis rather than the products of cancer itself [14]. Thus, samples collected prior to diagnosis are necessary for true early detection of cancer. It is possible that a colon cancer proteomic profile may appear a few years before the clinical detection of cancer where serial pre-diagnostic samples may aid in the earliest detection of a proteomic profile specific to colon cancer. Additionally, sample collection, storage, and processing might differ between clinically diagnosed cancer patients and controls [11, 15, 16], which may influence the validity of results. Furthermore, many previous studies on cancer proteomics used only blood samples [3, 4]. Demographic and epidemiologic characteristics may affect proteomic profiles and have the potential to improve detection and prediction of colon cancer.

The aim of this study was to identify serum protein profiles for detecting and predicting colon cancer using serial prediagnostic samples stored at the Department of Defense Serum Repository (DoDSR). Specifically, the study aimed to 1) investigate whether there are proteomic profiles in prediagnostic serum samples that discriminate colon cancer cases from controls; 2) use serial serum samples from the same study subject to detect the earliest time point at which identified proteomic profiles appear and assess whether profiles vary with time from diagnosis of colon cancer; as well as 3) evaluate whether there are epidemiological factors that may improve proteomic discrimination of colon cancer patients from controls.

Materials and Methods

We conducted a case-control study nested within the U.S. military population under surveillance by the Armed Forces Health Surveillance Center (AFHSC). The AFHSC hosts a central repository of longitudinally collected medical data for the US armed forces and the Department of Defense Serum Repository (DoDSR) that collects blood samples from all members approximately every other year [17]. This study was approved by the Uniformed Services University of the Health Sciences and the Walter Reed National Military Medical Center Institutional Review Boards.

Study subjects

Cases consisted of men and women diagnosed with colon cancer between January 1, 2001 and December 31, 2009 who were between 17 and 79 years of age at diagnosis of colon cancer, active duty at time of diagnosis, and had at least one pre-diagnostic serum sample of ≥ 1mL. Controls were men and women who did not have colon cancer, were active duty in the reference year (corresponding to diagnosis year of matched case), had at least one serum sample ≥ 1mL, and were alive at the start of the study. They were matched to cases on age (within one year of birth), racial background (European-American, African-American, and Other), and time points at which serum samples were collected (within ± 60 days of requested case sample). All controls were randomly selected from the AFHSC database.

At the first step of study subject enrollment, AFHSC identified and provided a list of all eligible cases according to the 9th Revision International Classification of Diseases (ICD-9) coding system. Before contacting eligible cases, we verified survival status through the Defense Manpower Data Center (DMDC), which manages and updates all personnel information, as well as CDC's National Death Index and the Social Security Administration. An informative letter describing the study purpose and procedures, informed consent document, and questionnaire packet were mailed to all surviving cases. For each case who participated, five potential controls were identified using procedures similar to cases. If the person did not respond or did not want to participate, the next control was contacted until successful enrollment.

After excluding deceased cases and individuals without confirmed colon cancer, there were 431 eligible surviving cases, out of whom 234 participated and completed a questionnaire (54%). We then contacted 560 controls who were matched to the surviving case participants. Out of the contacted controls, 222 participated (40%) and 215 had a matched case. The remaining 19 cases that did not have a matched control were subsequently matched to randomly selected controls from AFHSC and did not have questionnaire data. Additionally, 163 randomly selected controls were matched to the 163 deceased colon cancer cases without questionnaire data. Overall, there were 397 cases and 397 matched controls included in this study.

Data collection

Epidemiologic survey data—For surviving cases and matched controls who participated in the study, self-administered questionnaires were utilized to collect epidemiologic data, which included information on demographic variables, personal habits, lifestyle characteristics, medical history, personal feelings and general health, family histories of cancer and colorectal polyps, reproductive and contraceptive histories (for women only), and dietary intake. When applicable, questions on exposures at the time before first blood draw for military duty (time 1) and in the year before cancer diagnosis for cases or reference date for controls (time 2) were asked. The reference date corresponds to the month in which the matched case was diagnosed. The mean length of time between the two time periods was 10.4 years, which did not significantly differ between cases and controls.

Medical records and surveillance data—AFHSC surveillance data was collected for all case-control matched pairs who were included in serum analysis. We requested data on demographics, military occupation specialty, casualty events, diagnosed medical conditions, medical procedures performed, and vaccinations.

Cancer registry data—The Automated Central Tumor Registry (ACTUR) of the Armed Forces collects data from medical records of cancer patients diagnosed or treated at military treatment facilities. The data contain information on demographic and tumor characteristics, diagnostic procedures, treatment, and vital status. Quality assurance guidelines established by the North America Association of Central Cancer Registries for state registries were used to edit the data. Out of 397 (234 surviving and 163 deceased cases) colon cancer patients identified using medical records, 264 were confirmed with colon cancer in ACTUR data (66.2%). The absence of ACTUR data is most likely due to diagnosis or treatment not at military treatment facilities, incomplete ACTUR data, or case participants' misconception of his/her rectal cancer as colon cancer.

