Concordant Release of Glycolysis Proteins into the Plasma Preceding a Diagnosis of ER\textsuperscript{+} Breast Cancer

Lynn M. Amon\textsuperscript{1}, Sharon J. Pitteri\textsuperscript{4}, Christopher I. Li\textsuperscript{1}, Martin McIntosh\textsuperscript{1}, Jon J. Ladd\textsuperscript{1}, Mary Disis\textsuperscript{3}, Peggy Porter\textsuperscript{2}, Chee Hong Wong\textsuperscript{1}, Qing Zhang\textsuperscript{1}, Paul Lampe\textsuperscript{1}, Ross L. Prentice\textsuperscript{1}, and Samir M. Hanash\textsuperscript{1}

Abstract

Although the identification of peripheral blood biomarkers would enhance early detection strategies for breast cancer, the discovery of protein markers has been challenging. In this study, we sought to identify coordinated changes in plasma proteins associated with breast cancer based on large-scale quantitative mass spectrometry. We analyzed plasma samples collected up to 74 weeks before diagnosis from 420 estrogen receptor (ER\textsuperscript{+}) cases and matched controls enrolled in the Women’s Health Initiative cohort. A gene set enrichment analysis was applied to 467 quantified proteins, linking their corresponding genes to particular biologic pathways. On the basis of differences in the concentration of individual proteins, glycolysis pathway proteins exhibited a statistically significant difference between cases and controls. In particular, the enrichment was observed among cases in which blood was drawn closer to diagnosis (effect size for the 0–38 weeks prediagnostic group, 1.91; \textit{P} = 8.3E-05). Analysis of plasmas collected at the time of diagnosis from an independent set of cases and controls confirmed upregulated levels of glycolysis proteins among cases relative to controls. Together, our findings indicate that the concomitant release of glycolysis proteins into the plasma is a pathophysiologic event that precedes a diagnosis of ER\textsuperscript{+} breast cancer. Cancer Res; 72(8); 1935–42. ©2012 AACR.

Introduction

Despite the mortality reduction associated with mammography (1), breast cancer remains the second leading cause of cancer mortality among the U.S. women (2). Early detection of breast cancer could potentially be enhanced through the development of biomarkers in blood to complement mammography. However, the discovery of protein markers that exhibit a significant change in their plasma concentration at early stages of breast tumor development has been challenging due to the wide dynamic range of protein abundance in plasma (3). A further challenge stems from the substantial protein variations that occur in plasma unrelated to tumor development and may confound the search for markers (4).

The feasibility of discerning changes in sets of genes, when changes in individual genes are subtle based on gene set enrichment analyses (GSEA), is a well-established concept in gene expression studies (5–9). We have applied a gene set type of analysis to a large-scale tandem mass spectrometry (MS)-based quantitative proteomics study comparing plasma drawn from women before the diagnosis of breast cancer with plasma from matched controls. We sought to identify coordinated changes in proteins by linking their corresponding genes to particular biologic pathways and computing an aggregate test statistic for all proteins in a gene set and generation of a null distribution by permutation of either sample or gene labels. This type of analysis has yielded significant trends in microarray data that would otherwise have not been detected (10–13).

In this proteomic study, we uncovered a biologic pathway involving glycolysis that exhibited a statistically significant difference between cases and controls on the basis of differences in the concentration of individual proteins in this pathway.

