Published in final edited form as:

Proteomics. 2015 November ; 15(22): 3772–3783. doi:10.1002/pmic.201500002.

A quantitative proteomics analysis of MCF7 breast cancer stem and progenitor cell populations

Song Nie¹, Sean P. McDermott^{2,3}, Yadwinder Deol^{2,3}, Zhijing Tan¹, Max S. Wicha^{2,3}, and David M. Lubman¹

¹Department of Surgery, University of Michigan, Ann Arbor, MI, USA

² Department of Internal Medicine, Division of Hematology/Oncology, University of Michigan, Ann Arbor, MI, USA

³ Comprehensive Cancer Center, University of Michigan, Ann Arbor, MI, USA

Abstract

Accumulating evidence has demonstrated that breast cancers are initiated and develop from a small population of stem-like cells termed cancer stem cells (CSCs). These cells are hypothesized to mediate tumor metastasis and contribute to therapeutic resistance. However, the molecular regulatory networks responsible for maintaining CSCs in an undifferentiated state have yet to be elucidated. In this study, we used CSC markers to isolate pure breast CSCs fractions (ALDH+ and CD44+CD24– cell populations) and the mature luminal cells (CD49f–EpCAM+) from the MCF7 cell line. Proteomic analysis was performed on these samples and a total of 3304 proteins were identified. A label-free quantitative method was applied to analyze differentially expressed proteins. Using the criteria of greater than twofold changes and *p* value <0.05, 305, 322 and 98 proteins were identified as significantly different in three pairwise comparisons of ALDH+ versus CD44+CD24–, ALDH+ versus CD49f–EpCAM+ and CD44+CD24– versus CD49f–EpCAM+, respectively. Pathway analysis of differentially expressed proteins by Ingenuity Pathway Analysis (IPA) revealed potential molecular regulatory networks that may regulate CSCs. Selected differential proteins were validated by Western blot assay and immunohistochemical staining. The use of proteomics analysis may increase our understanding of the underlying molecular mechanisms of breast CSCs. This may be of importance in the future development of anti-CSC therapeutics.

Keywords

Breast cancer; Cancer stem cell; Quantitative proteomics; LC-MS/MS; Pathway analysis; Systems biology

Correspondence: Professor David M. Lubman, Department of Surgery, The University of Michigan Medical Center, 1150 West Medical Center Drive, Building MSRB1 Room A510B, Ann Arbor, MI 48109-0656, USA, dmlubman@umich.edu, **Fax:** 734-615-2088.

Additional supporting information may be found in the online version of this article at the publisher's web-site

Colour Online: See the article online to view Figs. 4, 6 and 7 in colour.

1 Introduction

The cancer stem cell (CSC) model postulates that most cancers are hierarchically organized and maintained by a small population of CSCs at the apex with self-renewal capacity [1]. Originally, it was shown that breast CSCs can be prospectively isolated using flow cytometry and exhibit a CD44+CD24–EpCAM+ phenotype [2]. Furthermore, it is possible to prospectively isolate normal and malignant breast stem cells based on high intracellular expression of aldehyde dehydrogenase (ALDH) using the Aldefluor assay, which detects cells based on ALDH activity on a fluorescent substrate [3]. Correlative histological studies have linked clinical outcome to the presence of ALDH1+ or CD44+CD24– cells [3,4]. As such, breast CSCs are predicted to be involved in metastasis and mediate tumor recurrence after chemotherapy and radiotherapy due to the relative inability of these modalities to effectively target breast CSCs [5, 6]. If this is the case, then breast CSCs must be efficiently targeted to achieve a true cure [7].

Emerging evidence suggests breast CSCs reside in different states based on epithelial or mesenchymal properties (i.e. ALDH+ versus CD44+CD24–) [8, 9]. Our gene expression profiles indicate that breast CSCs share similar profiles across the major molecular subtypes (e.g. luminal, triple-negative, HER2+) [9]. These results suggest the possibility that breast CSCs might be targeted with a common anti-CSC therapeutic in combination with standard therapy (e.g. anti-estrogen, chemotherapy, Herceptin). While these gene expression profiles provide valuable insight into the biology of breast CSCs, the difference of protein expressed abundance and PTM is known to regulate signaling pathways [10]. Therefore, development of anti-CSC therapeutics requires a greater molecular understanding of breast CSCs, including protein-signaling pathways, which requires an unbiased proteomics approach. The development of MS based proteomics technology has played an increasingly important role in addressing genomewide molecular regulatory networks at the protein level. Fundamental to these analyses is the ability to comprehensively and accurately quantify thousands of cellular proteins. Several quantitative proteomics methods have been developed, including stable isotope labeling, isobaric labeling and label-free methods [11, 12]. Both labeling methods and the label-free methods have certain strengths and weaknesses [12]. Although labeling strategies have been widely used in quantitative proteomics, label-free quantitation has also been shown to be a useful quantitative strategy and commonly used for measuring protein differential expression in proteomics analysis [13, 14]. It has been successfully applied even in very limited cell samples [15, 16].

Patient-derived xenografts would serve as an ideal source for breast CSC material since they faithfully recapitulate the original patient tumor [17]. However, the relatively low breast CSC population frequency hinders the use of xenografts as a source of breast CSCs for proteomic analysis. To overcome this limitation, we used a luminal breast cancer cell line, MCF7, as a source for ALDH+ and CD44+CD24– breast CSCs since most cell lines contain both CSC fractions [18–20]. Within the MCF7 cell line, most cells exhibit a mature luminal cell phenotype, CD49f–EpCAM+ [21–23]. If breast CSCs are more prognostic than bulk cancer cells, the proteins governing stem cell properties are likely to influence clinical outcome.

In the present study, we applied a proteomics method to analyze ALDH⁺, CD44⁺CD24[−] and CD49f[−]EpCAM⁺ cells isolated by flow sorting from MCF7 cells utilizing an Orbitrap Elite high resolution mass spectrometer. Differentially expressed proteins were selected by statistical analysis in three pairwise comparisons using a label-free quantitative method based on spectral counts. Underlying significant pathways were further analyzed among the differential proteins. Selected differential proteins were validated by Western blot and immunohistochemical staining. The results presented here not only demonstrate the potential of a label-free based proteomics strategy, but also could reveal molecular networks associated with breast cancer progression and metastasis, which will help to improve the molecular understanding of breast CSCs and the development of anti-CSC therapeutics.

