Identification of Differentially Abundant Plasma Proteins in Cancer Using the Corra Label-Free LC-MS Platform

Simon Letarte[1], Mi-Youn Brusniak[1], Alexander Schmidt[2], Hollis Lau[1], Paul Shannon[1], David Campbell[1], Lukas Mueller[2], Olga Vitek[3], James Eddes[1], Karen Spratt[4], Christopher Kemp[4], Julian Watts[1], Hui Zhang[1], Ruedi Aebersold[1,2]

1 Institute for Systems Biology, 1441 North 34th Street, Seattle, WA, 98103.
2 Institute for Molecular Systems Biology, ETH-Zurich, CH-8093 Zurich, Switzerland
3 Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 479
4 Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA, 98109

INSTITUTE FOR Systems Biology

Revolutionizing science. Enhancing life.

- *Plasma from 10 skin cancer and 10 non-cancer mice was enriched for glycopeptides.*
- *Glycopeptide-enriched plasma was analyzed by LC-MS on an ESI-TOF mass spectrometer.*
- *Data was imported into the Corra framework to perform peak picking and alignment.*
- *Aligned files contain m/z, retention time (RT) and charge state describing each feature, with the intensity measured in each sample.*
- *Corra generated a subset of differentially abundant features between the two disease states.*
- *Differentially abundant features were sequenced via an inclusion list on LTQ-FT mass spectrometer.*
- *About 80 peptides (50 proteins) were identified, mostly associated with host immune response, indicating that one population is seriously sick!*

INTRODUCTION

Cancer is responsible for about 25% of all deaths[1] in the United States. Finding early disease biomarkers would help diagnose it earlier and provide better treatment. The goal of this study is to demonstrate the feasibility of an LC-MS proteomics approach to biomarker discovery, by identifying peptides and proteins that are differentially abundant between cancer and non-cancer mice. Plasma is a rich body fluid, circulating throughout the whole body, picking up molecules from many organs. It could contain many biomarkers, molecules specific to certain diseases, if we can develop a methodology sensitive enough to get down to the low abundance molecules. Glycoproteins show a great potential for leached proteins from diseased tissues because a lot of them are located on the surface of cells and the sugar confers them increased solubility, which are more likely to be found in plasma. Glycan structures provide a handle that can be used for glycopeptide enrichment, and thus to simplify the mixture.

A paradigm shift is necessary to depart from the way traditional "shotgun" proteomics experiments are conducted. When automatic precursor selection is used to sequence peptides, there is a strong bias towards identification of high abundance peptides. The search for early disease biomarkers in plasma is a situation where there are more peptides than the sampling capacity of the instrument and where the interesting molecules are presumed to be of low intensity. In order to concentrate our efforts on the interesting features, our approach compares LC-MS profiles in order to find statistically significant changes in peptide abundance between different populations. Once those peptides are found, they are made into an inclusion list on a tandem mass spectrometer to be sequenced and identified.

METHODS

The mice used for this study was the ARF-deficient model[2, 3], where half of the population was exposed to a potentially oncogenic hyperproliferative topical agent that resulted in the appearance of skin cancers. Plasma from the mice was extracted and formerly N-glycosylated tryptic peptides were generated from glycoprotein isolated using hydrazide chemistry[4]. Peptides were separated on an 1100 Series HPLC (Agilent, Santa Clara, CA) equipped with a nanoflow pump, and a monolithic C18 10 cm column (Merck, Darmstadt, Germany). Mass analysis was performed on a micrOTOF electrospray time-of-flight mass spectrometer (Bruker, Billerica, MA) with a mass accuracy of 5 ppm or better, determined by use of an external standard to recalibrate the mass range between each run. The data was imported into the *Corra* framework, illustrated in Figure 1, to perform statistical analysis to reveal which peptides are differentially expressed between the two sample populations. The discriminatory peptides were made into an inclusion list for sequencing on an LTQ-FT (Thermo, San Jose, CA) mass spectrometer, with a mass window of 25 ppm to take into account the combined error of the 2 mass spectrometers. The LTQ-FT was hyphenated to an identical 1100 Series HPLC, but used a self-packed capillary column instead of the monolith used on the other system. Cytoscape (www.cytoscape.org) was used to plot GO molecular functions and protein physical interactions.

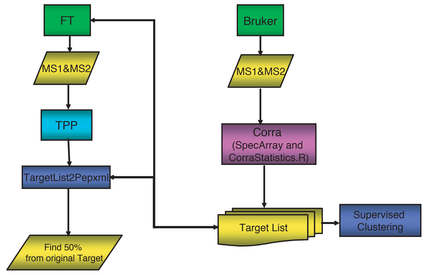

FIGURE 1: Corra framework workflow. MS1data acquired on the Bruker ESI-TOF is fed into the statistical processing module to produce a target list. High resolution FT-ICR instrument was used to produce the MS2 identifications.