Patients and Methods

Human subjects

We conducted a nested case–control study within the Women’s Health Initiative Observational Study (WHI OS). The WHI OS is a prospective cohort of 93,676 postmenopausal women ages 50 to 79 years conducted through 40 clinical centers throughout the United States (3, 14). In the WHI OS, blood specimens were collected at 2 time points, at enrollment (baseline) and at year 3 of follow-up. A total of 420 women clinically diagnosed with estrogen receptor (ER\textsuperscript{+}) invasive breast cancer within 17 months of either their baseline or year 3 blood draw without a prior history of breast cancer were identified and selected for our study. Controls were individually matched 1:1 to cases on age at enrollment (±1 year), race/ethnicity (white, black, Hispanic, Asian/Pacific Islander, or...
other), blood draw date (± 1 year), and clinical center of enrollment. Matching was done in a time forward manner to ensure that each control had at least as much follow-up time following her blood draw as the time from blood draw to breast cancer diagnosis of the case to which she was matched. All blood samples were obtained in the fasting state (at least 12 hours) and maintained at 4°C for up to 1 hour until plasma or serum was separated from cells. Centrifuged aliquots were put into −70°C freezers within 2 hours of collection. Because of the need for in-depth proteomic analysis which engenders a throughput limitation (15), plasma specimens were pooled into 12 distinct case pools consisting of 35 breast cancer cases and 12 corresponding control pools of the same size (Table 1). The pools were stratified by progesterone receptor (PR) status, whether breast cancer was lobular or ductal and whether the blood was drawn close to diagnosis (0–38 weeks before diagnosis) or farther from diagnosis (38–74 weeks before diagnosis). Two additional paired sets of pools were also interrogated.

To determine plasma changes at the time of diagnosis, 1 set of each pool was immunodepleted of the 6 most abundant proteins with a Hu-6 column (Agilent Technologies). Samples from each pool were reduced with dithiothreitol and cysteine residues were alkylated with iodoacetamide. The peptides were then digested with trypsin. Each of the 14 sets of paired pools of plasma was evaluated by independent large-scale quantitative proteomics experiments using previously published methods (15–17). Briefly, each pool was immunodepleted of the 6 most abundant proteins with a Hu-6 column (Agilent Technologies). Samples were reduced with diithiothreitol and cysteine residues were alkylated with isotopically heavy acrylamide (1,2,3-C13) or light (C12) acrylamide to distinguish cases from control. Each case–control set of pooled samples were mixed and subjected to intact protein separation by anion exchange chromatography followed by reverse-phase chromatography, yielding 96 distinct fractions for each case–control pool. Each fraction was digested with trypsin.

**Mass spectrometry**

For each experiment, 96 fractions were each subjected to high-resolution MS-MS by a LTQ-Orbitrap mass spectrometer coupled with NanoLC-1D high-performance liquid chromatography. Liquid chromatography separation of the peptides was conducted with a 90-minute linear gradient from 5% to 40% acetonitrile in 0.1% formic acid. Spectra were acquired in data-dependent mode (m/z range, 400–1,800) and included selection of the 5 most abundant doubly or triply charged ions of each mass spectrometry (MS) spectrum for MS-MS analysis. The spectra were searched using Mascot against the human International Protein Index (IPI) database (v. 3.69). Quantitative information was extracted from acrylamide-labeled peptides using previously published quantitative tools (18) for any confident peptide (minimum PeptideProphet; ref. 19; probability >0.95, maximum fractional delta mass = 20 ppm, and had > one MS1 scan). To reduce the potential for bias due to misidentifications, we included only those peptides that were observed in both isotopic states over all experiments. ProteinProphet (20) was used to assign peptides to groups of indistinguishable proteins (due to peptide homology) and to calculate protein ratios from the geometric means of all the individual peptides. Proteins that had been removed from the IPI version 3.69, or were immunodepleted, were removed from the analysis. Mean log2 ratios of light/heavy intensities for each experiment were median-normalized.

**Gene set analyses**

Each protein groups identified by MS was assigned to a gene symbol using the IPI database or from annotation from National Center for Biotechnology Information (NCBI), if no
IPI gene symbol was available. To use canonical gene sets as functional groups, we mapped our protein groups onto gene symbols so that each gene symbol was represented only once. For protein groups that are ambiguously assigned to multiple gene symbols, we selected the one with the most annotation provided by the NCBI. For gene symbols that are represented by more than one protein group (e.g., multiple isoforms), we used the group with the highest number of peptides detected by MS. Proteins that were measured in only one preclinical plasma experiment were removed. Modulated test-statistics testing whether each protein ratio was different than 0 were computed using the limma package (ref. 21) from bioconductor.org and each observation was weighted using the number of quantitated events (peptides).