2 Materials and methods

2.1 Reagents

MCF7 breast cancer cells were obtained from ATCC (Manassas, VA). MCF7 cells were cultured in EMEM (ATCC) supplemented with 10% FBS, 10 µg/mL insulin (Sigma I02643), 50 units/mL penicillin and 50 µg/mL Streptomycin (Invitrogen). Anti-Mouse Ig CompBead Plus (#560497), anti-Rat Ig CompBead Plus (#560499), EpCAM-PE (#347198), CD44-APC (#559942), CD49f-PE-Cy5 (#551129), Isotype-PE (#349043), Isotype-APC (#340442), Isotype PE-Cy5 (#553931) and Isotype-PE-Cy7 (#348788) antibodies were from BD Bio-sciences (San Jose, CA). CD24 PE-Cy7 (#311120) was from Biolegend (San Diego, CA). Aldefluor reagent (BAAA-DA) was synthesized by the Vahlteich Medicinal Chemistry Core (University of Michigan) from BODIPY FL succinimidyl ester (Invitrogen #D2184), purified by TLC, structure verified by HPLC, aliquoted at 50 µg per micro-centrifuge tube and stored at −20°C (unpublished data). DAPI (D8417) and Live-Dead near-IR (L10119) were from Invitrogen and resuspended at 1 mg/mL in dd H₂O or 200 µL DMSO, respectively. The ArC Bead Kit (A10346) was from Invitrogen. Aldefluor assay buffer (#01072) and DEAB (#01705) were from STEM-CELL Technologies (Vancouver, BC, Canada).

2.2 Flow cytometry

Breast CSCs were quantified by staining with Aldefluor reagent, CD44, CD24, CD49f and EpCAM antibodies. Trypsinized MCF7 cells were filtered into FACS tubes (BD 352235), centrifuged and re-suspended in Aldefluor staining buffer. For cell sorting experiments, $1-3 \times 10^7$ cells were stained in 1000 µL total volume with 5 µL Aldefluor reagent, 0.3 ng/µL EpCAM-PE, 1 ng/µL CD49f-PE-Cy5, 1 ng/µL CD24-PE-Cy7, 0.3 ng/µL CD44-APC and 1% Live-Dead near-IR. Half of the cells were immediately transferred to a tube containing 5 µL DEAB. Control cells in 200 µL total Aldefluor staining buffer were stained with 1 µL Aldefluor reagent, 1 ng/µL Isotype-PE, 1 ng/µL Isotype-PE-Cy5, 1 ng/µL Isotype-PE-Cy7, 1 ng/µL Iso-APC and 1% Live-Dead near-IR. Compensation beads were used by staining one drop of positive and one drop negative beads with 1 µL of each antibody per manufacturer instructions. ArC beads were used with Live-Dead near-IR reagent per manufacturer instructions. After 30 min staining at 37°C, cells were washed, centrifuged and re-suspended with Aldefluor staining buffer. Cell sorting was performed using a MoFlo XDP (350, 488, 561, 640 nm lasers) or MoFlo Astrios (350, 405, 488, 561, 594, 640 nm lasers) (Beckman

Coulter). Cells were collected into PBS/3% FBS, spun down at 300 g in low protein-binding microcentrifuge tubes, washed with cold PBS, and frozen at -80°C in a minimal volume of PBS.

2.3 Cell protein extraction and tryptic digestion by FASP

Filter-aided sample preparation method (FASP) was used to prepare protein samples including protein extraction and digestion, which has proved to be a useful sample preparation method, especially for limited cell samples [24]. The FASP Protein Extraction kit and Digestion kit were from Expedeon, Inc. (Knoxville, TN). The same number of cells from different cell populations were lysed in 30 μL of lysis buffer, which contains 4% SDS, 0.1M DTT and protease inhibitors, in a 100°C water bath for 5 min. The lysed samples were centrifuged at 15 000 g for 10 min and the supernatant was transferred to another vial for further tryptic digestion. For trypsin digestion, the FASP digestion procedure was followed. After a serial buffer change and alkylation, trypsin (Promega, Madison, WI) was added in 1:30 ratio and incubated overnight at 37°C . Finally, the digest was cleaned using C18 ZipTipTM (Millipore, Billerica, MA).

2.4 LC-MS/MS analysis

The digested samples were analyzed in triplicates by LC-MS/MS. The LC-MS/MS analysis was performed on an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific) coupled with a Proxeon EASY nLC II System (Thermo). Peptides were separated on an in house packed 25 cm column (inner diameter 75 μm , Magic C18 AQ 100 Angstrom, 5 μm) at flow rate of 400 nL/min. A 155 min linear gradient from 2 to 35% ACN (0.1% formic acid) was used. MS spectra were acquired in data dependent mode. Full mass scans (m/z 400–2000) were acquired in the Orbitrap analyzer with resolution $R = 120\,000$ at m/z 400, and the 20 most intense ions in the full mass scan were selected for MS/MS analysis by CID in the linear ion trap. The normalized collision energy was set at 35% for MS/MS. Dynamic exclusion was set to exclusion duration of 60 s.

2.5 Data analysis

All collected raw data were searched against a database generated from the UniProtKB/Swiss-Prot human protein knowledge base (released in May 2012) using SEQUEST (version 1.1.0.263) in Proteome Discoverer 1.1 (Thermo). Searches were performed using the following settings: precursor ion m/z tolerance: ± 10 ppm; fragment ion m/z tolerance: ± 0.6 Da; two missed cleavages allowed; static modification: carbamidomethyl of cysteine (+57.02146 Da, C) and dynamic modifications: oxidation (+15.99492 Da, M). The search results were filtered by 1% false discovery rate (FDR) at peptide-level. If multiple proteins share the same peptide sequences, they will be reported as a protein group. Label free quantification was carried out using spectral counting method, which is based on counting the total number of matched peptide spectra for a given protein and has been shown to be a useful and practical method for quantitative analysis of large-scale proteomics data [25]. The spectral count of each individual protein was normalized by the sum of the spectral counts per LC-MS/MS run [26].

