Published in final edited form as:

Cancer Res. 2012 April 15; 72(8): 1935–1942. doi:10.1158/0008-5472.CAN-11-3266.

Concordant release of glycolysis proteins into the plasma preceding a diagnosis of ER+ breast cancer

Lynn M. Amon¹, Sharon J. Pitteri², Christopher I. Li¹, Martin McIntosh¹, Jon Ladd¹, Mary Disis³, Peggy Porter⁴, Chee-Hong Wong¹, Qing Zhang¹, Paul Lampe¹, Ross L. Prentice¹, and Samir M. Hanash¹

¹Fred Hutchinson Cancer Research Center, Public Health Sciences

²Stanford University, School of Medicine

³University of Washington, School of Medicine,

⁴Fred Hutchinson Cancer Research Center, Human Biology Division

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 prior to diagnosis from 420 ER+ 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 biological pathways. Based upon 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 pre-diagnostic group: 1.91; p-value: 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 pathophysiological event that precedes a diagnosis of ER+ breast cancer.

Introduction

Despite the mortality reduction associated with mammography (1) breast cancer remains the second leading cause of cancer mortality among 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 that 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 prior to a diagnosis of breast cancer to plasma from matched controls. We sought to identify coordinated changes in proteins by linking their corresponding genes to particular biological 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 biological pathway involving glycolysis that exhibited a statistically significant difference between cases and controls based on differences in the concentration of individual proteins in this pathway.

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 50–79 years of age conducted through 40 clinical centers throughout the United States (3, 14). In the WHI OS blood specimens were collected at two time points, at enrollment (baseline) and at year 3 of follow-up. A total of 420 women clinically diagnosed with ER+ 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 one hour until plasma or serum was separated from cells. Centrifuged aliquots were put into –70 degree Celsius freezers within two hours of collection. Due to 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. The pools were stratified by PR status, whether breast cancer was lobular or ductal and whether the blood was drawn close to diagnosis (0–38 weeks prior to diagnosis) or farther from diagnosis (38–74 weeks prior to diagnosis). Two additional paired sets of pools were also interrogated. To determine plasma changes at the time of diagnosis, one set consisted of case plasmas obtained from 30 women newly diagnosed with stage 1 ER+ ductal breast cancer patients. Control plasmas were from age-matched cancer-free controls. To determine variability in protein ratios among controls, another pair of plasma pools (control-control) was analyzed that consisted of plasma from 10 control subjects in each pool, that were WHI participants with no history of cancer, matched on age and ethnicity.

Sample preparation

Each of the 14 sets of paired pools of plasma was evaluated by independent large-scale quantitative proteomics experiments using previously described methods (15–17). Briefly, each pool was immuno-depleted of the six most abundant proteins using a Hu-6 column (Agilent Technologies). Samples were reduced with DTT and cysteine residues were alkylated with isotopically heavy acrylamide(1,2,3- C_{13}) or light (C_{12}) acrylamide to distinguish case from control. Each case-control set of pooled samples were mixed and subjected to intact protein separation by anion exchange chromatography followed by reversed 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 tandem mass spectrometry using a LTQ-Orbitrap mass spectrometer coupled with NanoLC-1D high performance liquid chromatography. Liquid chromatography separation of the peptides was performed 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 five most abundant doubly or triply charged ions of each 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 (19) probability > 0.95, maximum fractional delta mass = 20 ppm and had more than 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 since 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 group identified by MS was assigned to a gene symbol using the IPI database or from annotation from NCBI if no IPI gene symbol was available. In order 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 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 pre-clinical plasma experiment were removed. Moderated test-statistics testing whether each protein ratio was different than 0 were computed using the limma package (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 KEGG gene sets downloaded from MSigDb v2.5 (22). Only gene sets that included at least five proteins

quantitated in our experiments, were retained. Of the 200 KEGG gene sets analyzed, 24 included at least five proteins with a measurable case/control ratio in at least two of the pre-clinical plasma experiments. Significance 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 then the Wilcoxon rank-sum test was used to determine if the ranks from genes in one set were higher or lower on average compared to the remainder proteins not in the set. The resulting Wilcoxon p-values were used to compute false discovery rates (FDRs) (23) for all the gene sets. Statistics were generated for the set of all 12 experiments and two subsets: blood drawn 0–38 weeks to diagnosis (6 pools) and blood drawn 38–74 weeks to diagnosis (6 pools). For the subset analysis, t-statistics for each protein were computed by dividing the samples into two groups in the limma design matrix. Wilcoxon p-values were also calculated using the mean log₂ ratio of case/control (weighted by the number of quantitated peptides), rather than the t-statistic, for the set of all pre-clinical plasma experiments, the 0–38 week subset, the 38–74 week subset, the one early stage clinical experiment and the one control experiment.