Gene symbols from the protein groups were matched to those in 200 Kyoto Encyclopedia of Genes and Genomes (KEGG) gene sets downloaded from MSigDb v2.5 (22). Only gene sets that included at least 5 proteins quantitated in our experiments were retained. Of the 200 KEGG gene sets analyzed, 24 included at least 5 proteins with a measurable case–control ratio in at least 2 of the preclinical plasma experiments. Significant enrichment of each of these 24 gene sets was computed as follows: t-statistics measured in the limma analysis were used to rank individual gene symbol and then the Wilcoxon rank-sum test was used to determine whether the ranks from genes in one set were higher or lower on average compared with the remainder proteins not in the set. The resulting Wilcoxon P values were used to compute false discovery rates (FDR; ref. 23) for all the gene sets. Statistics were generated for the set of all 12 preclinical plasma experiments, 2 of them showed statistically significant enrichment (FDR < 0.05; Table 2). The complement and coagulation gene set was enriched for proteins that were downregulated in cancer compared with control plasma. The magnitude of this enrichment was modestly higher among cases where blood was drawn closer to diagnosis (within 0–38 weeks) than in those, where blood samples were drawn further from diagnosis (38–74 weeks before; effect size, −0.98 vs. −0.43). The glycolysis and gluconeogenesis gene set was enriched for proteins that were upregulated in cases compared with controls that consisted primarily of proteins in the glycolysis pathway (Fig. 1). Supplementary Table S2 summarizes the results for each of the subsets of preclinical plasma experiments. Table 3 shows the fold changes for individual proteins of the glycolysis and gluconeogenesis gene set that were measurable in these experiments for the 4 different types of plasma assayed and t-statistics for the 2 types of preclinical samples. The identification of glycolysis proteins by MS was based on multiple peptides with cysteine-containing peptides used for quantification (Fig. 2). For the glycolysis protein set, the enrichment was observed among cases whose blood was drawn closer to diagnosis [effect size for the 0–38 weeks group, 1.91 (P = 8.3E-05) vs. 0.30 (P = 0.20 for the 38–74 weeks group)]. In the 0 to 38-week samples, all, but one, proteins were upregulated and the fold changes had a median log2 ratio of 0.31 and a maximum of 0.8 (fold change = 1.74). The samples taken further from diagnosis had smaller fold changes. In contrast, the control experiment yielded only 2 proteins, GAPDH and TPI1, with greater than 2-fold changes.

### Table 2. Wilcoxon P values and effect sizes for significant gene sets

| Complement and coagulation cascades (Number of proteins = 51) | Glycolysis and gluconeogenesis (Number of proteins = 12) |
|-----------------|-----------------|-----------------|
| Effect sizea    | P               | FDR             | Effect sizea   | P               | FDR             |
| All preclinical | −0.96           | 1.6E-07         | 3.8E-06        | 1.49            | 7.0E-04         | 8.4E-03         |
| 0–38 wk samples | −0.98           | 1.2E-06         | 2.8E-05        | 1.91            | 8.3E-05         | 9.9E-04         |
| 38–74 wk samples| −0.43           | 1.1E-03         | 0.021          | 0.30            | 0.20            | 0.47            |

*aEffect size = \log_2 (mean case–control t-statistic in gene set)/(mean case–control t-statistic not in set).*
The increase in effect size and statistical significance for samples drawn closer to diagnosis suggests that these concordant changes in protein concentration are disease-related. In the clinical group, all glycolysis proteins were upregulated (log2 ratio > 0) and most were increased greater than 2-fold. To test the hypothesis that increased levels of glycolysis proteins was related to disease, we computed Wilcoxon P values comparing log ratios measured in plasma collected at the time of breast cancer diagnosis versus controls, as well as in the control–control experiment in comparison with preclinical plasma findings (Fig. 3). While the receiver operating characteristic (ROC) curves for the complement and coagulation gene set for the 0 to 38 week, 38 to 74 week, and control samples were quite similar, all having an area under the curve (AUC) less than 0.5 with concordant similarities based on Wilcoxon P values, the glycolysis and gluconeogenesis gene set had a strikingly different pattern. As with the t-statistic analysis, the P-value for the 38 to 74 week samples was much greater than the 0 to 38 week samples. The differences in the distribution of protein ratios is reflected in the separation of the curves in Fig. 3B, whereas for the 0 to 38 week samples, 75% of the proteins in the set are in top 10% of all protein ratios. The 38 to 74 week curve, with an AUC of 0.62, is considerably closer to that of a uniform distribution of ratios. The trend in elevation of proteins in the glycolysis gene set in samples drawn closer to the time of diagnosis is supported by the results from the control–control and newly diagnosed data sets. As for the 0 to 38 week samples, the glycolysis protein ratios in the samples collected at the time of diagnosis rank near the top, with an increase in the AUC from 0.83 to 0.89 and a P value with an order of magnitude of greater significance. The values for the control–control samples were similar to those for the 38 to 74 week samples with AUC of 0.62 and no statistical significance between log2 ratios of proteins in the glycolysis and gluconeogenesis set and proteins that were not in the set.