2.6 Bioinformatics analysis

ANOVA and *t*-test statistical methods were used to identify significantly differentially expressed proteins. The *P*-value generated from ANOVA or *t*-test statistical analysis (< 0.05) was considered as a significant difference. The clustering method for x and y axis used here is K-means clustering. Euclidean method was used to calculate distance. GO annotation for cellular component and biological process of the identified proteins were performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) bioinformatics resources [27,28]. The molecular function and biological networks of the differentially expressed proteins were analyzed using the Ingenuity Pathway Analysis software (Ingenuity Systems, MountainView, CA).

2.7 Western blot analysis

15 μ g of proteins from the three cell populations were separated by 4–15% SDS-PAGE and transferred to polyvinylidene difluoride (PVDF) membranes (Bio-Rad, CA). The membranes were blocked with 2% milk 1h, and incubated with primary antibodies (1:500) overnight at 4°C. The membranes were washed with PBST three times, incubated with peroxidase-conjugated anti-IgG (H&L) secondary antibody (Abcam) for 1 h and washed another three times with PBST, followed by detection with Supersignal West Pico Chemiluminescent horseradish peroxidase substrate (Thermo Scientific, IL). The antibodies of protein HSPB1 (Heat shock protein beta-1) and ANXA3 (Annexin A3) were purchased from the Developmental Studies Hybridoma Bank and Abcam company.

2.8 Immunohistochemical analysis of tissue microarrays

The paraffin-embedded tissue arrays (BRN801a and BC081120, US Biomax, Rockville, MD) were dewaxed in xylene for 10 min twice and rehydrated through a series of alcohol solutions (200 proof, Sigma-Aldrich, St. Louis, MO) (100% ethanol twice, 90% ethanol, 70% ethanol, 5 min each) to water. Antigen retrieval was done by boiling the slides for 15 min in a citrate buffer (Teknova, U.S) at pH 6.0. Endogenous peroxidase activity was blocked with 3% H_2O_2 in methanol for 10 min after returning to room temperature. The TMAs were subsequently blocked with 2% BSA and incubated with monoclonal antibody (1:75) overnight at 4°C followed by incubation with a secondary antibody conjugated to horseradish peroxidase. Immunodetection was performed using DAB solution (Vector Laboratories, Burlingame, CA). Hematoxylin counterstain was used to visualize nuclei. The protein expression level in each tissue section was assessed in non-necrotic areas of three separate microscopic fields of view under a magnification of 20 $\times$ and was represented by the mean of the percentage of positive cells.

3 Results

3.1 Workflow and isolation of CSC populations

To investigate the molecular network including protein-signaling pathways in CSC, a quantitative proteomics analysis was performed on MCF7 CSCs and mature luminal cells. A schematic of the experimental workflow is shown in Fig. 1. ALDH+ CSCs, CD44+CD24– CSCs and CD49f-EpCAM+ mature luminal cells were collected. The cell protein content

was then extracted and digested by the FASP method. Tryptic peptides were analyzed by LC-MS/MS based on an Orbitrap Elite MS and a label-free quantitative method using spectral counts was applied to quantify differentially expressed proteins among the three different cell populations. Selected differential proteins were validated by Western blotting or immunohistochemical analysis.

Flow cytometry sorting was used to isolate MCF7 breast CSCs and mature luminal cells using CSC markers. As shown in Fig. 2, CD44+CD24⁻ stem cells were gated (P1) and sequentially gated for ALDH⁻ (P2) and not CD49f-EpCAM⁺ (P3) cells. ALDH⁺ stem cells were gated (P4) and sequentially gated for not CD44+CD24⁻ (P5) and not CD49f-EpCAM⁺ (P6) cells. CD49f-EpCAM⁺ cells were gated (P7) and sequentially gated for not CD44+CD24⁻ (P8) and ALDH⁻ (P9) cells. The frequencies represent values based on each parental gate. Flow cytometry analysis showed that CD44+CD24⁻ and ALDH⁺ stem cells were expressed in about 0.04 and 0.16% of MCF7 cells, respectively, whereas CD49f-EpCAM⁺ non stem cells account for about 83.5% of MCF7 breast cancer cells (Fig. 2). The breast CSCs were very limited in MCF7 cells. To collect enough cell samples for quantitative proteomics analysis, we isolated the three cell populations by FACS sorting 10 times and collected about 50 000 ALDH⁺ stem cells, 320 000 CD44+CD24⁻ stem cells and 7000 000 CD49f-EpCAM⁺ non-stem cells. The same number (50 000) of cells from each fraction was used for proteomics analysis based on the limited stem cell number.

3.2 Identification of proteins from the three different cell populations

A FASP method was used to extract and digest proteins from 50 000 cells of the three different cell populations. In the analysis of a small number of cell samples, a highly efficient sample preparation method including protein extraction and digestion is very critical to improve protein identification. The FASP method has been demonstrated to be such a highly efficient and reproducible sample preparation, especially for small cell or tissue samples [24,29]. The cell samples collected from several cell-sorting experiments were lysed and combined using lysis buffer. The same number of cell lysates from each of three different cell populations was digested using the FASP method and divided into four fractions for MS analysis. The resulting peptides of each sample were analyzed by three replicates of LC-MS/MS using an Orbitrap Elite mass spectrometer. A representative base peak chromatogram of the three samples is shown in Supporting Information Fig. S1.