RESULTS

LC-MS Reproducibility

Good mass accuracy on the Bruker ESI-TOF was achieved by injecting a small amount of glu-fibrinopeptide B between every sample injection and fragmenting it by raising the cone voltage to produce in-source CID. A Visual Basic macro would then automatically recalibrate the mass scale. Retention time (RT) reproducibility was achieved by careful control of the sample load, periodic maintenance and manual inspection of every sample acquired. RT was kept within 30 s over a 60 min. gradient for any given instrument. However, when using a combination of instruments, one should expect a larger variation. The correlation of the retention times of the peptides on the ESI-TOF and LTQ-FT was within 5 minutes, which was good considering that the systems used different analytical columns and the LTQ-FT did not use a pre-column. The manually validated results for one sample are illustrated in Figure 2.

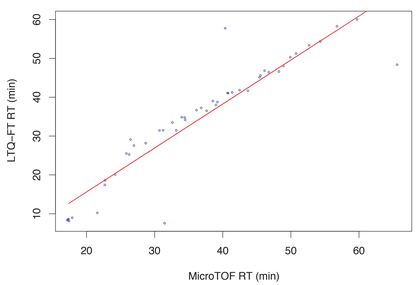

FIGURE 2: Correlation of the retention times between the MicrOTOF where the LC-MS patterns were generated and the LTQ-FT and the targetted analysis was performed. A number of early eluting, hydrophillic peptides are not matching because of the absence of a pre-column for sample trapping/desalting on the LTQ-FT.

Clustering and Inclusion Lists

Two different approaches were used within Corra to select the features of interest, using their clustering power or their differential abundance as a selection criteria. In the first method, principal components analysis (PCA) was used to construct an inclusion list representing the top 400 features with the more intense eigenvectors of the first two principal components. A second inclusion list, again of 400 features, was constructed using volcano plots. Both inclusion lists performed similarly, each identifying ~80 distinct peptides belonging to ~50 proteins with a ProteinProphet score above 0.9. About half of the features in each inclusion list were confidently identified (PeptideProphet score above 0.9).

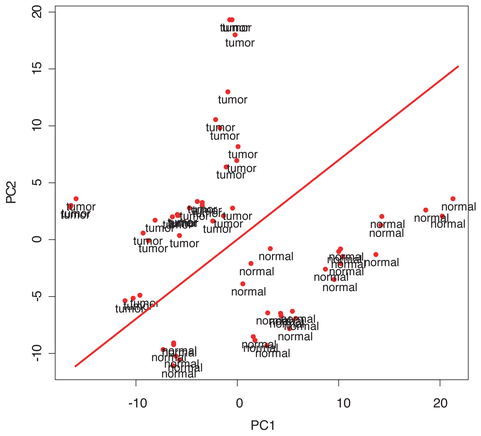

FIGURE 3: Principal components analysis (PCA) showing the clustering of the 60 samples (20 mice in triplicate) in two different clusters that can be linearly separated. Closer inspection reveals that the triplicate of the same individual also clusters together.

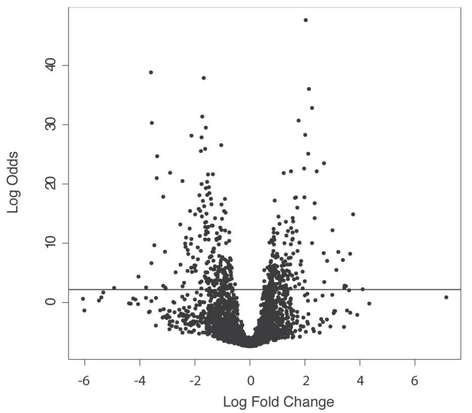

FIGURE 4: Volcano plot showing all the features (peptides not yet identified) that were present in every sample. Each point represents a feature; the higher on the plot (on the y axis) means a higher probability, more on the side (x axis) means greater differential abundance.

Identification and functions

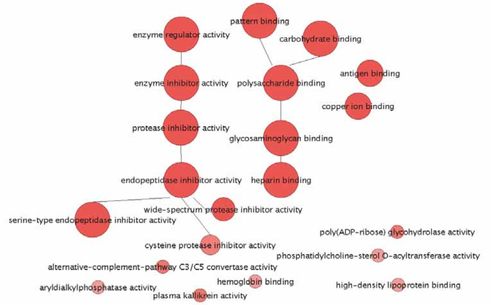

FIGURE 5: GO molecular function of the differentially abundant proteins identified in this experiment with a p-value higher than 0.5. The color intensity correlates with the p-value and the size of the node represents the number of proteins in each category.