Serum samples—Serum samples were obtained from the DoDSR. Samples were collected from active-duty and reserve personnel upon entry to the US military on average every one to two years starting in 1988 for HIV testing. We requested a maximum of 4 serum samples from each subject (n=794) with an overall serum sample count of n=2,752. According to standard procedures, blood specimens are drawn and shipped immediately at room temperature overnight to laboratories for HIV testing. After the samples are tested, they are frozen and shipped to DoDSR and stored in −30°C walk-in freezers. All procedures are generally completed within 48 hours and never thawed after they are frozen. Once retrieved, they were shipped on dry ice to our laboratory and immediately frozen at −80°C.

Lab measurements—MALDI-TOF MS profiling of weak cation exchange bead captured serum proteins was performed using a MB-WCX 100 Protein Profiling Kit (Bruker Daltonics, Billerica, MA). Serum proteins and peptides were eluted from the beads, concentrated and spotted 1:2 with CHCA matrix (α-Cyano-4-hydroxycinnamic acid; 10 mg/ml in 50% acetonitrile and 2.5% trifluoroacetic acid) for MALDI-TOF profiling as per manufacture instructions and previous studies [18, 19]. Ions in the mass range of 2000 to 15,000 *m/z* were collected using an AutoFlex III MALDI-TOF/TOF mass spectrometer and spectra for each sample were visualized and exported using DataAnalysis 4.0 software from Bruker Daltonics, Billerica, MA. Mass spectra were then further processed and exported using Progenesis MALDI software (Nonlinear Dynamics, Durham, NC). Total ion chromatogram (TIC) normalized peak heights were used for further analysis. Proteomic data were then separated into a 2/3 training set and 1/3 independent test set. The training set was used to create models (described below) that were qualified using the independent test set. This partitioning was performed randomly while maintaining matched pairs. Serum levels (ng/ml) of insulin-like growth factor-1 (IGF-1) and its binding protein, IGFBP-3 were determined using the commercial R&D Quantikine ELISA kit.

Data processing and analysis

We used a two-step procedure to identify discriminatory peaks using random forest (RF; 150 trees, min leaf = 1) feature selection and model optimization [20–22]. The ensemble nature of a RF as well as the internal out-of-bag validation results in excellent performance on external data across many data domains. A RF can utilize categorical data, which is important in later steps when epidemiologic and medical claims data were included in classifier development.

During feature selection, we harnessed two properties of RFs: (i) out-of-bag (oob) error, and (ii) variable importance. Similar to the proposed method by [23], we developed a two-step process to determine the most important variables in the training set. The first step was to construct a random forest using all variables (m) consisting of 150 trees and then extract the oobVI information, which was used to rank the variables in order of decreasing importance. The second step was to grow the random forest with the first k variables for $k = 1$ to m, calculating the oobError for all m forests. The model with the smallest error rate was selected that included k variables.

Model optimization was based on forward sequential feature selection with 10-fold cross-validation to generate models with minimum misclassification error and performance was qualified with the independent test set. An error rate <0.35 was considered acceptable for a model. All machine learning was performed using Matlab (v8.3.0.532; Mathworks, Natick, MA).

Based on the peaks selected during feature selection and model optimization using only the training set, models were next generated by including additional epidemiologic variables (demographics, lifestyle, medical history, as well as IGF-1 and IGFBP-3 biomarker concentrations), while keeping the test set held-out. We first included epidemiological variables from the Serum Repository (SR) only, which consisted of demographics, diagnosed medical conditions, medical procedures performed, and vaccinations. For the case-control pairs who completed the questionnaire, additional variables from the questionnaire were also used, which included race/ethnicity, highest education, service branch, marital status, religion, household income, number of people living in household, personal history of colon and rectal polyps, family history of cancer, physical activity, history of alcohol, history of smoking, deployment status, BMI (classified according to World Health Organization 2014), other health conditions, and treatments related to colon cancer. Since questionnaire data was collected for two time periods, variables at corresponding time points were respectively used in the models. As a result, there were three sets of epidemiologic and surveillance variable sets used for modeling: 1) SR only, 2) SR plus questionnaire information related to one year before diagnosis or reference date (SR+Q (diag)), and 3) SR plus questionnaire information related to the time before first blood draw (SR+Q (bfbd)). In all models, peaks selected during feature selection and model optimization were used as a base and additional epidemiologic and surveillance variables were added sequentially. Using this approach, sequential feature selection was used with 10-fold cross-validation to generate models with minimum misclassification error and performance was qualified with the independent test set.