The analysis of the mass spectrometry data shows that around 4400–7700 unique peptides, corresponding to 1500–2000 proteins, were identified in each single run. By combining the three replicates for each sample, around 2000–2600 proteins were identified. The number of proteins, peptides and unique peptides identified in each individual run are listed in Supporting Information Table S1. A total of 3304 unique proteins were identified across the three different cell populations, with 302 of these proteins unique to ALDH⁺ cells, 308 unique to CD44+CD24⁻ cells, 484 unique to CD49f-EpCAM⁺ cells and 1477 proteins common to the three cell populations (Fig. 3A). The detailed information on the total proteins and significantly different proteins across the three samples are listed in Supporting Information Tables S2 and S3. The reproducibility of three replicates of LC-MS/MS analysis was evaluated using R^2 value from linear regression for spectral counts of identified

proteins and the average CV (coefficient of variation) of each protein spectral count among three technical replicates. All the R^2 values and average CV for the analysis in the three cell samples were ~0.99 and less than 16%, respectively, which indicates that the LC-MS/MS analysis was highly reproducible in this work. The biological reproducibility has also been evaluated using the CD49f-EpCAM+ cell population. Each cell sample digestion obtained from an independent flow sorting experiment of 50 000 cells was analyzed in triplicate by MS. The combined data from three runs were applied to evaluate biological reproducibility. More than 70% of the proteins were identified in all three independent experiments (Supporting Information Fig. S2a). The percentage of overlapping proteins is very high. In addition, the spectral counts in different experiments were compared. The correlation was found to be around 0.98 (Supporting Information Fig. S2b).

GO analysis was further performed for identified proteins from the three cell samples using the DAVID bioinformatics tool and database [27]. GO cellular component annotation (Fig. 3B) showed that these proteins were primarily located to the cytosol, nucleus, membrane, mitochondrion and cytoskeleton. There were more cytoskeleton proteins in the ALDH+ cells compared with others. The GO biological process annotation (Fig. 3C) revealed that more than 5% of identified proteins are relevant to positive regulation of catalytic activity and regulation of cell death for ALDH+, CD44+CD24- stem cells, compared with 1% for CD49f-EpCAM+ non-stem cells. This result suggests a significant alteration in signaling regulation between ALDH+, CD44+CD24- stem cells and CD49f-EpCAM+ non-stem cells.

3.3 Differentially expressed proteins among the three different cell populations

Proteins identified from the ALDH+, CD44+CD24- stem cell populations and CD49f-EpCAM+ non-stem cell population were quantified using the spectral counting quantitative strategy. To identify the differences among the three fractions, the tryptic peptides from the same number of cells (12 500 cells) were first analyzed by LC-MS/MS using a long gradient. Then, the spectral counts for each protein were normalized by total spectral counts in each LC-MS/MS run to reduce the variation between samples and replicates [30]. The average normalized spectral counts for each target protein was used to calculate a ratio in each pairwise comparison and a Student's *t*-test was used to generate a *p*-value. The histograms (Fig. 4A) of protein ratios for three pairwise comparisons showed that all the proteins were reproducibly clustered in a normal distribution centered at the $\log_2$ (protein ratio) of ~0, which demonstrated that the protein loading amount from each sample is equivalent. The deviation of the $\log_2$ (protein ratio) distribution for ALDH+/CD44+CD24- and ALDH+/CD49f-EpCAM+ comparisons was larger than that of the CD44+CD24-/CD49f-EpCAM+ comparison (Fig. 4A).

Volcano plots based on $-\log_{10}(p\text{-value})$ and $\log_2$ (protein ratio) of protein quantification were used to demonstrate statistically significant changes between pairwise comparisons (Fig. 4B). Using the criteria of greater than twofold changes and *p*-value < 0.05 [31, 32], 305, 322 and 98 proteins were identified as significantly different in ALDH+/CD44+CD24-, ALDH+/CD49f-EpCAM+ and CD44+CD24-/CD49f-EpCAM+ comparisons, respectively. The differential proteins across the three samples were analyzed by an unsupervised hierarchical clustering (Fig. 4C). Consequently, the clustering algorithm

grouped all the replicates of each cell population together, which means high technical reproducibility was obtained in our LC-MS/MS analysis. As expected, the ALDH+ cell population was distinguished from the CD44+CD24– and CD49f-EpCAM+ cell populations.

3.4 Signal pathways analysis of differentially expressed proteins

To investigate biological and molecular networks underlying the three different cell populations, the significant differential protein lists from the three pairwise comparisons were investigated by Ingenuity Pathway Analysis (IPA). A threshold of p -value < 0.05 was used to filter the significant canonical pathways from the three differential protein lists. The significant canonical pathways are shown in Fig. 5, where the height of the bars presents the significance of the signaling pathways of which the differentially expressed proteins are involved.

Three major pathway categories of cell-cell junctions, glucose metabolism, and signal conduction and regulation were over-represented among the differential proteins. The enriched cell–cell junction pathways includes remodeling of Epithelial Adherens Junctions, Epithelial Adherens Junction Signaling Pathways, Germ Cell-Sertoli Cell Junction Signaling, Sertoli Cell-sertoli Cell Junction Signaling and Gap Junction Signaling pathway, which modulate cell motility and surface remodeling and thus regulate cell migration and invasion [33]. These pathways also play an important role in tumorigenesis and cancer migration [34]. eIF2 signaling conduction and regulation of eIF4 and p70S6K pathways, which regulate the initiation of protein synthesis and play a critical role in cell cycle progression were over-represented in ALDH+ versus CD49f-EpCAM+ [35]. Glucose metabolism pathways including glycolysis and gluconeogenesis were enriched in ALDH+/CD44+CD24–, ALDH+/CD49f-EpCAM+, but not in CD44+CD24–/CD49f-EpCAM+ comparisons. All eleven enzymes involved in the glycolysis pathway were identified and quantified across these three cell populations in this quantitative analysis (Supporting Information Table S5).

To further discover the functional relevance of these significantly expressed proteins, we generated connectivity networks by IPA. For the ALDH+/CD44+CD24– comparison, 29 proteins involved in cancer, dermatological disease and conditions, and hematological disease were grouped as the top network (Fig. 6A). The top biological network for differentially expressed proteins within the ALDH+/CD49f-EpCAM+ comparison includes 32 proteins involved in cell death and survival and neurological diseases (Fig. 6B). For the CD44+CD24–/CD49f-EpCAM+ comparison, the top network includes 29 proteins relevant to cellular compromise, hereditary disorder and neurological disease (Fig. 6C).