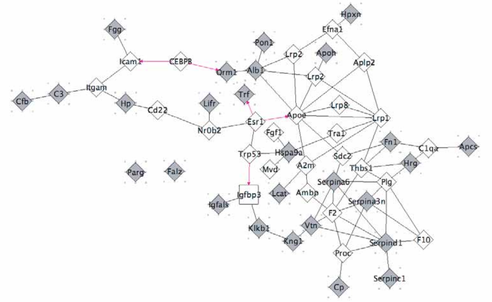

FIGURE 6: Protein physical interactions detected in the experiment. Protein-protein interactions are represented by black arrows; protein-DNA interactions by magenta arrows. Observed proteins are in gray.

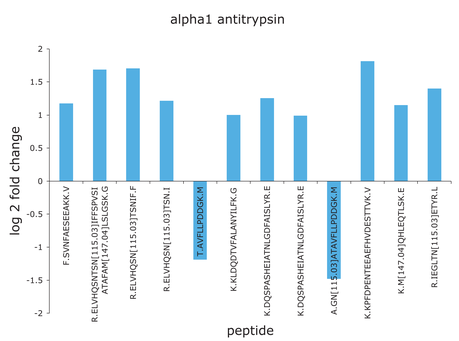

FIGURE 7: Example of peptide fold-change for a given protein in cancer vs non-cancer mice.

TABLE I : Example of differentially abundant proteins detected.

Name	Cancer-related annotation
Haptoglobin	
Serine protease inhibitor A3K	
Hemopexin	
Alpha-1-antitrypsin 1	
Fibrinogen, gamma polypeptide	
Corticosteroid-binding globulin	
Leukemia inhibitory factor receptor	LIF implicated in cancer incl. skin cancer
Ceruloplasmin	
Clusterin	
plasma paraoxonase	
Leucine-rich alpha-2-glycoprotein	Marker for Neutrophil differentiation
Alpha-1-antitrypsin 1-2	
Serine protease inhibitor A3F	
Immunoglobulin J chain	
Alpha-2-macroglobulin	
Receptor for viral-encoded semaphorin protein (plexin-C1; CD232)	Expressed on skin Langerhans cells and can trigger IL6/IL8 secretion
Insulin-like growth factor-binding protein complex acid labile chain	
Vitronectin	Cancer metastasis link via urokinase receptor and plasminogen activation
Apolipoprotein H	
Sal-like protein 3 (Spalt-like protein 3; SALL3)	Transcription factor - locus a site of aberrant DNA methylation
Serine protease inhibitor A3N	
Ig h-1a protein	
Beta-2-glycoprotein I	
Serine protease inhibitor A3M	
Ig gamma-2A chain C region	

CONCLUSIONS

This study demonstrates that it is possible to start from a complex plasma sample and perform label-free LC-MS quantitative proteomics to identify a list of differentially abundant proteins. *Corra* makes it possible to conduct such studies with different analytical instruments and to integrate the results to perform targeted peptide sequence analysis. The two methods to generate inclusion lists, PCA and volcano plots, had ~70 % overlap in terms of the peptides identified. This indicates that both methods are capable to identify the features different in the two populations.

It is possible to use *Corra* to measure fold-change in peptide abundance between different sample populations. However, to accurately deconvolute peptide abundance into protein abundance will require some more work due to shared peptides, peptide modifications and various other processes that cause certain peptides from a given protein to follow different trends. The proteins identified are obviously not all skin cancer biomarkers. Many are related to immune response, indicating that something is happening in the cancer subjects. One possible explanation would be that since the tumors were allowed to grow to very large sizes, it would generate a very strong response from the host. In order to detect early disease biomarkers, it would be interesting to repeat the experiment with smaller tumors.

ACKNOWLEDGEMENTS

This work was supported in part with federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract N01-HV-28179 (to R.A.) and with federal funds from the National Cancer Institute, National Institutes of Health, under contract No. N01-CO-12400 (to J.W.).

REFERENCES

(1) Jemal, A.; Murray, T.; Ward, E.; Samuels, A.; Tiwari, R. C.; Ghafoor, A.; Feuer, E. J.; Thun, M. J. CA Cancer J Clin 2005, 55, 10-30.
(2) Kamijo, T.; Bodner, S.; van de Kamp, E.; Randle, D. H.; Sherr, C. J. Cancer Res 1999, 59, 2217-2222.
(3) Kelly-Spratt, K. S.; Gurley, K. E.; Yasui, Y.; Kemp, C. J. PLoS Biol 2004, 2, E242.
(4) Zhang, H.; Li, X. J.; Martin, D. B.; Aebersold, R. Nat Biotechnol 2003, 21, 660-666.