Published in final edited form as:

Proteomics Clin Appl. 2011 April ; 5(3-4): 179–188. doi:10.1002/prca.201000037.

Proteome and Transcriptome Profiles of a Her2/Neu-driven Mouse Model of Breast Cancer

Regine M. Schoenherr^{‡,1}, Karen S. Kelly-Spratt^{‡,1}, ChenWei Lin^{‡,1}, Jeffrey R. Whiteaker^{‡,1}, Tao Liu^{‡,2}, Ted Holzman¹, Ilsa Coleman¹, Li-Chia Feng¹, Travis D. Lorentzen¹, Alexei L. Krasnoselsky¹, Pei Wang¹, Yan Liu¹, Kay E. Gurley¹, Lynn M. Amon¹, Athena A. Schepmoes², Ronald J. Moore², David G. Camp II², Lewis A. Chodosh³, Richard D. Smith², Peter S. Nelson¹, Martin W. McIntosh^{*,1}, Christopher J. Kemp^{*,1}, and Amanda G. Paulovich^{*,1}

¹Fred Hutchinson Cancer Research Center, Seattle, Washington 98109

²Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354

³Department of Cancer Biology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104

Abstract

Purpose—We generated extensive transcriptional and proteomic profiles from a Her2-driven mouse model of breast cancer that closely recapitulates human breast cancer. This report makes these data publicly available in raw and processed forms, as a resource to the community. Importantly, we previously made biospecimens from this same mouse model freely available through a sample repository, so researchers can obtain samples to test biological hypotheses without the need of breeding animals and collecting biospecimens.

Experimental design—Twelve datasets are available, encompassing 841 LC-MS/MS experiments (plasma and tissues) and 255 microarray analyses of multiple tissues (thymus, spleen, liver, blood cells, and breast). Cases and controls were rigorously paired to avoid bias.

Results—In total, 18,880 unique peptides were identified (PeptideProphet peptide error rate ≤1%), with 3884 and 1659 non-redundant protein groups identified in plasma and tissue datasets, respectively. Sixty-one of these protein groups overlapped between cancer plasma and cancer tissue.

Conclusions and clinical relevance—These data are of use for advancing our understanding of cancer biology, for software and quality control tool development, investigations of analytical variation in MS/MS data, and selection of proteotypic peptides for MRM-MS. The availability of these datasets will contribute positively to clinical proteomics.

Keywords

Breast cancer; Her2; mouse; proteome; transcriptome

*Corresponding authors: Dr. Amanda G. Paulovich, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, WA 98109, USA. apaulovi@fhcrc.org, Office 206-667-1912, Fax 206-667-2277. Dr. Christopher J. Kemp, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, WA 98109, USA. cjtemp@fhcrc.org, Office 206-667-4075, Fax 206-667-5815. Dr. Martin W. McIntosh, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, WA 98109, USA. mmcintos@fhcrc.org, Office 206-667-4612, Fax 206-667-7264..

[‡]These authors contributed equally.

The authors have declared no conflict of interest.

Discovering and verifying cancer biomarkers directly in human samples is tremendously difficult due to considerable genetic, behavioral, and environmental heterogeneity. Mouse model studies, on the other hand, can be conducted under stringent genetic and environmental control, which has proven beneficial when investigating the fundamentals of cancer biology, developing and evaluating therapeutic agents, and developing and refining technologies for biomarker discovery and verification.[1-5] Mouse models have also become more sophisticated in recent years, moving from cell lines to xenografts to genetically engineered mice, and they can mimic human cancers to an ever greater extent. The application of mouse models is therefore highly promising for expanding our understanding of cancer biology and thereby moving closer to enhancing the diagnosis, prognosis, and treatment of cancer.

In line with these goals, we have generated molecular profiles from a Her2/Neu breast cancer mouse model, in which an activated Neu oncogene is conditionally expressed in the mammary epithelium of bitransgenic mice *via* the tetracycline regulatory system, and in which mammary carcinomas develop with 100% penetrance.[6] We have previously demonstrated that samples from this mouse model can be successfully used to evaluate mass spectrometry- and immunoaffinity-based technologies and to discover and confirm biomarkers of breast cancer in mice, with some of our findings being translatable to human biomarker candidates.[5] The data presented here encompass six proteome and six transcriptome datasets, and we are making the raw and processed data freely available to the public as a resource. This is in concordance with recommendations developed in a 2008 International Summit regarding the release and sharing of proteomics data (the Amsterdam principles).[7]

The diversity and depth of data derived from the experiments reported here can serve several purposes: i) The data may aid in shedding further light into cancer biology and assist in the search of biomarkers for breast cancer. We note that the samples used in this work were obtained from a mouse model repository[8] that we established as part of the NCI Mouse Proteomic Technologies Initiative (<http://proteomics.cancer.gov/programs/mouse/overview.asp>); from this repository, plasma and tissue samples from 300 rigorously paired case and control mice are available upon request from the Her2/Neu mice, and researchers can use these to test biological hypotheses generated from the data. ii) Since the number of liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based quantitative proteomics studies has rapidly increased in recent years, the need for standardization has been recognized[9] and growing efforts have been directed towards achieving such standardization and towards developing a suite of quality control tools.[10-14] These publicly available data include extensive LC-MS/MS technical replicates of the same sample, and these can be mined and used to refine and expand such tools. iii) Targeted mass spectrometry (MS) methods such as multiple reaction monitoring-MS (MRM-MS) are emerging as an important means of verifying the hundreds-to-thousands of biomarker candidates that have resulted from biomarker discovery efforts; [5,15-17]; however, it is still not trivial to predict which peptides of a specific protein are the best proteotypic peptides, that is, peptides of a specific protein that are robustly detected by MS, and which combinations of parent and fragment ion mass-to-charge values (transitions) will yield optimal sensitivity. The extensive catalog of peptide MS/MS spectra provided here will provide empirical data for murine spectral libraries and will also support selection of proteotypic peptides and transitions for the development of MRM-MS-based methods.