3.5 Validation of selected differential proteins

To validate the quantitative data, two significantly different proteins (HSPB1, ANXA3) were selected for verification of their differential expression in the three different cell populations by Western blotting analysis. The selection of these proteins was based on the MS results and their important roles in cell death and survival. Based on the MS quantification results, HSPB1 (Heat shock protein beta-1) protein has higher expression

level in ALDH⁺ and CD44⁺CD24⁻ CSCs compared to CD49f-EpCAM⁺ bulk cells. The fold change of ALDH⁺/CD49f-EpCAM⁺ and CD44⁺CD24⁻/CD49f-EpCAM⁺ are 2.6 and 1.75, respectively. The expression of ANXA3 (Annexin A3) was up-regulated about 3.5 fold in ALDH⁺ CSCs compared to bulk cells. However, the spectral counts of ANXA3 in the CD44⁺CD24⁻ CSCs sample were too low for quantification. Consistent with the proteomics quantitative results, Western blot analysis also showed the two proteins were up-regulated in CSCs compared to bulk cells, which demonstrates that the quantitative results observed by MS analysis was reliable (Fig. 7A).

ANXA3 is associated with a poor prognosis in breast cancer and participates in the modulation of apoptosis in vitro by affecting the Bcl-2/Bax balance [36]. Western blot analysis revealed that ANXA3 was expressed up-regulated in stem cells compared to bulk cells. The differential expression of the protein ANXA3 was further confirmed by immunohistochemical staining on two breast tissue microarrays. In the 70 normal and cancer adjacent normal samples, there is a weak ANXA3 expression in about 73% cases (51 cases), with only four samples exhibiting a strong ANXA3 protein expression. In the 100 invasive ductal breast carcinoma microarray, 48% of the samples expressed strong ANXA3 protein (48 cases), another 32% have a medium expression level of ANXA3 protein (32 cases). These results apparently show that ANXA3 expression is significantly higher in breast cancers compared to normal breast tissue. A representative figure is shown in Fig. 7B.

4 Discussion

Breast cancer is the most common cancer in women, and remains a major cause of mortality in women worldwide [37]. Increasing evidence has supported that breast cancers are initiated and developed from a small population of CSC, which are hypothesized to mediate tumor metastasis and contribute to the failure of breast cancer therapy due to their therapeutic resistance [38]. Breast CSCs have been characterized as exhibiting a CD44⁺/CD24⁻ or ALDH⁺ phenotype. Here we applied an integrated label-free approach for quantitative analysis of the proteome of CD44⁺/CD24⁻, ALDH⁺ CSCs and CD49f-EpCAM⁺ + mature luminal cells to discover the difference between breast CSC with mature breast differentiated cells as well as increase our understanding of the underlying molecular network involved in CSC's characteristics. MCF7 cells were selected as they contain a higher percentage of ALDH⁺ EpCAM⁺ and CD44⁺/CD24⁻/EpCAM⁺ populations of CSCs as well as differentiated epithelial cells which can then be sorted in numbers sufficient to allow proteomic studies. Other cell lines representing different subtypes of breast cancer do not contain a considerable percentage of both types of stem cells. Moreover, our group has previously shown that ALDH⁺ and CD44⁺/CD24⁻ populations of CSCs isolated from cell lines as well as patient samples and normal breast tissue are similar [9], therefore, we believe that data generated from MCF7 CSCs could be extended to other cell lines/tumor sub-types as well.

We have used the well-characterized MCF7 breast cancer cell line to fractionate breast CSCs. About 0.04% of MCF7 cells are CD44⁺CD24⁻ CSCs and 0.16% of cells are ALDH⁺ CSCs based on flow cytometry results. Due to the low frequency, the collection of a sufficient number of stem cells to perform quantitative proteomics analysis presented

difficulty. After FACS sorting over ten experiments, we collected only about 50 000 ALDH⁺ cells, from which we estimated that we extracted about 4~5 µg cell protein. Only one-fourth of the prepared sample was loaded for LC-MS/MS analysis for each run. Although the limited cell samples were applied here, we still identified 3304 proteins across the three different fractions. Among them, 305, 322 and 98 proteins were identified as significantly different in ALDH⁺/CD44⁺CD24⁻, ALDH⁺/CD49f⁻ EpCAM⁺ and CD44⁺CD24⁻/CD49f⁻ EpCAM⁺ comparisons based on twofold change and *p*-value <0.05, respectively. Although the number of proteins identified was limited by the small number of cells, comparison of these stem cells with their progenitor cells revealed profound differences among them. Therefore, these differentially expressed proteins may be potential factors important for the maintenance of stem cell self-renewal and pluripotent states.

The loss of cell polarity and cell-cell junctions and the partial disassembly of epithelial tissues are changes associated with cancer progression to an invasive, metastatic state. Several junction associated pathways were enriched significantly in the differential proteins from the three pairwise comparisons of ALDH⁺/CD44⁺CD24⁻, ALDH⁺/CD49f⁻EpCAM⁺ and CD44⁺CD24⁻/CD49f⁻EpCAM⁺, which modulate cell motility and surface remodeling and thus regulate cell migration and invasion [33]. These functions of junction pathway are highly associated with CSC features. For example, Claudin proteins, major integral membrane proteins of tight junctions, were identified in our experiment (Claudin-3 and -7).

CD44⁺/CD24⁻ CSCs are characterized by expression of epithelial-mesenchymal transition (EMT) genes, invasive and largely quiescent, whereas mesenchymal-epithelial-like (MET) stem cells are characterized by expression of ALDH, epithelial morphology and self-renewal [8, 9]. We recently reported that breast CSCs are capable of inter-converting between two cellular states in a process regulated in part by the tumor microenvironment [9]. Cell-cell junction pathways play important roles in the epithelial-mesenchymal transition, which is an important contributor to the invasion and metastasis of epithelial-derived cancers [34, 39, 40]. Thomson et al. measured a marked and consistent down-regulation of cell-polarity, tight-junction, adherens-junction, desmosomal-junction and basement membrane adhesion components which was correlated with a mesenchymal EMT state [39].