Peptide Identification

Next we attempted to identify peptides of interest using high-resolution tandem mass spectrometry (Thermo Orbitrap Elite). In order to identify these peptides, we first selected serum samples with higher than average intensities of discriminatory peaks used in modeling and then generated a single case-control pooled sample. The pooled sample was injected into a C18 reversed phase nano-LC on a 75-μM × 15-cm capillary column packed in-house (YMC ODS-AQ 120Å S5; Waters Corporation, Beverly, MA) using a 120 minute linear gradient from 5% to 50% buffer B (97.8% acetonitrile, 2% HPLC H2O, .2% formic acid). Chromatography was carried out at 200nl/min using an Ultimate 3000 nanoflow system (Dionex, Sunnyvale, CA) directly interfaced to an Orbitrap Elite tandem mass spectrometer (Thermo Fisher Scientific, San Jose, CA). Data dependent acquisition was set to select the top 10 ions for MS/MS CID fragmentation at normalized collision energy of 35% along with dynamic exclusion, enabled with a repeat count of three, a repeat duration of 30 s, and exclusion duration of 180. Acquired data files were converted to peak lists using ProteomeDiscover (v1.4) and these resulting Mascot generic format files were searched using the Mascot algorithm (v2.4.1). The database used the UniProtKB 2015_05 release comprised of the SwissProt, SwissProt varsplic, and TrEMBL releases (167,678 entries); taxonomy *Homo sapiens* was specified. The decoy search parameter was specified within Mascot to calculate local FDRs with the following variable modifications: no enzyme, N-term pyroGlu, Met oxidation, and Asn/Gln deamidation with a precursor tolerance of 20 ppm and fragment ion tolerance of 0.8 Da. There were 107 protein families identified in the control pool at 4.65% local FDR and 102 identified protein families in the case pool at 5.17% local FDR. Using the N-term acetyl variable modification, there were 96 protein families identified in the control pool at 0.00% local FDR and 104 identified protein families in the case pool at 2.13% local FDR. The results were mined to find confidently assigned peptide sequences that aligned with the peaks of interest identified during the MALDI-TOF profiling.

Results

Table 1 shows the distribution of demographic characteristics by case-control status among individuals with serum samples (SR) and those with serum samples who also completed the questionnaire (SR+Q). For both the SR and SR+Q groups, cases and controls were similar in the distribution of matching variables (age, gender, race, and number of serum samples). Among participants in the SR+Q group, cases were more likely to be in the Army and Air Force and less likely to be in the Navy and Marines compared to controls. The proportion of protestant and subjects of unknown religion seemed higher in cases than controls while the proportion of those with no religion and Jewish religion tended to be lower among cases. Additionally, cases were more likely to have a household income of $45,000–$59,999, but less likely to have $60,000 or more than controls. The two groups were similar in ethnicity and highest education received.

Table 2 presents the *m/z* values of proteomic peaks identified from feature selection and model optimization for each time interval. After optimization, the only error rate lower than the threshold level of 0.35 was observed for one year prior to colon cancer diagnosis. The peaks identified had *m/z* values of 3119.32, 2886.67, 2939.23, and 5078.81 and a sensitivity of 69% and specificity of 67%.

Epidemiologic factors were then assessed in addition to final identified peaks from time interval 1 (Table 3). For the SR group, all factors selected into the model were vaccination related. However, sensitivity and specificity remained similar to the model without these factors. When questionnaire information at the collection of the first blood sample were considered in addition to factors from medical records, histories of other cancers and tonsillectomy were retained in the final model, which improved the specificity from 67% to 76%. When questionnaire data at colon cancer diagnosis were assessed, smoking at diagnosis was found to be significant and increased the sensitivity from 69% to 76%.

Table 4 shows the identification of peptides using serum samples with higher intensities of the identified peaks. Three peaks (2886.670, 2939.235, and 3119.317 *m/z*) out of the four were identified; two (2886.67 and 3119.32 *m/z*) were histone acetyltransferases and one (2939.24 *m/z*) was a transporting ATPase subunit.

Discussion

Using serial pre-diagnostic serum samples and collected epidemiological information, this study identified certain proteomic profiles that have the potential to discriminate colon cancer patients from randomly-selected matched controls within the military. Proteomic peaks (2886.67, 2939.24, 3119.32, and 5078.81 *m/z*) were identified one year prior to colon cancer diagnosis with a sensitivity of 69% and a specificity of 67%. When epidemiologic information was also considered, factors at colon cancer diagnosis reduced the error rate and increased the sensitivity while factors at first blood sample collection were found to increase specificity.