A detailed description of the materials and methods used can be found in Supporting Information 1, while a summary of the methods is given here. All mouse work was performed under IACUC regulations as approved by the Fred Hutchinson Cancer Research Center's animal use committee (accreditation number IR 1311). A previously described

doxycycline-inducible, bitransgenic MMTV-rtTA/TetO-NeuNT (Her2/Neu) mouse model[6] was used. Corresponding control mice were transgenic for TetO-NeuNT only and were littermates of the bitransgenic mice, and the tumor-bearing and control mice were carefully paired with respect to cage and environmental conditions to minimize noise and bias. In this model, an activated form of the rat Her2 oncogene is conditionally expressed in the mammary epithelium under control of a doxycycline regulatable element. When doxycycline is added to the drinking water, 100% of the mice develop a breast cancer that closely mimics the human disease. Disease progression starts with a premalignancy, followed by a localized cancer, then local invasiveness, and finally distant metastasis to the lungs. Just as when human patients are treated with the monoclonal antibody Herceptin, which targets the Her2 gene product, the mouse tumors regress when doxycycline is removed from the drinking water and the oncogene is turned off. For these studies, all mice received doxycycline (2 mg/mL + 5% sucrose) in the drinking water starting at 8 weeks of age. The biospecimens for ten of twelve datasets were collected when the tumor size had reached 1 cm (Table 1 and Supporting Figure 1 in **Supporting Information 2**); at this point the tumors were clearly noticeable, yet the mice did not appear to be otherwise ill. (Each tumor-bearing mouse and its corresponding control littermate were euthanized by CO₂ inhalation.) Additionally, for the remaining two datasets (the 3 day MARS and 6 day MARS experiments), biospecimens were collected at either 3 days or 6 days following withdrawal of the doxycycline, when tumors were visibly regressing[6] (Supporting Figure 1, **Supporting Information 2**). Whole blood was collected by cardiac puncture, and plasma was isolated by centrifugation. Aliquots were transferred to cryovials and frozen in liquid nitrogen. Tissue samples were collected when the tumors reached ~1 cm size and the samples were snap-frozen in liquid nitrogen. Although cellularity can vary amongst tumor tissues, the tumors arising in this particular mouse model are densely epithelial, as shown in Supporting Figure 2 in **Supporting Information 2**.

For the breast tissue proteome dataset (Table 1), normal and tumor tissues were harvested and processed separately from 10 control and 10 tumor-bearing mice, see the right-most panel of Figure 1A for a sample processing overview (the underlined words in Figure 1 give the uniquely identifying names of the datasets, see also Tables 1 and 2). Briefly, the tissues were homogenized and normal and tumor lysates were separately pooled by equal mass. The pools were denatured using methanol, reduced with dithiothreitol, alkylated, trypsin digested, dried, and finally resuspended in ammonium bicarbonate to an estimated 6 pmol/ μ L (0.3 μ g/ μ L) original protein concentration based on a Bradford QuickStart Assay (Bio-Rad Laboratories, Hercules, CA). Further details on the methods are given in Supporting Information 1.

In addition to the tissue samples, a total of ten mouse plasma pools were generated for this study. Figures 1A (the three left-most panels) and **1B** illustrate the numbers of mice comprising each plasma pool and the sample processing workflows for the different datasets. For the plasma samples in Figure 1A, pooling was based on equal protein mass from each mouse, and each pool was separately depleted of the three most abundant proteins (albumin, IgG, and transferrin) using an MS-3 column of the mouse Multiple Affinity Removal System (MARS, Agilent Technologies, Santa Clara, CA). For the MARS, 3 day MARS, and 6 day MARS samples, MARS-depleted plasma pools were separately denatured, reduced, alkylated, and trypsin-digested using the same methanol-based method as was used for the tissue samples above. For samples that were fractionated by strong cation exchange chromatography (SCX) in Figure 1A, the MARS-depleted plasma pools were separately denatured, reduced, alkylated, and trypsin-digested using a urea-based method, which was followed by a desalting step. Four SCX separations, each using 400 μ g of digested plasma based on total original protein concentration, were performed for each of four sample pools (2 cancer pools and 2 normal pools). A total of 14 pooled fraction samples

resulted for each sample pool. Pilot LC-MS/MS experiments determined that the four SCX pooled fractions with the highest numbers of unique peptide identifications (with PeptideProphet scores ≥ 0.95) were pooled fractions 2, 3, 8, and 9, and these fractions were subsequently analyzed by LC-MS and LC-MS/MS. Please refer to Supporting Information 1 for further experimental details.