Cancer progression is often accompanied by the alterations in metabolic pathways, including enhanced glycolysis, lipogenesis and/or autophagy in highly proliferative cancer cells, including CSCs [41]. The altered metabolic pathways may supply the energy and nutrients necessary for their sustained growth and survival under normoxic and hypoxic conditions [42]. In our study, pathway analysis showed that glycolysis and gluconeogenesis pathways are significantly enriched in the differentially expressed proteins from the ALDH⁺/CD44⁺CD24⁻, ALDH⁺/CD49f⁻EpCAM⁺ comparison (Fig. 5). All 11 enzymes involved in glycolysis were identified and quantified across the three fractions in this quantitative analysis (Supporting Information Table S5). Seven of the eleven enzymes are found to be more than 1.5 fold up-regulated in ALDH⁺ CSCs, but not in the CD44⁺CD24⁻ CSCs. These results are also consistent with characteristics of ALDH⁺ CSCs, which have high intracellular expression of ALDH [3]. Beta-enolase (ENO3) protein was identified only in ALDH⁺ CSCs (Supporting Information Table S5). Some studies have indicated that there is

a transition from the alpha/alpha homodimer to the alpha/beta heterodimer in some tissues during ontogenesis [43].

Annexins are a family of intracellular proteins that bind membrane phospholipids in a Ca(2+) concentration-dependent manner. Several Annexins play important roles during tumor progression. Ten Annexin proteins were identified in our data (Supporting Information Table 4). For example, the ANXA3 protein is associated with a poor prognosis in breast cancer and participates in the modulation of apoptosis in vitro by affecting the Bcl-2/Bax balance [36]. In our experiments, ANXA3 was quantified as an up-regulated protein in ALDH+ CSCs compared to CD44+CD24- CSCs and CD49f-EpCAM+ cancer cells by MS analysis, which was demonstrated by Western blot analysis and immunohistochemical staining. In addition, microarray analysis also identified Annexin A3 to be consistently up-regulated in ALDH+ CSCs compared to matched CD44+CD24- CSCs (data not shown). Based on the proteomics data, Annexin A2 (ANXA2) has the same trend of change as ANXA3. Some reports have indicated ANXA2 plays a role in the regulation of cellular growth and in the signal transduction pathway [44]. However, three other members of the Annexin family, Annexin A5, A6 and A11, were found up-regulated in CD49f-EpCAM+ cancer cells (mature luminal cells) compared to the two CSC populations. Annexin A5 has been reported to be involved in an antiproliferative mechanism in endometrial cancer cells [45] and inhibits the apoptotic body formation [46]. The results are consistent with the characteristics of CSCs with increased ability for proliferation and growth. Song et al. also demonstrated that Annexin A11 is directly involved in cell proliferation and cisplatin resistance of ovarian cancer [47].

In summary, a quantitative proteomics analysis of breast CSCs and progenitor cells was performed. More than 3000 proteins were identified across the three cell populations. 305, 322 and 98 proteins were identified as significantly different in ALDH+/CD44+CD24-, ALDH+/CD49f-EpCAM+ and CD44+CD24-/CD49f-EpCAM+ comparisons, respectively. Most of the differential proteins among the three fractions are involved in tumorigenesis, cancer migration, energy metabolism, cancer invasion and cancer progression. Of ten Annexin proteins identified in our data, five of them showed significant differences among the three types of cell populations, of which the differential expression of ANXA3 was confirmed by Western blot analysis and immunohistochemical staining. The Annexin proteins play important roles during tumor progression. These differentially expressed proteins represent promising targets to further decipher molecular characteristics of breast CSCs and to help understand molecular mechanisms of CSC therapy resistance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge support by the National Institutes of Health through Grant GM01 GM49500 (D.M.L.) and Promise Award (KG120001) from the Susan G Komen foundation.

S.N, S.P.M performed research; M.S.W, D.M.L provided research support; S.N, S.P.M, Z.J.T, M.S.W, D.M.L wrote the paper.

Potential conflict of interest: M.S.W. holds equity in OncoMed Pharmaceuticals, receives research support from MedImmune and Dompe, and is a scientific advisor to MedImmune, Verastem, Cerulean, and Paganini. S.P.M receives research support from Verastem.