Quantification of proteins was based on peptides with isotopically labeled cysteine residues in LC-MS experiments. Only proteins measured in at least two pre-clinical pools were retained, resulting in a total of 467 protein log ratios that were evaluated. As described above, these proteins were associated with a single gene symbol in order to compare to the gene sets in the KEGG biological pathway database. The complete table (Supplementary Table S1) is available as supplementary data.

Results

467 quantified proteins were subjected to analysis of gene sets represented in the KEGG biological pathway database. In the analysis of the twelve pre-clinical pools, effect sizes and p-values were computed comparing the test statistics of the case versus control comparisons for all proteins in a gene set to all other proteins measured. Of the 24 gene sets identified that included at least five proteins with a measurable case/control ratio in at least two of the pre-clinical plasma experiments, two of them showed statistically significant enrichment (FDR<0.05) (Table 2). The complement and coagulation gene set was enriched for proteins that were down-regulated in cancer compared to control plasma. The magnitude of this enrichment was modestly higher among cases where blood was drawn closer to diagnosis (within 0–38 weeks) compared to those where blood samples were drawn further from diagnosis (38–74 weeks prior) (effect size: –0.98 vs. –0.43). The glycolysis and gluconeogenesis gene set was enriched for proteins that were upregulated in cases compared to controls that consisted primarily of proteins in the glycolysis pathway (Figure 1). Supplementary Table S2 summarizes the results for each of the subsets of pre-clinical 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 four different types of plasma assayed and t-statistics for the two types of pre-clinical samples. The identification of glycolysis proteins by mass spectrometry was based on multiple peptides with cysteine containing peptides used for quantification (Figure 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-value: 8.3E–05) vs. 0.30

(p-value: 0.20 for the 38–74 weeks group)]. In the 0–38 week samples, all but one protein was up-regulated 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 two proteins, GAPDH and TPI1, with greater than two-fold changes.

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 up-regulated (log₂ ratio > 0) and most were increased greater than two-fold. To test the hypothesis that the 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 vs controls, as well as in the control-control experiment in comparison with pre-clinical plasma findings (Figure 3). Whereas the ROC curves for the complement and coagulation gene set for the 0–38 week, 38–74 week and control samples were quite similar, all having an AUC < 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 38–74 week samples was much greater than the 0–38 week samples. The differences in the distribution of protein ratios is reflected in the separation of the curves in Figure 3b whereby for the 0–38 week samples, 75% of the proteins in the set are in top 10% of all protein ratios. The 38–74 week curve, with an AUC=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 0–38 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 greater significance. The values for the control-control samples, were similar to those for the 38–74 week samples with an AUC=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.

To assess the impact of individual experiments on the overall results we computed Wilcoxon two-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 0–38 week samples had p-values <0.05. However for glycolysis and gluconeogenesis, five of the six 0–38 week experiments had p-values <0.05. Among the six 38–74 week experiments, the effect sizes were smaller, half of them had the opposite sign and only one pool has a significant p-value. These results indicate that the statistical significance of the enrichment for glycolysis and gluconeogenesis set in the 0–38 week samples experiments is not driven by a particular pool(s), while the significance of the enrichment of complement proteins does appear to be dominated by two 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 cells are deprived of oxygen. A role for