Of the four proteomic peaks, three were identified as peptides derived from KAT6A, ATP1A4, and EP300. Sodium/potassium-transporting ATPase subunit alpha-4 (ATP1A4) is encoded by the ATP1A4 gene and acts as a catalyst in the hydrolysis of ATP in moving sodium and potassium ions across the plasma membrane [24]. Additionally, histone acetyltransferase KAT6A encoded by KAT6A gene is a component of the MOZ/MORF complex and acetylates residues on histone H3 and H4 [24]. The KAT6A gene has been found to play a role in breast carcinogenesis [25] while the MOZ complex has been implicated in cellular senescence inhibition in mouse embryonic fibroblasts [26]. The EP300 protein (E1A binding protein P300) is encoded by tumor suppressor gene EP300 and is also involved in histone acetyltransferase activity and transcription regulation through chromatin remodeling [24]. One study found that frameshift mutations in EP300 occurred in gastric and colorectal cancers where EP300 expression was lost in 12–24% of patients [27].

Identification of proteomic markers from samples collected in the year prior to colon cancer diagnosis, but not in those collected earlier, is theoretically reasonable. Conceptually, the closer the collection date of a pre-diagnostic sample to colon cancer diagnosis, the higher the likelihood of identifying a cancer-related protein. To the best of our knowledge, there have been no proteomic studies on colon cancer detection using serial pre-diagnostic serum samples. In a study on ovarian cancer [28], pre-diagnostic serum samples analyzed using the CA 125 immunoassay and SELDI-TOF-MS were collected from ovarian cancer patients and age-matched controls. The study showed that CA125 was elevated in 40 of 65 (61.5%) serum samples collected less than one year prior to cancer diagnosis, but in only 1 of 50 (2%) samples collected more than one year prior to cancer diagnosis. Although this study was not on colon cancer, results suggest that protein biomarkers may be more detectable in the time period closer to cancer diagnosis.

As the only proteomic study on colon cancer utilizing prediagnostic samples with both biologic and epidemiologic data, this study had several strengths. First, the study is unique in its utilization of serial pre-diagnostic samples to identify the earliest time at which discriminatory proteomic profiles might occur and whether the profiles may vary by time prior to colon cancer diagnosis. Our results suggest the significance of measuring serial samples, providing a basis for further research on early detection and prediction of colon cancer. Second, this study recruited an appropriate and comparable control group. In many previous studies, controls were based on hospital-based convenience samples, and not defined clearly. Thus, they might not come from the same target population as cases [29]. As a result, biomarkers related to non-cancer conditions might be more prevalent among cancer patients and ascertained as cancer biomarkers. Additionally, the ambiguity on whether comparison groups were comparable might have limited the accuracy and reliability of results. In our study, the controls were from the same target population (active-duty members) and selected randomly based on certain matching criteria. Therefore, they were theoretically comparable to cases, except on colon cancer status. Third, in previous studies, sample collection, storage, storage time, and processing might differ between controls and cases, which might influence study results [15, 28, 30]. In our study, prediagnostic serum samples were collected and processed with standardized procedures without referring to case-control status by the DoDSR. We also matched cases and controls by time at sample collection (therefore, storage). This can preclude the possibility that identified biomarkers result from the differences between cancer patients and controls in sample collection, storage, and processing. Fourth, our study included various demographic, medical, and epidemiologic data, which are not often collected and used in analysis. Since protein levels are likely affected by various factors [31], proteomic detection/prediction of cancer should include additional demographic and epidemiological information. While the effects of these factors are not clear at the present time, our findings suggest the significance of considering epidemiological factors for improving early detection or prediction of colon cancer with identified and validated proteomic profiles.

While our study has several strengths, there are also limitations. First, we cannot exclude the possibility of protein degradation. While serum samples were frozen within 48 hours of sample collection, they were shipped at room temperature and stored in walk-in freezers at −30°C rather than −80°C. Although some studies showed that 1) the levels of high abundant proteins to which small proteins (e.g. cancer proteins) are bound and thus protected [32–34] remained similar after exposure to room temperature for 48 hours or several months [35–38] and 2) low abundant proteins were relatively stable after being stored at −20°C to −40°C for many years [39–41], protein degradation was likely. We applied quality control procedures to exclude samples with poor quality in order to reduce the potential effects of degradation. However, we cannot exclude the possibility of degradation, particularly because we found peaks one year prior to diagnosis, but not in other earlier time intervals. Nevertheless, the identified proteomic differences between cases and controls show our capability to detect proteomic changes. Second, some proteomic biomarkers with extremely low abundance might not be detected using our measurement methods. This may be particularly true for pre-diagnostic samples collected many years prior to cancer diagnosis when cancer cells are scarce. Since the analyte detection sensitivity for conventional mass spectrometry analysis is typically higher than 50pg/mL, biomarkers with concentrations below this limit cannot be detected [29]. Third, processing and analysis of high-throughput data are still in its infancy with many technical challenges [42–44]. Various pre-analytical data processing, algorithms, and model-building approaches can lead to low reproducibility of identified biomarkers. Finally, available analytical tools such as machine learning do not have built-in capabilities to analyze more complex data such as matched samples, samples from different time periods, and multiple samples within the same time period from the same person; all of which, occurred in our study.