References

- [1]. Dick JE. Stem cell concepts renew cancer research. *Blood*. 2008; 112:4793–4807. [PubMed: 19064739]
- [2]. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl. Acad. Sci. U S A*. 2003; 100:3983–3988. [PubMed: 12629218]
- [3]. Ginestier C, Hur MH, Charafe-Jauffret E, Monville F, et al. ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. *Cell Stem Cell*. 2007; 1:555–567. [PubMed: 18371393]
- [4]. Ricardo S, Vieira AF, Gerhard R, Leitao D, et al. Breast cancer stem cell markers CD44, CD24 and ALDH1: expression distribution within intrinsic molecular subtype. *J. Clin. Pathol*. 2011; 64:937–946. [PubMed: 21680574]
- [5]. Creighton CJ, Li X, Landis M, Dixon JM, et al. Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. *Proc. Natl. Acad. Sci. USA*. 2009; 106:13820–13825. [PubMed: 19666588]
- [6]. Liu R, Wang X, Chen GY, Dalerba P, et al. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *N. Engl. J. Med*. 2007; 356:217–226. [PubMed: 17229949]
- [7]. McDermott SP, Wicha MS. Targeting breast cancer stem cells. *Mol. Oncol*. 2010; 4:404–419. [PubMed: 20599450]
- [8]. Liu S, Clouthier SG, Wicha MS. Role of microRNAs in the regulation of breast cancer stem cells. *J. Mammary Gland Biol. Neoplasia*. 2012; 17:15–21. [PubMed: 22331423]
- [9]. Liu S, Cong Y, Wang D, Sun Y, et al. Breast cancer stem cells transition between epithelial and mesenchymal states reflective of their normal counterparts. *Stem Cell Rep*. 2014; 2:78–91.
- [10]. Lopez-Otin C, Hunter T. The regulatory crosstalk between kinases and proteases in cancer. *Nat. Rev. Cancer*. 2010; 10:278–292. [PubMed: 20300104]
- [11]. Nie S, Lo A, Zhu J, Wu J, et al. Isobaric protein-level labeling strategy for serum glycoprotein quantification analysis by liquid chromatography-tandem mass spectrometry. *Anal. Chem*. 2013; 85:5353–5357. [PubMed: 23638883]
- [12]. Collier TS, Sarkar P, Franck WL, Rao BM, et al. Direct comparison of stable isotope labeling by amino acids in cell culture and spectral counting for quantitative proteomics. *Anal. Chem*. 2010; 82:8696–8702. [PubMed: 20845935]
- [13]. Collier TS, Randall SM, Sarkar P, Rao BM, et al. Comparison of stable-isotope labeling with amino acids in cell culture and spectral counting for relative quantification of protein expression. *Rapid Commun. Mass Spectrom*. 2011; 25:2524–2532. [PubMed: 21818813]
- [14]. Nie S, Lo A, Wu J, Zhu J, et al. A glycoprotein biomarker for pancreatic cancer discovered by quantitative proteomics analysis. *J. Proteome Res*. 2014; 13:1873–1884. [PubMed: 24571389]
- [15]. Zhu J, Nie S, Wu J, Lubman DM. Target proteomic profiling of frozen pancreatic CD24+ adenocarcinoma tissues by immuno-laser capture microdissection and nano-LC-MS/MS. *J. Proteome Res*. 2013; 12:2791–2804. [PubMed: 23679566]
- [16]. He J, Zhu J, Liu Y, Wu J, et al. Immunohistochemical staining, laser capture microdissection, and filter-aided sample preparation-assisted proteomic analysis of target cell populations within tissue samples. *Electrophoresis*. 2013; 34:1627–1636. [PubMed: 23436586]
- [17]. Petrillo LA, Wolf DM, Kapoun AM, Wang NJ, et al. Xenografts faithfully recapitulate breast cancerspecific gene expression patterns of parent primary breast tumors. *Breast Cancer Res Treat*. 2012; 135:913–922. [PubMed: 22941572]
- [18]. Charafe-Jauffret E, Ginestier C, Iovino F, Wicinski J, et al. Breast cancer cell lines contain functional cancer stem cells with metastatic capacity and a distinct molecular signature. *Cancer Res*. 2009; 69:1302–1313. [PubMed: 19190339]

- [19]. Fillmore CM, Kuperwasser C. Human breast cancer cell lines contain stem-like cells that self-renew, give rise to phenotypically diverse progeny and survive chemotherapy. *Breast Cancer Res.* 2008; 10:R25. [PubMed: 18366788]
- [20]. Keller PJ, Lin AF, Arendt LM, Klebba I, et al. Mapping the cellular and molecular heterogeneity of normal and malignant breast tissues and cultured cell lines. *Breast Cancer Res.* 2010; 12:R87. [PubMed: 20964822]
- [21]. Eirew P, Kannan N, Knapp DJ, Vaillant F, et al. Aldehyde dehydrogenase activity is a biomarker of primitive normal human mammary luminal cells. *Stem Cells.* 2012; 30:344–348. [PubMed: 22131125]
- [22]. Eirew P, Stingl J, Raouf A, Turashvili G, et al. A method for quantifying normal human mammary epithelial stem cells with in vivo regenerative ability. *Nat. Med.* 2008; 14:1384–1389. [PubMed: 19029987]
- [23]. Lim E, Vaillant F, Wu D, Forrest NC, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat. Med.* 2009; 15:907–913. [PubMed: 19648928]
- [24]. Wisniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. *Nat. Methods.* 2009; 6:359–362. [PubMed: 19377485]
- [25]. Liu H, Sadygov RG, Yates JR 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 2004; 76:4193–4201. [PubMed: 15253663]
- [26]. Dong MQ, Venable JD, Au N, Xu T, et al. Quantitative mass spectrometry identifies insulin signaling targets in *C. elegans*. *Sciences (New York).* 2007; 317:660–663.
- [27]. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 2009; 4:44–57. [PubMed: 19131956]
- [28]. Huang da W, Sherman BT, Tan Q, Kir J, et al. DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* 2007; 35:W169–175. [PubMed: 17576678]
- [29]. Zhu J, Nie S, Wu J, Lubman DM. Target proteomic profiling of frozen pancreatic CD24+ adenocarcinoma tissues by immuno-laser capture microdissection and nano-LC-MS/MS. *J. Proteome Res.* 2012; 12:2791–2804. [PubMed: 23679566]
- [30]. Callister SJ, Barry RC, Adkins JN, Johnson ET, et al. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.* 2006; 5:277–286. [PubMed: 16457593]
- [31]. Savitski MM, Mathieson T, Zinn N, Sweetman G, et al. Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *J. Proteome Res.* 2013; 12:3586–3598. [PubMed: 23768245]
- [32]. Wright CA, Howles S, Trudgian DC, Kessler BM, et al. Label-free quantitative proteomics reveals differentially regulated proteins influencing urolithiasis. *Mol. Cell. Proteomics.* 2011; 10:M110 005686. [PubMed: 21474797]
- [33]. Azios NG, Krishnamoorthy L, Harris M, Cubano LA, et al. Estrogen and resveratrol regulate Rac and Cdc42 signaling to the actin cytoskeleton of metastatic breast cancer cells. *Neoplasia (New York).* 2007; 9:147–158.
- [34]. Man YG. Aberrant leukocyte infiltration: a direct trigger for breast tumor invasion and metastasis. *Int. J. Bio. Sci.* 2010; 6:129–132. [PubMed: 20224734]
- [35]. Clemens MJ. Translational regulation in cell stress and apoptosis. Roles of the eIF4E binding proteins. *J. Cell. Mol. Med.* 2001; 5:221–239. [PubMed: 12067482]
- [36]. Zeng C, Ke Z, Song Y, Yao Y, et al. Annexin A3 is associated with a poor prognosis in breast cancer and participates in the modulation of apoptosis in vitro by affecting the Bcl-2/Bax balance. *Exp. Mol. Pathol.* 2013; 95:23–31. [PubMed: 23631820]
- [37]. Hicks DG, Kulkarni S. HER2+ breast cancer: review of biologic relevance and optimal use of diagnostic tools. *Am. J. Clin. Pathol.* 2008; 129:263–273. [PubMed: 18208807]
- [38]. Li X, Lewis MT, Huang J, Gutierrez C, et al. Intrinsic resistance of tumorigenic breast cancer cells to chemotherapy. *J. Natl. Cancer Inst.* 2008; 100:672–679. [PubMed: 18445819]