glycolysis in cancer was first suggested by Otto Warburg more than fifty 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 utilized for tumor detection by positron emission tomography (PET) following intravenous injection of the glucose analogue (18)F-fluorodeoxy-glucose (18)FDG (27). Coordinated up-regulation 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 resulting in the identification of the glycolytic pathway as a major determinant of FDG uptake (32). Here, we describe the first study to identify 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 over 200 gene sets in the KEGG biological pathway database, we identified two gene sets, glycolysis and gluconeogenesis and complement and coagulation, for which proteins exhibited concentration differences as a group in plasma drawn from subjects prior to 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 prior to 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 down-regulated 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 vs control plasmas involving newly diagnosed subjects with breast cancer did not yield an enrichment in this protein set. Furthermore, the down-regulation 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. Additionally, the proteins in the complement and coagulation pathway are among the highest in abundance in plasma. They exhibit high inter-sample 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 0–38 week samples. The magnitude and direction of this change was seen consistently across the six 0–38 week experiments. Unlike the complement and coagulation gene set, the trend of coordinated elevation of glycolysis proteins was also demonstrated with breast cancer plasma samples collected at the time of diagnosis compared to controls. Further, among the 0–38 week samples, all but one protein 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 many fewer proteins whose abundance with low abundance in plasma reducing 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 performed to rule out systematic correlations as the source of the enrichment. First, we applied a de-correlation 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 0–38 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 six pre-clinical 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 up-regulation of the glycolysis and gluconeogenesis proteins in plasma are 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 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 two 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 utility 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 biological and technical variability complicating the identification of potential candidate markers that meet FDR requirements. The approach we have utilized in this study for in-depth quantitative proteomics includes isotopic labeling of proteins and extensive fractionation prior to tryptic digestion and separate analysis of individual fractions by mass spectrometry. As a result the approach is labor intensive with limited throughput. Therefore we utilized 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 biological 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by NCI grants 1 R21 161713-01 and U01 CA111273, NHLBI N01-WH-74313, Komen Foundation grant KG110259

References

1. Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. *JAMA*. 1995; 273:149–54. [PubMed: 7799496]
2. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. *CA Cancer J Clin*. 2009; 59:225–49. [PubMed: 19474385]
3. The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. *Control Clin Trials*. 1998; 19:61–109. [PubMed: 9492970]
4. Hanash SM, Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. *Nature*. 2008; 452:571–9. [PubMed: 18385731]
5. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003; 34:267–73. [PubMed: 12808457]
6. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005; 102:15545–50. [PubMed: 16199517]
7. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A*. 2005; 102:13544–9. [PubMed: 16174746]
8. Kim BS, Kim I, Lee S, Kim S, Rha SY, Chung HC. Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics*. 2005; 21:517–28. [PubMed: 15374865]
9. Jiang Z, Gentleman R. Extensions to gene set enrichment. *Bioinformatics*. 2007; 23:306–13. [PubMed: 17127676]
10. Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, et al. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet*. 2008; 40:499–507. [PubMed: 18443585]
11. Cha S, Imielinski MB, Rejtár T, Richardson EA, Thakur D, Sgroi DC, et al. In situ proteomic analysis of human breast cancer epithelial cells using laser capture microdissection: annotation by protein set enrichment analysis and gene ontology. *Mol Cell Proteomics*. 2010; 9:2529–44. [PubMed: 20739354]