For the plasma samples in Figure 1B, 40 μ L of plasma of each of 10 tumor-bearing mice and of 10 control Her2/Neu mice were pooled separately and then processed independently. The plasma pools were immunoaffinity depleted using an MS-3 MARS column, and both the flow-through and bound fractions of the MARS depletion were trypsin digested. The flow-through fraction digest was further processed by cysteinyl peptide enrichment as previously described[18, 19]. Briefly, the peptides resulting from the MARS flow-through protein digest were reduced and the sample was subsequently incubated with Thiopropyl Sepharose 6B thiol-affinity resin for 1 hour. The unbound, non-cysteinyl containing peptides were collected as flow-through of the column, and the cysteinyl-containing peptides were washed on-column and subsequently eluted using dithiothreitol (DTT). The cysteinyl-containing peptides were alkylated and both samples were desalted. Subsequently, 300 μ g of peptides (either the tryptic digest of MARS bound proteins or the cysteinyl peptides or non-cysteinyl peptides of MARS flow-through proteins) were separately fractionated by SCX. Up to 37 fractions were collected for each peptide population, and each fraction was analyzed separately by reversed-phase capillary LC-MS/MS. Additional details are given in Supporting Information 1.

Two types of mass spectrometry analyses were performed for the datasets given in Figure 1A: LC-MS experiments to normalize peptide loading between different samples of a dataset prior to LC-MS/MS analyses, and LC-MS/MS shotgun experiments. For the former, an Agilent 1100 system was connected to an LCT Premier time-of-flight (TOF) mass spectrometer (Waters Corporation, Milford, MA). LC-MS experiments of the cancer and normal samples were performed, and msInspect[20] was used to determine the number of peptide features in each experiment and the distribution of these peptide features' signal intensities (with a peptide feature constituting a peptide's isotopic envelope with a particular charge state). Normalization of sample concentration for each dataset was achieved by adjusting the sample injection volumes until the median feature intensities were aligned across cancer and normal samples. Approximately 2 μ g of digested protein was subsequently injected for each LC-MS/MS experiment. Data-dependent shotgun LC-MS/MS analyses of the samples were performed with an Agilent 1100 system connected to a linear ion trap mass spectrometer (LTQ, Thermo Scientific, Waltham, MA). The 5 most abundant ions of an MS scan were selected for MS/MS fragmentation. See Supporting Information 1 for further details.

For the plasma proteome dataset given in Figure 1B, LC-MS/MS analyses were performed using a custom-built high-pressure capillary LC system[21] coupled on-line to a linear ion trap mass spectrometer (LTQ; ThermoElectron) *via* an in-house-manufactured electrospray ionization interface, see also Supporting Information 1. Here, each MS scan was followed by MS/MS scans of the 10 most abundant ions.

The MARS, MARS+SCX, 3 day MARS, and 6 day MARS plasma proteome samples, as well as the breast tissue proteome samples (Table 1), underwent only limited sample processing to yield relatively few samples for analysis by LC-MS/MS, which made it practical to analyze the samples using multiple technical repeat LC-MS/MS injections. Acquiring data based on multiple technical repeats increases the number of peptide and protein identifications and can aid in downstream quantitative analysis based on, for example, spectral counting approaches.[22,23] For the MARS dataset, the plasma pools

were MARS-depleted, digested, and analyzed by LC-MS/MS. To allow for greater sampling depth yet still be able to acquire multiple technical repeats, we also generated the MARS +SCX dataset for which the same plasma was not only depleted and digested, but also fractionated by strong cation exchange, and four (of 14) fraction pools that contained the greatest number of peptides (determined by preliminary LC-MS/MS experiments, data not shown) were analyzed by LC-MS/MS. In addition to these datasets, the aim of the MARS +Cys+SCX dataset was to obtain exquisite sampling depth and thereby as broad a molecular characterization of mouse plasma as possible. To this end, both the MARS bound and the MARS flow-through fractions were analyzed. The proteome coverage, especially of presumably low abundance proteins, was further enhanced by performing cysteinyl peptide enrichment[18,19] on the MARS flow-through fraction samples. The three resulting sample types, the MARS-bound fraction and the cysteinyl and non-cysteinyl peptide fractions of the MARS flow-through, were in turn subjected to SCX fractionation to even further reduce the complexity of each sample that would be analyzed by LC-MS/MS, and thereby obtain greater proteome coverage. A total of 841 LC-MS/MS runs were performed for the six proteome datasets, see Table 2.