- [39]. Thomson S, Petti F, Sujka-Kwok I, Mercado P, et al. A systems view of epithelial-mesenchymal transition signaling states. *Clin. Exp. Metastasis*. 2011; 28:137–155. [PubMed: 21194007]
- [40]. Kojima T, Sawada N. Regulation of tight junctions in human normal pancreatic duct epithelial cells and cancer cells. *Ann. NY Acad. Sci.* 2012; 1257:85–92. [PubMed: 22671593]
- [41]. Guido C, Whitaker-Menezes D, Lin Z, Pestell RG, et al. Mitochondrial fission induces glycolytic reprogramming in cancer-associated myofibroblasts, driving stromal lactate production, and early tumor growth. *Oncotarget*. 2012; 3:798–810. [PubMed: 22878233]
- [42]. Mimeault M, Batra SK. Hypoxia-inducing factors as master regulators of stemness properties and altered metabolism of cancer-and metastasis-initiating cells. *J. Mol. Med.* 2013; 17:30–54.
- [43]. Comi GP, Fortunato F, Lucchiari S, Bordoni A, et al. Beta-enolase deficiency, a new metabolic myopathy of distal glycolysis. *Ann. Neurol.* 2001; 50:202–207. [PubMed: 11506403]
- [44]. He KL, Deora AB, Xiong H, Ling Q, et al. Endothelial cell annexin A2 regulates polyubiquitination and degradation of its binding partner S100A10/p11. *J. Bio. Chem.* 2008; 283:19192–19200. [PubMed: 18434302]
- [45]. Yamamoto H, Sato H, Shibata S, Murata M, et al. Involvement of annexin V in the antiproliferative effect of GnRH agonist on cultured human uterine leiomyoma cells. *Mol. Human Reprod.* 2001; 7:169–173.
- [46]. Kenis H, van Genderen H, Bennaghmouch A, Rinia HA, et al. Cell surface-expressed phosphatidylserine and annexin A5 open a novel portal of cell entry. *J. Bio. Chem.* 2004; 279:52623–52629. [PubMed: 15381697]
- [47]. Song J, Shih Ie M, Chan DW, Zhang Z. Suppression of annexin A11 in ovarian cancer: implications in chemoresistance. *Neoplasia (New York)*. 2009; 11:605–614. 601 p following 614.

Figure 1. Schematic of experimental workflow. ALDH+, CD44+CD24- and CD49f-EpCAM+ cell populations were sorted from MCF7 cells. Then the extracted proteins from about 50 000 cells were digested by the FASP method. LC-MS/MS based on Orbitrap Elite MS and spectral counting methods were applied to analyze and quantify differentially expressed proteins among the three different fractions.

Figure 2.

Flow cytometry sorting process to isolate MCF7 CSCs and mature luminal cells. Left) CD44+CD24- cells were gated (P1) and sequentially gated for ALDH- (P2) and not CD49f-EpCAM+ (P3) cells. Middle) ALDH+ cells were gated (P4) and sequentially gated for not-CD44+CD24- (P5) and not CD49f-EpCAM+ (P6) cells. Right) CD49f-EpCAM+ cells were gated (P7) and sequentially gated for not-CD44+CD24- (P8) and ALDH- (P9) cells. The frequencies represent values based on each parental gate.

Figure 3. Proteome analysis of identified results. (A) The overlap of proteins identified from ALDH+, CD44+CD24- stem cell populations and CD49f- EpCAM+ non-stem cell population. Comparison of Gene Ontology annotations for cellular component (B) and biological process (C) of proteins identified among the three cell populations using DAVID bioinformatics database.

Figure 4.

(A) Histograms of log₂-transformed protein ratios for pairwise comparisons. The ratios for ALDH+/CD44+CD24- and ALDH+/CD49f-EpCAM+ show larger deviation than those for CD44+CD24-/CD49f-EpCAM+ comparison. (B) Volcano plot of ratios and *p*-value of the three pairwise comparisons. Proteins with *p* < 0.05 and above/below twofold changes are identified as proteins with significant changes. (C) Hierarchical cluster analysis of differentially expressed proteins among the three samples. The rows represent each protein and the column means different runs. Three replicates were clustered together. The clustering method for x and y axis used here is K-means clustering. Euclidean method was used to calculate distance.

Figure 5. Canonical signaling pathways enriched in differentially expressed proteins among ALDH+, CD44+CD24– stem cell populations and CD49f-EpCAM+ non-stem cell populations. A *p*-value threshold of 0.05 is applied (dotted line).

Figure 6. Top biological network generated by Ingenuity Pathway Analysis for differentially expressed proteins of (A) ALDH+/CD44+CD24-, (B) ALDH+/CD49f-EpCAM+ and (C) CD44+CD24-/CD49f-EpCAM+. Red and green represent over- and underexpressed proteins identified in this work; white indicates proteins that were not in the differentially expressed protein lists but related to the network. Solid arrows represent known direct interactions, and dotted arrows mean indirect interactions.

Figure 7.

(A) Western blotting analysis of HSPB1 and ANXA3 in the three cell samples. Beta-actin served as control. (B) Immunohistochemical staining analysis of ANXA3 in normal and breast cancer tissue microarrays. A representative figure is presented (Magnification 20 $\times$). In normal breast tissue, ANXA3 had weak, medium or strong expression level in 73%, 21% and 6% of samples, respectively. In breast cancer samples, ANXA3 had weak, medium or strong expression level in 20%, 32% and 48% of samples, respectively.