12. Oh JJ, Taschereau EO, Koegel AK, Ginther CL, Rotow JK, Isfahani KZ, et al. RBM5/H37 tumor suppressor, located at the lung cancer hot spot 3p21.3, alters expression of genes involved in metastasis. *Lung Cancer*. 2010; 70:253–62. [PubMed: 20338664]
13. Choi J, Southworth LK, Sarin KY, Venteicher AS, Ma W, Chang W, et al. TERT promotes epithelial proliferation through transcriptional control of a Myc- and Wnt-related developmental program. *PLoS Genet*. 2008; 4:e10. [PubMed: 18208333]
14. Hays J, Hunt JR, Hubbell FA, Anderson GL, Limacher M, Allen C, et al. The Women's Health Initiative recruitment methods and results. *Ann Epidemiol*. 2003; 13:S18–77. [PubMed: 14575939]
15. Pitteri SJ, Amon LM, Busald Buson T, Zhang Y, Johnson MM, Chin A, et al. Detection of elevated plasma levels of epidermal growth factor receptor before breast cancer diagnosis among hormone therapy users. *Cancer Res*. 2010; 70:8598–606. [PubMed: 20959476]
16. Katayama H, Paczesny S, Prentice R, Aragaki A, Faca VM, Pitteri SJ, et al. Application of serum proteomics to the Women's Health Initiative conjugated equine estrogens trial reveals a multitude of effects relevant to clinical findings. *Genome Med*. 2009; 1:47. [PubMed: 19402886]
17. Pitteri SJ, Hanash SM, Aragaki A, Amon LM, Chen L, Busald Buson T, et al. Postmenopausal estrogen and progestin effects on the serum proteome. *Genome Med*. 2009; 1:121. [PubMed: 20034393]
18. Faca V, Coram M, Phanstiel D, Glukhova V, Zhang Q, Fitzgibbon M, et al. Quantitative analysis of acrylamide labeled serum proteins by LC-MS/MS. *J Proteome Res*. 2006; 5:2009–18. [PubMed: 16889424]
19. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*. 2002; 74:5383–92. [PubMed: 12403597]
20. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*. 2003; 75:4646–58. [PubMed: 14632076]
21. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004; 3:Article3. [PubMed: 16646809]
22. Broad Institute. www.broadinstitute.org
23. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995; 57:289–300.
24. Warburg O. On the origin of cancer cells. *Science*. 1956; 123:309–14. [PubMed: 13298683]
25. Toschi A, Lee E, Thompson S, Gadir N, Yellen P, Drain CM, et al. Phospholipase D-mTOR requirement for the Warburg effect in human cancer cells. *Cancer Lett*. 2010; 299:72–9. [PubMed: 20805015]
26. Sun Q, Chen X, Ma J, Peng H, Wang F, Zha X, et al. Mammalian target of rapamycin up-regulation of pyruvate kinase isoenzyme type M2 is critical for aerobic glycolysis and tumor growth. *Proc Natl Acad Sci U S A*. 2011; 108:4129–34. [PubMed: 21325052]
27. Rigo P, Paulus P, Kaschten BJ, Hustinx R, Bury T, Jerusalem G, et al. Oncological applications of positron emission tomography with fluorine-18 fluorodeoxyglucose. *Eur J Nucl Med*. 1996; 23:1641–74. [PubMed: 8929320]
28. Bi X, Lin Q, Foo TW, Joshi S, You T, Shen HM, et al. Proteomic analysis of colorectal cancer reveals alterations in metabolic pathways: mechanism of tumorigenesis. *Mol Cell Proteomics*. 2006; 5:1119–30. [PubMed: 16554294]
29. Unwin RD, Craven RA, Harnden P, Hanrahan S, Totty N, Knowles M, et al. Proteomic changes in renal cancer and co-ordinate demonstration of both the glycolytic and mitochondrial aspects of the Warburg effect. *Proteomics*. 2003; 3:1620–32. [PubMed: 12923786]
30. Perroud B, Lee J, Valkova N, Dhirapong A, Lin PY, Fiehn O, et al. Pathway analysis of kidney cancer using proteomics and metabolic profiling. *Mol Cancer*. 2006; 5:64. [PubMed: 17123452]
31. Isidoro A, Casado E, Redondo A, Acebo P, Espinosa E, Alonso AM, et al. Breast carcinomas fulfill the Warburg hypothesis and provide metabolic markers of cancer prognosis. *Carcinogenesis*. 2005; 26:2095–104. [PubMed: 16033770]

32. Palaskas N, Larson SM, Schultz N, Komisopoulou E, Wong J, Rohle D, et al. 18F-Fluorodeoxy-glucose Positron Emission Tomography Marks MYC-Overexpressing Human Basal-Like Breast Cancers. *Cancer Res.* 2011; 71:5164–74. [PubMed: 21646475]
33. Corzett TH, Fodor IK, Choi MW, Walsworth VL, Turteltaub KW, McCutchen-Maloney SL, et al. Statistical analysis of variation in the human plasma proteome. *J Biomed Biotechnol.* 2010; 2010:258494. [PubMed: 20130815]
34. Gong Y, Li X, Yang B, Ying W, Li D, Zhang Y, et al. Different immunoaffinity fractionation strategies to characterize the human plasma proteome. *J Proteome Res.* 2006; 5:1379–87. [PubMed: 16739989]
35. Gundry RL, Fu Q, Jelinek CA, Van Eyk JE, Cotter RJ. Investigation of an albumin-enriched fraction of human serum and its albuminome. *Proteomics Clin Appl.* 2007; 1:73–88. [PubMed: 20204147]
36. Migneco G, Whitaker-Menezes D, Chiavarina B, Castello-Cros R, Pavlides S, Pestell RG, et al. Glycolytic cancer associated fibroblasts promote breast cancer tumor growth, without a measurable increase in angiogenesis: evidence for stromal-epithelial metabolic coupling. *Cell Cycle.* 2010; 9:2412–22. [PubMed: 20562527]
37. Pavlou MP, Kulasingam V, Sauter ER, Kliethermes B, Diamandis EP. Nipple aspirate fluid proteome of healthy females and patients with breast cancer. *Clin Chem.* 2010; 56:848–55. [PubMed: 20348408]
38. Koukourakis MI, Kontomanolis E, Giatromanolaki A, Sivridis E, Liberis V. Serum and tissue LDH levels in patients with breast/gynaecological cancer and benign diseases. *Gynecol Obstet Invest.* 2009; 67:162–8. [PubMed: 19088479]
39. Seth LR, Kharb S, Kharb DP. Serum biochemical markers in carcinoma breast. *Indian J Med Sci.* 2003; 57:350–4. [PubMed: 12944692]
40. Papaldo P, Di Cosimo S, Ferretti G, Vici P, Marolla P, Carlini P, et al. Effect of filgrastim on serum lactate dehydrogenase and alkaline phosphatase values in early breast cancer patients. *Cancer Invest.* 2004; 22:650–3. [PubMed: 15565824]
41. Pitteri SJ, Faca VM, Kelly-Spratt KS, Kasarda AE, Wang H, Zhang Q, et al. Plasma proteome profiling of a mouse model of breast cancer identifies a set of up-regulated proteins in common with human breast cancer cells. *J Proteome Res.* 2008; 7:1481–9. [PubMed: 18311905]