All LC-MS/MS shotgun data were submitted to database searching using X!Tandem[24,25] including a scoring plug-in[26] compatible with PeptideProphet[27] to obtain peptide identifications. The data processing and storage were handled *via* the Computational Proteomics Analysis System (CPAS)[28], and the search parameters used were (i) tryptic enzyme constraint, (ii) up to 2 missed cleavages, (iii) ± 2 Da MH⁺ mass tolerance, (iv) alkylated cysteine as fixed modification, and (v) oxidized methionine as variable modification. Non-tryptic peptides were not included. All data were searched against the mouse International Protein Index (IPI) database version 3.65 (containing 56,775 IPI entries) that was released on October 2, 2009. (Since searching all LC-MS/MS runs together in CPAS was not possible due to the large number of runs (841 runs total), the data were split into two groups for searching.) The resulting peptide identifications were filtered by PeptideProphet score to maintain an overall PeptideProphet peptide error rate (false discovery rate, FDR) $\leq 1\%$ (that is, the $\leq 1\%$ error rate cut-off was applied at the peptide level). The corresponding PeptideProphet cutoffs for the searches were: search group 1, 0.92, search group 2, 0.95. For each peptide passing the $\leq 1\%$ error rate filter, all corresponding IPI numbers and Entrez Gene IDs were identified by matching the peptide sequence to all sequences in the mouse IPI database version 3.65. Note that only IPIs associated with a reported sequence in the IPI database were used in this mapping; that is, additional (alias) IPIs listed in each record for the IPI were not used. To reduce protein redundancy and ambiguity due to single peptides mapping to multiple proteins, the data from each separate dataset were further analyzed by using ProteinProphet[29], which was used as a clustering tool to yield non-redundant protein groups.[30,31] To this end, the probability scores of peptides with PeptideProphet error rate cut-offs $\leq 1\%$ were set equal to 1.0, and ProteinProphet analysis was subsequently performed on only those peptides. (The resulting protein groups had protein probabilities that were either very high (≥ 0.98) or zero (due to only high-confidence peptides having been included; no ProteinProphet error rates are associated with the protein groups when using this particular analysis). The protein groups having zero probability represented redundant protein groups, whereas the protein groups with ≥ 0.98 probability represented non-redundant protein groups.) An average sequence coverage percentage was calculated from the sequence coverages of the individual IPIs in a protein group. The sequence coverage of individual IPIs was calculated as the percentage of the sum of amino acids of identified unique peptide sequences (with PeptideProphet error rate $\leq 1\%$) of an IPI to the total amino acid length of the IPI. Table 2 gives a summary of the numbers of peptide and non-redundant protein group identifications, and Supporting Information 3 includes complete lists of protein group identifications for each of the proteome datasets along with information on the numbers of peptides identified

for each protein group, average sequence coverages, and gene identifications. As expected, the MARS+Cys+SCX dataset yielded higher numbers of non-redundant protein groups than the other plasma datasets that underwent less fractionation (Table 2). Of note is that the numbers of peptide and protein group identifications for the MARS datasets are relatively modest when compared with e.g. the 3 and 6 day MARS datasets (even though the same sample preparation procedure was followed for these three datasets (Figure 1A)). This is most likely due to the use of different MS acquisition parameters for the 3 and 6 day MARS datasets than for the MARS datasets, see Supporting Information 1; an increased MS scan range (m/z 300-1600 vs. m/z 400-1600) for the 3 and 6 day MARS vs. MARS data allowed peptides with lower m/z values to be sampled, and shorter scan times and fewer microscans resulted in the acquisition of an increased number of MS and MS/MS scans for the 3 and 6 day MARS data. Of note is also that the numbers of peptides and protein groups in the tissue dataset (Table 2) are markedly higher in the tumor tissue than in normal tissue. We observed similarly disproportionate numbers in our earlier work that was based on tissue samples from a different cohort of mice,[5] and hypothesize that these numbers reflect the greater cellularity of the tumor samples (see also Supporting Figure 2 in **Supporting Information 2**) when compared with normal tissue.

An evaluation of the global overlap between protein groups detected in cancer and normal samples from all proteome datasets (based on Supporting Information 3) is given in **Panel A** of Supporting Figure 3 (**Supporting Information 2**). A total of 1,918 protein groups were found exclusively in the cancer samples, and a break-down of these proteins between plasma and tissue is shown in **Panel B**, Supporting Figure 3 (**Supporting Information 2**). Of these, 61 protein groups were overlapping between cancer plasma and cancer tissue and could be further investigated as potential biomarker candidates as blood-based cancer diagnostics.

Another analysis that would provide an initial, crude indication of proteins that were differentially detected in normal and cancer samples would be to assess the fold differences of the numbers of unique peptides for each protein group in Supporting Information 3. For example, 15 unique peptides were identified for IPI00830803 of Fibulin 2 (Fbln2) in cancer tissue, whereas none were identified in normal tissue. Similar numbers were obtained for the Fbln2 protein group containing IPI00132067 and IPI00750260 (14 and 0 unique peptides in cancer and normal tissue, respectively). Indeed, Fibulin 2 is a biomarker that we verified in our previous work in this same mouse model.[5]

More extensive quantitative analyses of the raw data, such as with spectral counting approaches[22,23] or with algorithms such as SASPECT[5], will be able to provide false discovery rate assessments and an indication as to which of the differentially detected proteins are attributable to the neoplastic nature of the tissue and which simply due to the difficulty of identifying the same protein in two or more replicates of the same sample. This difficulty is inherent to untargeted shotgun mass spectrometry experiments, in which undersampling of a complex biological sample by the mass spectrometer results in some lack of reproducibility when the same sample is analyzed multiple times, see Supporting Figure 4 (**Supporting Information 2**). Undersampling in shotgun experiments has been discussed by Tabb *et al.*[12] and Whiteaker *et al.*[5], and algorithms such as SASPECT were specifically designed to accommodate the undersampling problem.

In addition to the proteome datasets, six datasets were generated based on transcriptional microarray analyses, see Table 1. Five of these analyzed whole tissue samples, while one analyzed epithelial cells from laser capture microdissected (LCM) tumor and normal tissues. The whole tissue analyses were performed on several tissues from both tumor-bearing and control Her2/Neu model mice. Specifically, a cohort of 25 tumor-bearing and 25 rigorously matched control female mice were treated with doxycycline for an average of 18 weeks.