Figure 1. Quantified proteins in the glycolytic pathways. Color representation consists of red (increased levels), green (decreased levels), and yellow (no significant quantitative change) based on log2 ratios for the 0 to 38 week prediagnostic samples relative to matched controls.
To assess the impact of individual experiments on the overall results, we computed Wilcoxon 2-sample $P$-values for each pool using the log2 ratio (Supplementary Table S2). For the complement and coagulation gene set, only three of the six 0 to 38 week samples had $P$-values $< 0.05$. However, for glycolysis and gluconeogenesis, five of the six 0 to 38 week experiments had $P$-values $< 0.05$. Among the six 38 to 74 week experiments, the effect sizes were smaller, half of them had the opposite sign, and only one pool had a significant $P$ value. These results indicate that the statistical significance of the enrichment for glycolysis and gluconeogenesis set in the 0 to 38 week samples experiments is not driven by a particular pool(s), whereas the significance of the enrichment of complement proteins does seem to be dominated by 2 pools.

**Discussion**

Glycolysis is the initial stage of glucose metabolism, the conversion of glucose into pyruvate and generation of energy in the form of ATP. This route to ATP production rather than oxidative phosphorylation occurs primarily when the cells are deprived of oxygen. A role for glycolysis in cancer was first suggested by Otto Warburg more than 50 years ago, who noted that tumor cells rely on glycolysis for ATP production even in the presence of oxygen (24). Recent reports indicate that mTOR activation is a key regulator of the Warburg effect leading to upregulation of glycolytic enzymes (25, 26). The "Warburg effect" is used for tumor detection by positron emission tomography (PET) following intravenous injection of the glucose analogue (18), $2[18F]$fluoro-2-deoxy-D-glucose (FDG; refs. 18, 27). Coordinated upregulation of glycolysis pathway proteins has been detected in several different tumor types (28–30) including breast cancer tumors (31). In a recent breast cancer study, the transcriptomic signature of human cancer cell lines and primary tumors for metabolic pathways associated with (18) FDG radiotracer uptake resulted in the identification of the glycolytic pathway as a major determinant of FDG uptake (32). Here, we describe the first study to identify

<p>| Table 3. Log$_2$ ratios and $t$-statistics for each of the 12 proteins in the glycolysis and gluconeogenesis gene set that were quantified in our experiments |</p>
<table>
<thead>
<tr>
<th>Protein</th>
<th>Number of preclinical pools measured</th>
<th>0-38 wk samples</th>
<th>38-74 wk samples</th>
<th>Clinical</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\text{Log}_2$ ratio</td>
<td>$t$-statistic</td>
<td>$\text{Log}_2$ ratio</td>
<td>$t$-statistic</td>
</tr>
<tr>
<td>ALDOA</td>
<td>11</td>
<td>0.26</td>
<td>2.38</td>
<td>0.03</td>
<td>0.28</td>
</tr>
<tr>
<td>ALDOB</td>
<td>6</td>
<td>0.10</td>
<td>0.76</td>
<td>-0.13</td>
<td>-1.44</td>
</tr>
<tr>
<td>ALDOC</td>
<td>5</td>
<td>0.80</td>
<td>4.90</td>
<td>0.22</td>
<td>1.63</td>
</tr>
<tr>
<td>ENO1</td>
<td>7</td>
<td>0.38</td>
<td>1.66</td>
<td>0.10</td>
<td>0.51</td>
</tr>
<tr>
<td>ENO2</td>
<td>3</td>
<td>0.59</td>
<td>1.96</td>
<td>0.10</td>
<td>0.24</td>
</tr>
<tr>
<td>ENO3</td>
<td>9</td>
<td>0.35</td>
<td>2.83</td>
<td>0.12</td>
<td>0.87</td>
</tr>
<tr>
<td>GAPDH</td>
<td>12</td>
<td>0.15</td>
<td>1.20</td>
<td>-0.02</td>
<td>-0.17</td>
</tr>
<tr>
<td>LDHA</td>
<td>6</td>
<td>0.41</td>
<td>2.50</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>LDHB</td>
<td>8</td>
<td>0.39</td>
<td>2.92</td>
<td>0.08</td>
<td>0.47</td>
</tr>
<tr>
<td>PGK1</td>
<td>9</td>
<td>0.27</td>
<td>3.69</td>
<td>0.12</td>
<td>1.19</td>
</tr>
<tr>
<td>PKM2</td>
<td>2</td>
<td>0.03</td>
<td>0.16</td>
<td>0.11</td>
<td>0.55</td>
</tr>
<tr>
<td>TPI1</td>
<td>12</td>
<td>-0.11</td>
<td>-0.67</td>
<td>-0.05</td>
<td>-0.32</td>
</tr>
</tbody>
</table>