Using the unique resource of serial pre-diagnostic serum samples stored at the DoD Serum Repository, this nested case-control study identified potential proteomic biomarkers for early detection and prediction of colon cancer, although the sensitivities and specificities were not particularly high. Four peptide masses were found to be useful in discriminating colon cancer and since we have assigned putative identifications to all but one mass, future assay development is feasible using either SRM mass spectrometry based assays or immunosorbent assays. This investigation also showed the possible significance of including epidemiological factors in discriminant and prediction models. The processing and analysis of high-throughput data, especially those with multiple samples from not only different time periods, but also within the same period, are challenging, which warrants continuing and un-remitted efforts in further analysis.

Acknowledgments

Source of Funding: This study was supported by grant # 5R01CA118707 from the National Cancer Institute. K. Zhu was the recipient of this grant.

The project team thanks the following institutes for their contributions to or support for the project: Uniformed Services University of the Health Services, Armed Forces Health Surveillance Center, DoD Automated Tumor Registry, Henry F. Jackson Foundation, Defense Manpower Data Center, Former Armed Forces Institute of Pathology, Walter Reed National Military Medical Center, and Tricare Management Activity. We also thank Dr. I-Jen Pan, Ms. Kanchana Perera, and Dr. Maciek Sasinowski for their contributions to the early stage of the project.

References

1. Wu W, Hu W, Kavanagh JJ. Proteomics in cancer research. International journal of gynecological cancer : official journal of the International Gynecological Cancer Society. 2002; 12:409–423. [PubMed: 12366655]
2. Jin X, Lin M, Zhang H, Han Y, He Y, Zhang Q, et al. Serum biomarkers of colorectal cancer with AU and NP20 chips including a diagnosis model. Hepato-gastroenterology. 2012; 59:124–129. [PubMed: 22260829]
3. Fan NJ, Kang R, Ge XY, Li M, Liu Y, Chen HM, et al. Identification alpha-2-HS-glycoprotein precursor and tubulin beta chain as serology diagnosis biomarker of colorectal cancer. Diagnostic pathology. 2014; 9:53. [PubMed: 24618180]
4. Liu C, Pan C, Shen J, Wang H, Yong L. MALDI-TOF MS combined with magnetic beads for detecting serum protein biomarkers and establishment of boosting decision tree model for diagnosis of colorectal cancer. International journal of medical sciences. 2011; 8:39–47. [PubMed: 21234268]
5. Zhu D, Wang J, Ren L, Li Y, Xu B, Wei Y, et al. Serum proteomic profiling for the early diagnosis of colorectal cancer. Journal of cellular biochemistry. 2013; 114:448–455. [PubMed: 22961748]
6. Albrethsen J, Bogebo R, Gammeltoft S, Olsen J, Winther B, Raskov H. Upregulated expression of human neutrophil peptides 1, 2 and 3 (HNP 1–3) in colon cancer serum and tumours: a biomarker study. BMC cancer. 2005; 5:8. [PubMed: 15656915]
7. Murakoshi Y, Honda K, Sasazuki S, Ono M, Negishi A, Matsubara J, et al. Plasma biomarker discovery and validation for colorectal cancer by quantitative shotgun mass spectrometry and protein microarray. Cancer science. 2011; 102:630–638. [PubMed: 21199170]
8. Huijbers A, Mesker WE, Mertens BJ, Bladergroen MR, Deelder AM, van der Burgt YE, et al. Case-controlled identification of colorectal cancer based on proteomic profiles and the potential for screening. Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland. 2014; 16:907–913. [PubMed: 25243779]
9. Shin J, Kim HJ, Kim G, Song M, Woo SJ, Lee ST, et al. Discovery of melanotransferrin as a serological marker of colorectal cancer by secretome analysis and quantitative proteomics. Journal of proteome research. 2014; 13:4919–4931. [PubMed: 25216327]
10. Wang Y, Song G, Wang Y, Qiu L, Qin X, Liu H, et al. Elevated serum levels of circulating immunoinflammation-related protein complexes are associated with cancer. Journal of proteome research. 2014; 13:710–719. [PubMed: 24295561]
11. Alvarez-Chaver P, Otero-Estevez O, Paez de la Cadena M, Rodriguez-Berrocal FJ, Martinez-Zorzano VS. Proteomics for discovery of candidate colorectal cancer biomarkers. World journal of gastroenterology. 2014; 20:3804–3824. [PubMed: 24744574]
12. Babel I, Barderas R, Diaz-Uriarte R, Martinez-Torrecuadrada JL, Sanchez-Carbayo M, Casal JI. Identification of tumor-associated autoantigens for the diagnosis of colorectal cancer in serum using high density protein microarrays. Molecular & cellular proteomics : MCP. 2009; 8:2382–2395. [PubMed: 19638618]
13. Wang Q, Shen J, Li ZF, Jie JZ, Wang WY, Wang J, et al. Limitations in SELDI-TOF MS whole serum proteomic profiling with IMAC surface to specifically detect colorectal cancer. BMC cancer. 2009; 9:287. [PubMed: 19689818]
14. Check E. Proteomics and cancer: running before we can walk? Nature. 2004; 429:496–497. [PubMed: 15175721]