Mice were sacrificed when mammary tumors reached $\geq 1 \text{ cm}^3$ size in the tumor-bearing mice, and tissues were snap-frozen and stored in liquid nitrogen until use. RNA samples were prepared for 5 tissues in all 50 mice: thymus, spleen, liver, blood cells (blood with plasma fraction removed), and breast, and a total of 250 individual microarray analyses were performed using Affymetrix GeneChip Mouse Genome 430 2.0 arrays (Affymetrix, Santa Clara, CA). The analyses of laser capture microdissected epithelium were based on only breast tissue. Specifically, neoplastic epithelial cells from breast tissue were captured and analyzed; hence, there is no contribution of normal tissue or stroma in the tumor-bearing samples that would confound the results. Approximately 5000 epithelial cells from each of three Her2/Neu case mice and benign breast from two strain-matched control mice were separately laser capture microdissected using the Arcturus Veritas Microdissection System as previously described.[32] RNA extraction and amplification of LCM samples as well as a reference standard RNA for use in two-color oligonucleotide arrays were performed by using standard procedures described in Supporting Information 1. Subsequent microarray analyses were performed using Agilent 44K whole mouse genome expression oligonucleotide microarray slides. Five microarray analyses were performed in total for the LCM experiments. See Supporting Information 1 for further details on all microarray analyses. The microarray data presented here focused on the transcriptional profiling of tissues; however, analyses of RNA in the circulation, specifically also of microRNAs,[33] have become possible in recent years and such analyses could provide additional biological information on this mouse model.

To make the data as widely accessible and useful to the public as possible, the data are available on three levels:

1. The raw and mzXML data files of all proteomics experiments have been deposited in the TRANCHE data repository (<https://proteomecommons.org/tranche/>) and are publicly available. **Supporting Information 4** provides a list of the identifying TRANCHE hashes for each of the datasets. Readers can download the data and perform data processing using a search algorithm of their choice, such as SEQUEST or Mascot. The raw and mzXML data can aid in the development of statistical quality control and data analysis tools, the need for which continues to expand,[7,9-13] and can be used for alternative data analyses[34]. We have also uploaded the raw transcript profiles onto Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>) to enable alternative and additional processing of those data. The accession numbers for the whole and the LCM tissue datasets are GSE20465 and GSE20280, respectively.
2. We are making the searched proteome data available on the CPAS data repository system[28] (<https://proteomics.fhcrc.org/CPAS>) under the “Published Experiments” folder in the “Her2-Neu mouse breast 2010” subfolder. The data in the “MS2 Experiment Runs” section includes results such as peptide sequences, peptide spectra, and PeptideProphet scores, and the data can be filtered and searched in a variety of ways. We also deposited PeptideProphet and ProteinProphet output files in the same subfolder on CPAS, in the “Files” section. (Of note is that the ProteinProphet output files in this latter section are the output files of our ProteinProphet analysis for which we used only peptides with PeptideProphet error rate cut-offs $\leq 1\%$, and whose primary purpose was to use ProteinProphet as a clustering tool to yield non-redundant protein groups. This was a more stringent analysis than the ProteinProphet analysis that is automatically performed by CPAS, which uses peptides with PeptideProphet probability scores ≥ 0.20 (corresponding here to an $\sim 20\%$ error rate cut-off), and whose results are contained in the data of the “MS2 Experiment Runs” section.) The peptide sequences and spectra can be mined to facilitate downstream MS-based verification

of proteins of interest by quantitative selected or multiple reaction monitoring (SRM- or MRM-MS) experiments, similar to the data provided by Yang and Lazar[35] and Kline *et al.*[17] for human samples. An analysis that might aid in increasing the confidence in selecting proteotypic peptides that would perform well in MRM assays might involve comparing whether identical peptides were identified in the different datasets, especially in relation to proteins identified with only one, two, or three peptides.

3. All processed proteome and transcriptome results (based on the ProteinProphet analyses that used only peptides with PeptideProphet error rates $\leq 1\%$, see above, and the microarray statistical analyses) have been aggregated into one spreadsheet, see Supporting Information 3, and it allows queries to be performed across all datasets. The table includes information such as protein group identifications, average protein group sequence coverages, peptide counts, and gene IDs for the proteome data, in addition to log2 ratios, FDR, and q-values that indicate differential expression for the transcriptome data. Since the table is in a spreadsheet format, it can be easily queried or filtered in a variety of ways, and comprehensive information from all datasets can be obtained. The table also includes human ortholog and gene ontology (GO) cellular localization annotations where possible.

In conclusion, mouse models of cancer have come to play an important role in understanding cancer biology and developing technologies in the quest for identifying biomarkers since their genetic backgrounds and environmental factors can be tightly controlled. Characterizing the molecular make-up of mouse models can aid tremendously in being able to take full advantage of the information they present. To this end, several laboratories collaborated to amass the substantial amount of proteome and transcriptome data presented here, which we now make available to the public without restrictions. We envision that the information we provide, from raw to processed data, will be a useful resource for researchers from many different fields, from statisticians, to chemists, to biologists. The combined resource of these datasets, coupled to our repository of high quality biospecimens from this Her2 mouse breast cancer model[8], provide fertile grounds for facile testing of biological hypotheses.