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 0–38 week pre-diagnostic samples relative to matched controls.

Figure 2. Extensive peptide coverage for glycolysis proteins

Representative data for the identification of three 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 mass spectrometry 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.

Figure 3. Gene set enrichment analysis in relation to time to diagnosis

The x-axis represents the rank of each protein's log₂ ratio in descending order while the y-axis is the proportion of proteins represented in the gene set for each rank as utilized for gene set enrichment scores (6) which are normalized such that cumulative score is zero. The area under the curve (AUC) serves as a measure of how the protein ratios in one gene set are distributed among all proteins measured.

Table 1

Description of plasma pools

Pool #	ER status	PR status	Histology	Blood draw timing (weeks prior to diagnosis)	Pool with heavy isotopic label
1	+	-	ductal	0-38	control
2	+	-	ductal	38-74	control
3	+	+	lobular	0-38	case
4	+	+	lobular	38-74	case
5	+	+	ductal	0-38	case
6	+	+	ductal	0-38	case
7	+	+	ductal	0-38	control
8	+	+	ductal	0-38	control
9	+	+	ductal	38-74	control
10	+	+	ductal	38-74	control
11	+	+	ductal	38-74	case
12	+	+	ductal	38-74	case
stage 1	+		ductal	drawn at diagnosis	control
control				healthy controls only	Control

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 size *	p-value	FDR	effect size *	p-value	FDR
All pre-clinical	-0.96	1.6E-07	3.8E-06	1.49	7.0E-04	8.4E-03
0-38 week samples	-0.98	1.2E-06	2.8E-05	1.91	8.3E-05	9.9E-04
38-74 week samples	-0.43	1.1E-03	0.021	0.30	0.20	0.47

* effect size = $\log_2(\text{mean case/control t-statistic in gene set})/(\text{mean case/control t-statistic not in set})$

Log2 ratios and t-statistics for each of the 12 proteins in the glycolysis and gluconeogenesis gene set that were quantified in our experiments

Table 3

Protein	Number of pre-clinical pools measured in	0–38 week samples		38–74 week samples		clinical		control
		log2 ratio	t-statistic	log2 ratio	t-statistic	log2 ratio	log2 ratio	
ALDOA	11	0.26	2.38	0.03	0.28	2.11		-0.21
ALDOB	6	0.10	0.76	-0.13	-1.44	0.44		-0.27
ALDOC	5	0.80	4.90	0.22	1.63	2.02		
ENO1	7	0.38	1.66	0.10	0.51	1.28		0.18
ENO2	3	0.59	1.96	0.10	0.24	0.12		0.18
ENO3	9	0.35	2.83	0.12	0.87	1.11		0.02
GAPDH	12	0.15	1.20	-0.02	-0.17	1.35		1.07
LDHA	6	0.41	2.50	0.01	0.06	1.50		-0.09
LDHB	8	0.39	2.92	0.08	0.47	1.50		-0.09
PGK1	9	0.27	3.69	0.12	1.19	0.96		
PKM2	2	0.03	0.16	0.11	0.55			-0.05
TPI1	12	-0.11	-0.67	-0.05	-0.32	0.45		1.63