15. Diamandis EP. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. Journal of the National Cancer Institute. 2004; 96:353–356. [PubMed: 14996856]

16. Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC bioinformatics. 2003; 4:24. [PubMed: 12795817]

17. Rubertone MV, Brundage JF. The Defense Medical Surveillance System and the Department of Defense serum repository: glimpses of the future of public health surveillance. American journal of public health. 2002; 92:1900–1904. [PubMed: 12453804]

18. Karbassi ID, Nyalwidhe JO, Wilkins CE, Cazares LH, Lance RS, Semmes OJ, et al. Proteomic expression profiling and identification of serum proteins using immobilized trypsin beads with MALDI-TOF/TOF. Journal of proteome research. 2009; 8:4182–4192. [PubMed: 19603828]

19. Schaub NP, Jones KJ, Nyalwidhe JO, Cazares LH, Karbassi ID, Semmes OJ, et al. Serum proteomic biomarker discovery reflective of stage and obesity in breast cancer patients. Journal of the American College of Surgeons. 2009; 208:970–978. discussion 978–980. [PubMed: 19476873]

20. Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC bioinformatics. 2006; 7:3. [PubMed: 16398926]

21. Cima I, Schiess R, Wild P, Kaelin M, Schuffler P, Lange V, et al. Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. Proc Natl Acad Sci U S A. 2011; 108:3342–3347. [PubMed: 21300890]

22. Chen T, Cao Y, Zhang Y, Liu J, Bao Y, Wang C, et al. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. Evidence-based complementary and alternative medicine : eCAM. 2013; 2013:298183. [PubMed: 23573122]

23. Genuer R, Poggi J, Tuleau-Malot C. Variable selection using random forests. Pattern Recognition letters, Elsevier. 2010; 31:2225–2236.

24. UniProt C. UniProt: a hub for protein information. Nucleic acids research. 2015; 43(Database issue):D204–D212. [PubMed: 25348405]

25. Turner-Ivey B, Guest ST, Irish JC, Kappler CS, Garrett-Mayer E, Wilson RC, et al. KAT6A, a chromatin modifier from the 8p11-p12 amplicon is a candidate oncogene in luminal breast cancer. Neoplasia. 2014; 16:644–655. [PubMed: 25220592]

26. Sheikh BN, Phipson B, El-Saafin F, Vanyai HK, Downer NL, Bird MJ, et al. MOZ (MYST3, KAT6A) inhibits senescence via the INK4A-ARF pathway. Oncogene. 2015

27. Kim MS, Lee SH, Yoo NJ, Lee SH. Frameshift mutations of tumor suppressor gene EP300 in gastric and colorectal cancers with high microsatellite instability. Human pathology. 2013; 44:2064–2070. [PubMed: 23759652]

28. Moore LE, Pfeiffer RM, Zhang Z, Lu KH, Fung ET, Bast RC Jr. Proteomic biomarkers in combination with CA 125 for detection of epithelial ovarian cancer using prediagnostic serum samples from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. Cancer. 2012; 118:91–100. [PubMed: 21717433]

29. Liotta LA, Petricoin EF 3rd. Omics and cancer biomarkers: link to the biological truth or bear the consequences. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2012; 21:1229–1235.

30. Sorace J, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC bioinformatics. 2003; 4:24. [PubMed: 12795817]

31. Chen ST, Pan TL, Tsai YC, Huang CM. Proteomics reveals protein profile changes in doxorubicin--treated MCF-7 human breast cancer cells. Cancer Lett. 2002; 181:95–107. [PubMed: 12430184]

32. Liotta LA, Ferrari M, Petricoin E. Clinical proteomics: written in blood. Nature. 2003; 425:905. [PubMed: 14586448]

33. Mehta AI, Ross S, Lowenthal MS, Fusaro V, Fishman DA, Petricoin EF 3rd, et al. Biomarker amplification by serum carrier protein binding. Dis Markers. 2003; 19:1–10. [PubMed: 14757941]

34. Zolotarjova N, Martosella J, Nicol G, Bailey J, Boyes BE, Barrett WC. Differences among techniques for high-abundant protein depletion. Proteomics. 2005; 5:3304–3313. [PubMed: 16052628]

35. Donnelly JG, Soldin SJ, Nealon DA, Hicks JM. Stability of twenty-five analytes in human serum at 22 degrees C, 4 degrees C, and −20 degrees C. Pediatr Pathol Lab Med. 1995; 15:869–874. [PubMed: 8705197]