Statement of Clinical Relevance

Characterization of mouse models of human diseases has had tremendous clinical impact. Studies in the mouse can be conducted under stringent genetic and environmental control, which has proven beneficial when investigating cancer biology, developing therapeutic agents, and developing technologies for biomarker discovery and verification. [1-5] Mouse models have become more sophisticated in recent years, moving from xenografts to genetically engineered mice, and they can mimic human cancers to an ever greater extent. The application of mouse models is therefore highly promising for expanding our understanding of cancer biology and thereby moving closer to enhancing the diagnosis, prognosis, and treatment of cancer.

In line with these goals, we have generated extensive transcriptional and proteomic profiles from a Her2-driven mouse model of breast cancer that closely recapitulates the pathology and natural history of human breast cancer. The purpose of this report is to make all of these data publicly available in raw and processed forms, as a resource to the community. Importantly, we have also made high quality biospecimens from this same mouse model freely available through a sample repository, so researchers can readily obtain samples to test biological hypotheses without the need of breeding animals and collecting biospecimens.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by NCI / SAIC subcontract 23XS144A to Drs. Amanda G. Paulovich and Martin W. McIntosh, by a grant from The Paul G. Allen Family Foundation to Drs. Peter S. Nelson, Martin W. McIntosh, Christopher J. Kemp, and Amanda G. Paulovich, by the Entertainment Industry Foundation (EIF) and the EIF Women's Cancer Research Fund to the Breast Cancer Biomarker Discovery Consortium (Dr. Amanda G. Paulovich), and by generous gifts from the Keck Foundation and the Canary Foundation to Dr. Amanda G. Paulovich. Portions of this research were supported by the NIH National Center for Research Resources (RR018522; Richard D. Smith, PI), and the Environmental Molecular Science Laboratory [a national scientific user facility sponsored by the U.S. Department of Energy (DOE) Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory (PNNL)]. PNNL is operated by Battelle Memorial Institute for the DOE under contract DE-AC05-76RLO-1830. We thank Dr. Todd R. Golub, Casey Gates, and Diane Gage of the Broad Institute's microarray facility for generating microarray data. We thank members of our advisory board for advice: Eleftherios P. Diamandis, Ruedi Aebersold, Leigh N. Anderson, Nicole D. Urban, Valeri I. Vasioukhin, Frederick Appelbaum, Mark T. Groudine, James Roberts, and Daniel E. Gottschling. We thank all members of the Allen Project team: Lee Hartwell, Samir M. Hanash, Sharon J. Pitteri, Hong Wong, Denny Liggitt, Daniel B. Martin, Ted Whitmore, Amelia Peterson, Robyn L. Prueitt, Matthew Fitzgibbon, Jimmy K. Eng, Damon H. May, Abby Stimmel, Yuzheng Zhang, Sara L. Zriny, Ruth Dumpit, Mary M. Trute, Liming Hou, Izabela Sokal, Jacob Kennedy, and Uliana J. Voytovich. We also thank Matthew Monroe and Samuel Purvine for their assistance.

Abbreviations

CPAS	Computational Proteomics Analysis System
FDR	false discovery rate
IPI	International Protein Index
LCM	laser capture microdissection
LTD	linear ion trap MS
MARS	Multiple Affinity Removal System
SCX	strong cation exchange chromatography

References

1. Quagliano E, Mastini C, Forni G, Cavallo F. ErbB2 transgenic mice: a tool for investigation of the immune prevention and treatment of mammary carcinomas. *Curr. Protoc. Immunol.* 2008 *Chapter* 20, Unit 20.9.1-20.9-10.
2. Bouwman P, Jonkers J. Mouse models for BRCA1 associated tumorigenesis: from fundamental insights to preclinical utility. *Cell. Cycle.* 2008; 7:2647–2653. [PubMed: 18728395]
3. Frese KK, Tuveson DA. Maximizing mouse cancer models. *Nat. Rev. Cancer.* 2007; 7:645–658. [PubMed: 17687385]
4. Vargo-Gogola T, Rosen JM. Modelling breast cancer: one size does not fit all. *Nat. Rev. Cancer.* 2007; 7:659–672. [PubMed: 17721431]
5. Whiteaker JR, Zhang H, Zhao L, Wang P, et al. Integrated pipeline for mass spectrometry-based discovery and confirmation of biomarkers demonstrated in a mouse model of breast cancer. *J. Proteome Res.* 2007; 6:3962–3975. [PubMed: 17711321]
6. Moody SE, Sarkisian CJ, Hahn KT, Gunther EJ, et al. Conditional activation of Neu in the mammary epithelium of transgenic mice results in reversible pulmonary metastasis. *Cancer. Cell.* 2002; 2:451–461. [PubMed: 12498714]
7. Rodriguez H, Snyder M, Uhlen M, Andrews P, et al. Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: the Amsterdam principles. *J. Proteome Res.* 2009; 8:3689–3692. [PubMed: 19344107]