Figure 2. Extensive peptide coverage for glycolysis proteins. Representative data for the identification of 3 glycolysis proteins, ALDOA, ENO1, and GAPDH, in a case-control experiment illustrate the extent of coverage achieved in the identification of glycolysis proteins in plasma by MS in this study. For each protein, the peptides identified are displayed below the predicted tryptic peptides. Orange indicates cysteine-containing peptides used for determination of case-control ratios.
increased levels of glycolysis proteins in plasmas of women with breast cancer. Given the greater frequency of ER− disease, particularly among postmenopausal women, relatively few women with ER+ disease were represented in the WHI cohort. We had sufficient power for analyses restricted to ER+ breast cancer.

Through this gene set analysis, probing more than 200 gene sets in the KEGG biologic pathway database, we identified 2 gene sets, glycolysis and gluconeogenesis and complement and coagulation, for which proteins exhibited concentration differences as a group in plasma drawn from subjects before being diagnosed with breast cancer relative to matched controls. Most of the path from glucose to pyruvate and then lactate is regulated by proteins that were elevated in the plasma of patients before diagnosis and even further elevated in the samples drawn at the time of diagnosis.

The complement and coagulation pathway gene set was significantly enriched for downregulated proteins when all sample pools were considered, but less significant in the samples taken further from diagnosis (38–74 weeks). However, several aspects of this relationship limit the significance of this finding. A comparison of case versus control plasmas involving newly diagnosed subjects with breast cancer did not yield enrichment in this protein set. Furthermore, the downregulation of proteins in this pathway was not consistently observed across the pools. Of note, the related gene set is relatively large including 69 genes, 52 of which had proteins measured in this study. This is more than twice as many as the next largest gene set. The large number of proteins contributes to the significance measures observed. In addition, the proteins in the complement and coagulation pathway are among the highest in abundance in plasma. They exhibit high intersample variability both across individuals and within samples taken from the same individual (33) due, in part, to sample preparation methodology (34, 35). In contrast, our findings with respect to the glycolysis pathway are supported by substantial evidence. First, there is an association between increased levels of proteins in this set and breast tumor development and progression. The related gene set was highly statistically significant in samples drawn closer to diagnosis, with the FDR dropping from 0.017 for all samples to 7e-3 in the 0 to 38 week samples. The magnitude and direction of this change was seen consistently across the six 0 to 38 week experiments. Unlike the complement and coagulation gene set, the trend of coordinated elevation of glycolysis proteins was also shown with breast cancer plasma samples collected at the time of diagnosis compared with controls. Furthermore, among the 0 to 38 week samples, all, but one, proteins had elevated levels among cases, with all proteins exhibiting elevated levels at the time of diagnosis. In contrast to the complement and coagulation gene set, this gene set includes fewer proteins whose low abundance in plasma reduces the chances that a statistical significance would be achieved by chance alone, as was observed in the control–control experiment for complement proteins.