36. Clark S, Youngman LD, Palmer A, Parish S, Peto R, Collins R. Stability of plasma analytes after delayed separation of whole blood: implications for epidemiological studies. Int J Epidemiol. 2003; 32:125–130. [PubMed: 12690023]

37. Kubasik NP, Ricotta M, Hunter T, Sine HE. Effect of duration and temperature of storage on serum analyte stability: examination of 14 selected radioimmunoassay procedures. Clin Chem. 1982; 28:164–165. [PubMed: 7034999]

38. Rai AJ, Gelfand CA, Haywood BC, Warunek DJ, Yi J, Schuchard MD, et al. HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. Proteomics. 2005; 5:3262–3277. [PubMed: 16052621]

39. Whittemore AS, Cirillo PM, Feldman D, Cohn BA. Prostate specific antigen levels in young adulthood predict prostate cancer risk: results from a cohort of Black and White Americans. J Urol. 2005; 174:872–876. discussion 876. [PubMed: 16093978]

40. Richardson B, Peck J, Wormuth J. Mean arterial pressure, pregnancy induced hypertension, and preeclampsia. Evaluation as independent risk factors and as surrogates for high maternal serum alpha-protein in estimating breast cancer risk. Ann Epidemiol. 2000; 10:464.

41. Richardson BE, Hulka BS, Peck JL, Hughes CL, van den Berg BJ, Christianson RE, et al. Levels of maternal serum alpha-fetoprotein (AFP) in pregnant women and subsequent breast cancer risk. Am J Epidemiol. 1998; 148:719–727. [PubMed: 9786226]

42. Martin SF, Falkenberg H, Dyrlund TF, Khoudoli GA, Mageean CJ, Linding R. PROTEINCHALLENGE: crowd sourcing in proteomics analysis and software development. Journal of proteomics. 2013; 88:41–46. [PubMed: 23220569]

43. Hoopmann MR, Moritz RL. Current algorithmic solutions for peptide-based proteomics data generation and identification. Current opinion in biotechnology. 2013; 24:31–38. [PubMed: 23142544]

44. Cappadona S, Baker PR, Cutillas PR, Heck AJ, van Breukelen B. Current challenges in software solutions for mass spectrometry-based quantitative proteomics. Amino acids. 2012; 43:1087–1108. [PubMed: 22821268]

Table 1

Selected characteristics of all patients diagnosed with colon cancer, 2002–2009, and their matched controls

	Serum Repository N=794		**Serum Repository + Questionnaire N=430**	
Characteristics	**Cases n (%)**	**Controls n (%)**	**Cases n (%)**	**Controls n (%)**
Age[a] (mean years ± stdev)	39.47 (8.69)	39.54 (8.70)	41.12 (8.07)	41.22 (8.09)
Gender				
Male	331 (83.4)	331 (83.4)	177 (82.3)	177 (82.3)
Female	66 (16.6)	66 (16.6)	38 (17.7)	38 (17.7)
Race				
White	276 (69.5)	276 (69.5)	162 (75.4)	162 (75.4)
Black	84 (21.2)	84 (21.2)	33 (15.4)	33 (15.4)
Other/Unknown	37 (9.3)	37 (9.3)	20 (9.3)	20 (9.3)
# of serum samples				
1	23 (5.8)	23 (5.8)	13 (6.1)	13 (6.1)
2	44 (11.1)	44 (11.1)	20 (9.3)	20 (9.3)
3	54 (13.6)	55 (13.9)	29 (13.5)	29 (13.5)
4	276 (69.5)	275 (69.3)	153 (71.2)	153 (71.2)
Vital Status				
Alive	227 (57.1)	397 (100)	208 (96.8)	215 (100)
Deceased	170 (42.9)	0	7 (3.3)	0
Ethnicity				
Hispanic	-	-	11 (5.1)	7 (3.3)
Not Hispanic	-	-	204 (94.9)	208 (96.7)
Unknown	-	-	0	0
Highest Education[a]				
High School or Less	-	-	19 (8.8)	22 (10.2)
Some College, Technical, Vocational	-	-	82 (38.2)	73 (34.0
College	-	-	47 (21.9)	49 (22.8)
Graduate, Professional	-	-	67 (31.2)	70 (32.6)
Unknown	-	-	0	1 (0.5)
Service Branch[a]				
Army	-	-	68 (31.6)	49 (22.8)
Navy	-	-	64 (29.8)	108 (50.2)
Air Force	-	-	55 (25.6)	24 (11.2)
Marines	-	-	16 (7.4)	31 (14.4)
Other/Unknown	-	-	12 (5.6)	3 (1.4)
Religion				
None	-	-	31 (14.4)	40 (18.6)
Protestant	-	-	122 (56.7)	107 (49.8)
Jewish	-	-	2 (0.9)	4 (1.9)