8. Kelly-Spratt KS, Kasarda AE, Igra M, Kemp CJ. A mouse model repository for cancer biomarker discovery. *J. Proteome Res.* 2008; 7:3613–3618. [PubMed: 18624399]
9. Mischak H, Apweiler R, Banks RE, Conaway M, et al. Clinical proteomics: a need to define the field and to begin to set adequate standards. *Proteomics Clin Appl.* 2007; 1:148–156. [PubMed: 21136664]
10. Xu H, Freitas MA. Automated diagnosis of LC-MS/MS performance. *Bioinformatics.* 2009; 25:1341–1343. [PubMed: 19304874]
11. Mirzaei H, Brusniak MY, Mueller LN, Letarte S, et al. Halogenated peptides as internal standards (H-PINS): introduction of an MS-based internal standard set for liquid chromatography-mass spectrometry. *Mol. Cell. Proteomics.* 2009; 8:1934–1946. [PubMed: 19411281]
12. Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* 2010; 9:761–776. [PubMed: 19921851]
13. Paulovich AG, Billheimer D, Ham AJ, Vega-Montoto L, et al. Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell. Proteomics.* 2010; 9:242–254. [PubMed: 19858499]
14. Rudnick PA, Clauser KR, Kilpatrick LE, Tchekhovskoi DV, et al. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol. Cell. Proteomics.* 2010; 9:225–241. [PubMed: 19837981]
15. Kuhn E, Wu J, Karl J, Liao H, et al. Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and ¹³C-labeled peptide standards. *Proteomics.* 2004; 4:1175–1186. [PubMed: 15048997]
16. Keshishian H, Addona T, Burgess M, Mani DR, et al. Quantification of cardiovascular biomarkers in patient plasma by targeted mass spectrometry and stable isotope dilution. *Mol. Cell. Proteomics.* 2009; 8:2339–2349. [PubMed: 19596694]
17. Kline KG, Frewen B, Bristow MR, Maccoss MJ, Wu CC. High quality catalog of proteotypic peptides from human heart. *J. Proteome Res.* 2008; 7:5055–5061. [PubMed: 18803417]
18. Liu T, Qian WJ, Strittmatter EF, Camp DG 2nd, et al. High-throughput comparative proteome analysis using a quantitative cysteinyl-peptide enrichment technology. *Anal. Chem.* 2004; 76:5345–5353. [PubMed: 15362891]
19. Liu T, Qian WJ, Chen WN, Jacobs JM, et al. Improved proteome coverage by using high efficiency cysteinyl peptide enrichment: the human mammary epithelial cell proteome. *Proteomics.* 2005; 5:1263–1273. [PubMed: 15742320]
20. Bellew M, Coram M, Fitzgibbon M, Igra M, et al. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics.* 2006; 22:1902–1909. [PubMed: 16766559]
21. Livesay EA, Tang K, Taylor BK, Buschbach MA, et al. Fully automated four-column capillary LC-MS system for maximizing throughput in proteomic analyses. *Anal. Chem.* 2008; 80:294–302. [PubMed: 18044960]
22. 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]
23. Ishihama Y, Oda Y, Tabata T, Sato T, et al. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics.* 2005; 4:1265–1272. [PubMed: 15958392]
24. Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* 2003; 17:2310–2316. [PubMed: 14558131]
25. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics.* 2004; 20:1466–1467. [PubMed: 14976030]
26. MacLean B, Eng JK, Beavis RC, McIntosh M. General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics.* 2006; 22:2830–2832. [PubMed: 16877754]

27. 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; 74:5383–5392. [PubMed: 12403597]
28. Rauch A, Bellew M, Eng J, Fitzgibbon M, et al. Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* 2006; 5:112–121. [PubMed: 16396501]
29. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2003; 75:4646–4658. [PubMed: 14632076]
30. Petyuk VA, Qian WJ, Hinault C, Gritsenko MA, et al. Characterization of the mouse pancreatic islet proteome and comparative analysis with other mouse tissues. *J. Proteome Res.* 2008; 7:3114–3126. [PubMed: 18570455]
31. Zhang X, Zhou JY, Chin MH, Schepmoes AA, et al. Region-specific protein abundance changes in the brain of MPTP-induced Parkinson's disease mouse model. *J. Proteome Res.* 2010; 9:1496–1509. [PubMed: 20155936]
32. True L, Coleman I, Hawley S, Huang CY, et al. A molecular correlate to the Gleason grading system for prostate adenocarcinoma. *Proc. Natl. Acad. Sci. U. S. A.* 2006; 103:10991–10996. [PubMed: 16829574]
33. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, et al. Circulating microRNAs as stable blood-based markers for cancer detection. *Proc. Natl. Acad. Sci. U. S. A.* 2008; 105:10513–10518. [PubMed: 18663219]
34. Menon R, Omenn GS. Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers. *Cancer Res.* 2010; 70:3440–3449. [PubMed: 20388783]
35. Yang X, Lazar IM. MRM screening/biomarker discovery with linear ion trap MS: a library of human cancer-specific peptides. *BMC Cancer.* 2009; 9:96. [PubMed: 19327145]

Figure 1. Sample processing workflows for the different proteome datasets

A. Sample preparation for the MARS, MARS+SCX, 3 day MARS, 6 day MARS plasma datasets, and the breast tissue dataset. **B.** Sample preparation for the MARS+Cys+SCX plasma dataset. The underlined words are the uniquely identifying names of the datasets (see also Tables 1 and 2).

Table 1
Overview of the proteome and transcriptome datasets from the Her2/Neu breast cancer mouse model

Note that exactly the same plasma pools were used for the MARS and MARS+SCX datasets. None of the mice for the other datasets were the same.