Several additional analyses were conducted to rule out systematic correlations as the source of the enrichment. First, we applied a decorrelation procedure that has been devised for analysis of microarray data (32) and observed that the statistical significance of the glycolysis gene set was maintained among the 0 to 38 week samples (FDR < 0.005). To rule out correlation due to peptides common to multiple proteins, we removed these peptides and recalculated the protein ratios. We also restricted proteins to those that were observed in at least 6 preclinical pools in an effort to remove any technical sources of correlation. We found that even in this analysis in which the number of proteins linked to this gene set was reduced from 12 to 7, the statistical significance was maintained (FDR = 0.016).
These findings suggest that coordinated upregulation of the glycolysis and gluconeogenesis proteins in plasma is associated with disease rather than a random correlation. The source of increased glycolysis proteins in plasmas obtained from breast cancer subjects remains to be determined. While breast tumor cells and cell lines exhibit upregulated glycolysis and likely release glycolysis proteins into the extracellular environment through cell turnover, other cell populations may contribute to the increase in glycolysis proteins in the circulation. A model has been proposed in which glycolysis is also upregulated in non-tumor cells in breast cancer (36). Proteomic studies have identified glycolytic proteins in the nipple aspirates of breast cancer subjects and controls, supporting release of the proteins into the microenvironment through cell turnover (37). Moreover, lactate dehydrogenase (LDH) assays in serum have provided evidence of increased levels in breast cancer (38, 39). Increased levels of glycolytic proteins in plasma may also result from a host response to the presence of tumor. Increased blood leukocyte counts and/or leukocyte infiltration of tumor tissue may also contribute to increased levels of glycolysis proteins in circulation (40). Plasma proteome profiling of a mouse model of breast cancer collected at 2 time points during tumor development and progression identified elevated levels of several glycolysis proteins in tumor-bearing mice (41).

Novel approaches that complement mammography for the early detection of breast cancer could have a substantial impact on strategies for effective breast cancer screening. Discovery of blood-based biomarkers that have use for the early detection of breast cancer represents a substantial challenge. Comprehensive proteomic studies of a complex fluid such as plasma are hampered by a wide dynamic range of protein abundance and biologic and technical variability, complicating the identification of potential candidate markers that meet the FDR requirements. The approach we have used in this study for in-depth quantitative proteomics includes isotopic labeling of proteins and extensive fractionation before tryptic digestion and separate analysis of individual fractions by MS. As a result, the approach is labor intensive with limited throughput. Therefore, we used a sample pooling design given the large number of samples available for the study. A sample size of 35 cases and 35 controls per pool was selected because with pooling, the component of variance due to biologic variability decreases inversely with the pool size, whereas that due to the measurement process is expected to be independent of pool size. Therefore, by conducting 14 deep plasma profiling experiments with pool pairs of size 35, we estimated that this design had equivalent power to an individual level study, approximately 71% in size. This modest reduction in sample size and statistical power was determined to be acceptable and sufficient for our study’s purposes, given the preference for in-depth quantitative analysis rather than high-throughput analysis with a limited reach into the plasma proteome.

The gene set approach we have followed in this study has allowed us to uncover coordinated changes in proteins in the glycolysis pathway which at the individual protein level would not have achieved significance after correction for multiple hypotheses testing, yet taken together indicate a significant elevation in this pathway.

Disclosure of Potential Conflicts of Interest
M.L. Disclosures: commercial research grant, Glasso: ownership interest (including patents), University of Washington, Seattle, WA; and consultant/advisory board member, VentiRx. No potential conflicts of interests were disclosed by the other authors.

Grant Support
The study was supported by NCI grants 1 R21 161713-01 and U01 CA111273, NHLBI N01-WH-74313, and Komen Foundation grant KG110259.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received September 30, 2011; revised January 10, 2012; accepted January 31, 2012; published OnlineFirst February 24, 2012.

References
Concordant Release of Glycolysis Proteins into the Plasma Preceding a Diagnosis of ER + Breast Cancer

Lynn M. Amon, Sharon J. Pitteri, Christopher I. Li, et al.

*Cancer Res* 2012;72:1935-1942. Published OnlineFirst February 24, 2012.