	Serum Repository N=794		**Serum Repository + Questionnaire N=430**	
Characteristics	**Cases n (%)**	**Controls n (%)**	**Cases n (%)**	**Controls n (%)**
Catholic	-	-	51 (23.7)	57 (26.5)
Other	-	-	5 (2.3)	5 (2.3)
Unknown	-	-	4 (1.8)	2 (0.9)
Household Income[a]				
<$29,999	-	-	16 (7.4)	12 (5.6)
$30,000 to $44,999	-	-	36 (16.7)	38 (17.7)
$45,000 to $59,999	-	-	60 (27.9)	45 (20.9)
>$60,000	-	-	102 (47.4)	119 (55.4)
Unknown	-	-	1 (0.5)	1 (0.5)
People in Household[a]				
1	-	-	41 (19.1)	43 (20.0)
2	-	-	44 (20.5)	43 (20.0)
3	-	-	54 (25.1)	45 (20.9)
4	-	-	48 (22.3)	52 (24.2)
>4	-	-	27 (12.6)	32 (14.9)
Unknown	-	-	1 (0.5)	0

Stdev= standard deviation

[a] At diagnosis for cases and corresponding reference date for controls

Table 2

Final proteomic peaks identified after random forest feature selection and optimization for each time period, separately

	Feature Selection	**Optimization**		
Time Interval[a] (yrs.)	**protein peaks *(m/z)***	**protein peaks *(m/z)***	**Sensitivity/Specificity**	**Error rate**
1	3288.69, 2818.53, 3119.32, 2886.67, 2835.08, 3257.54, 2939.23, 5078.81, 4227.07, 5621.97, 3579.74	3119.32, 2886.67, 2939.23, 5078.81	0.691 / 0.673	0.318
2	4247.51, 1999.89, 2754.29, 6186.56, 5427.29	4247.51, 2754.29, 6186.56, 5427.29	0.652 / 0.370	0.489
3	6004.53, 2285.10	6004.53, 2285.10	0.452 / 0.548	0.500
4–5	3682.93, 2522.61, 2506.06, 2909.06, 3523.29, 4247.51, 3642.04, 2080.68, 2133.24	2909.06, 4247.51, 2080.68	0.308 / 0.590	0.551
6–8	3650.80, 2818.53, 4185.21, 4339.98, 3443.46, 3086.22, 2273.42, 5671.62, 4518.12, 2313.33, 2990.83, 2583.94, 3097.90, 7186.26, 5471.09, 2430.14, 3257.54, 3241.97	2990.83	0.500 / 0.344	0.578
8+	4128.75, 2293.86, 3147.55, 4339.98, 5804.00, 4185.21, 7562.00, 2835.08, 4247.51, 6785.21, 3916.55, 5015.53, 6186.56	3147.55, 4185.21, 7562.00, 6186.56	0.612 / 0.510	0.439

Note: Unique samples from matched case-control pairs with case confirmation in ACTUR were used

[a] Time before cancer diagnosis (cases) or reference date (controls)

Table 3

Model optimization with additional epidemiologic information from the Serum Repository (SR) or questionnaire data for selected proteomic peaks in the year before cancer diagnosis

Model	Error rate	Sensitivity/Specificity	Final variables[a]
SR	0.33	0.691 / 0.655	3119.316895
			2886.669922
			2939.234619
			5078.806641
			Adenovirus vaccine
			Anthrax vaccine
			Influenza H1N1 vaccine
SR + Q (bfbd)	0.28	0.680 / 0.760	3119.316895
			2886.669922
			2939.234619
			5078.806641
			Other cancer
			Tonsillectomy
SR + Q (diag)	0.28	0.760 / 0.680	3119.316895
			2886.669922
			2939.234619
			5078.806641
			Smoking at diagnosis

SR= Serum Repository; Q=questionnaire; bfbd=before first blood draw; diag=before diagnosis

[a] Peaks included in each model are from time period 1 in Table 2

Table 4

Identification of peptides from model feature selection and model optimization

Peak (*m/z*)	UniProt ID	Description	M_{calc}	Δ mass (Da)	Sequence
2886.670	Q92794	Histone acetyltransferase (KAT6A)	2886.386	0.72	PSAVAMQAGPRALAVQRGMNMGVNLMPT + 3 Deamidated (NQ); Oxidation (M)
2939.235	Q13733	Sodium/potassium-transporting ATPase subunit alpha-4 (ATP1A4)	2937.5243	0.71	LRTELRPGETLNVNFLLRMDRAHE + Acetyl (N-term); Oxidation (M)
3119.317	Q7Z6C1	Histone acetyltransferase p300 (EP300)	3118.3464	0.03	QQGSPQMGGQTGLRGPQPLKMGMMNNPNP + 4 Deamidated (NQ); 4 Oxidation (M)
5078.807	-	-	-	-	-