Plasma Proteome Datasets		
Dataset Name	Sample type of tumor-bearing mice	Experiment
MARS ^a	1 cm tumor	MARS depletion + trypsin digestion + LC-MS/MS
MARS+SCX ^a	1 cm tumor	MARS depletion + trypsin digestion + SCX fractionation + LC-MS/MS
MARS+Cys+SCX	1 cm tumor	3 sample preparation workflow branches: ^b
		1) trypsin digestion of MARS-bound fraction + SCX fractionation + LC-MS/MS
		2) MARS depletion + trypsin digestion + SCX fractionation of non-cysteiny1 peptides + LC-MS/MS
		3) MARS depletion + trypsin digestion + SCX fractionation of cysteiny1 peptides + LC-MS/MS
3 day MARS	3 day regression	MARS depletion + trypsin digestion + LC-MS/MS
6 day MARS	6 day regression	MARS depletion + trypsin digestion + LC-MS/MS

Tissue Proteome Dataset		
Dataset Name	Sample type of tumor-bearing mice	Experiment
Breast Tissue	1 cm tumor	homogenization + trypsin digestion + LC-MS/MS

Tissue Transcriptome Datasets		
Dataset Name	Sample type of tumor-bearing mice	Experiment
Whole breast array	1 cm tumor	homogenization + total RNA extraction + microarray analysis of whole breast tissue
Whole liver array	1 cm tumor	homogenization + total RNA extraction + microarray analysis of whole liver tissue
Whole spleen array	1 cm tumor	homogenization + total RNA extraction + microarray analysis of whole spleen tissue
Whole blood array	1 cm tumor	homogenization + total RNA extraction + microarray analysis of whole blood tissue (blood with plasma fraction removed)
Whole thymus array	1 cm tumor	homogenization + total RNA extraction + microarray analysis of whole thymus tissue
LCM breast array	1 cm tumor	homogenization + total RNA extraction + microarray analysis of laser capture microdissected (LCM) breast tissue

^aThe MARS and MARS+SCX datasets were based on two sample sets, with each set consisting of one normal and one cancer plasma pool (4 plasma pools total). All other proteome datasets were based on only one sample set, with each set again consisting of one normal and one cancer plasma pool. See also Figures 1A and B.

^bSee Figure 1B.

Table 2
Summary of identified peptides and protein groups for the proteome datasets

All numbers are based on a PeptideProphet error rate cut-off $\leq 1\%$.

Dataset	Numbers of LC-MS/MS Technical Repeats	Unique Peptides ^a	Non-redundant Protein Groups ^{a,b}	Non-redundant Protein Groups ^{a,b}		
				w/ 1 Identified Peptide	w/ 2 Identified Peptides	w/ ≥ 3 Identified Peptides
MARS - Sample Pool 1						
Cancer	30	1581	200	92	19	89
Normal	28	1607	208	105	18	85
Total		1918	282			
Common to both		1270	126			
MARS - Sample Pool 2						
Cancer	29	1718	210	96	20	94
Normal	30	1560	211	107	17	87
Total		1955	291			
Common to both		1323	130			
MARS+SCX - Sample Pool 1						
Cancer	Fxn2=16 ^c , Fxn3=25, Fxn8=28, Fxn9=26	1337	316	211	22	83
Normal	Fxn2=16, Fxn3=21, Fxn8=24, Fxn9=20	1348	301	188	23	90
Total		1643	435			
Common to both		1042	182			
MARS+SCX - Sample Pool 2						
Cancer	Fxn2=16, Fxn3=23, Fxn8=19, Fxn9=17	1283	308	202	26	80
Normal	Fxn2=16, Fxn3=20, Fxn8=23, Fxn9=17	1483	338	215	35	88
Total		1643	471			
Common to both		1029	175			
MARS+Cys+SCX						
Cancer	87 ^d	8000	1774	1063	184	527

Dataset	Numbers of LC-MS/MS Technical Repeats	Unique Peptides ^a	Non-redundant Protein Groups ^{a,b}	Non-redundant Protein Groups ^{a,b}		
				w/ 1 Identified Peptide	w/ 2 Identified Peptides	w/ ≥3 Identified Peptides
Normal	86 ^d	7292	1766	1068	184	514
Total		10304	2687			
Common to both		4988	853			
3 day MARS						
Cancer	32	3734	647	380	55	212
Normal	32	3453	616	375	48	193
Total		4219	892			
Common to both		2968	371			
6 day MARS						
Cancer	30	3391	594	372	50	172
Normal	30	3315	576	363	53	160
Total		3862	820			
Common to both		2844	350			
Breast Tissue						
Cancer	50	7112	1380	501	228	651
Normal	50	3945	744	339	109	296
Total		8702	1659			
Common to both		2355	465			

^a excluding peptides and protein groups pertaining to trypsin.

^b The numbers of non-redundant protein groups are based on the numbers of non-redundant protein groups in Supporting Information 3 (more than one IPI might be part of a group due to the ambiguity due to e.g. isoforms, but each group is counted as "one" protein).

^c Fxn = SCX pooled fraction, see the main text.

^d The numbers of LC-MS/MS runs for the MARS+Cys+SCX do not refer to technical repeats of the same samples, but rather to a single technical repeat LC-MS/MS injection of each SCX fraction upon SCX fractionation of the three different samples (the MARS bound sample, the cysteinyl-containing peptide sample of the MARS flow-through, and non-cysteinyl-containing peptide sample of the MARS flow-through).