Extracting Pathway-level Signatures from Proteogenomic Data in Breast Cancer Using Independent Component Analysis

Authors
Wenke Liu, Samuel H. Payne, Sisi Ma, and David Fenyo

Correspondence
David@FenyoLab.org; sisima@umn.edu

In Brief
Independent component analysis was applied to human breast cancer proteogenomic data, and pathway-level signatures were further integrated with clinical information. Our results demonstrated that ICA can be used to extract biological relevant signals from multi-omics data in an unsupervised manner.

Graphical Abstract

Highlights
• Unsupervised feature extraction from proteogenomics data.
• Pathway level integration of multi-omics data based on clinical features.
Recent advances in the multi-omics characterization necessitate knowledge integration across different data types that go beyond individual biomarker discovery. In this study, we apply independent component analysis (ICA) to human breast cancer proteogenomics data to retrieve mechanistic information. We show that as an unsupervised feature extraction method, ICA was able to construct signatures with known biological relevance on both transcriptome and proteome levels. Moreover, proteome and transcriptome signatures can be associated by their respective correlation with patient clinical features, providing an integrated description of phenotype-related biological processes. Our results demonstrate that the application of ICA to proteogenomics data could lead to pathway-level knowledge discovery. Potential extension of this approach to other data and cancer types may contribute to pan-cancer integration of multi-omics information. Molecular & Cellular Proteomics 18: S169–S182, 2019. DOI: 10.1074/mcp.TIR119.001442.

Breast cancer is the most common cancer among women, and although targeted therapies have helped to significantly reduced breast cancer mortality rate in the past decade, further improvement will require a comprehensive understanding of the molecular mechanisms of the disease (1, 2). Recently, deep mass spectrometry based proteomic characterization of genomically annotated breast cancer samples by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) has marked the initial step of a proteogenomic integrative approach, in which recurrent mutations and copy number variations on the genomic level, expression profiles on the transcriptomic level and protein abundance and functional manifestations on proteomic level were measured for the same group of patient samples and examined in the same framework (3–5). The collection of high-quality multi-omics data immediately led to the discovery of concordant gene amplification and protein phosphorylation in key pathways (3). At the same time, there is increasing demand for analytical approaches that could incorporate all data types and extract pathway level signatures. Because in all human patients “-omics” data sets the number of features far exceeds the number of samples, analysis of any single data type is already susceptible to “the curse of dimensionality,” and integration by simple concatenation of multi-omics data would be an even less desirable option. Our previous work has benchmarked the predictive power of multi-omics data sets for classifying breast cancer patients into different survival groups and showed that combined multi-omics data sets produced with data-driven fusion techniques were not able to outperform proteomic data alone (6). This result highlighted the possible redundancy among information contained in different biological levels and motivates us to explore other data fusion techniques that extract both concordant and complementary features from high-dimensional multi-omics data.

In the current study, we applied independent component analysis to proteomic and transcriptomic data of 77 breast cancer samples to extract pathway-level molecular signatures. Independent component analysis (ICA) is an unsupervised learning method widely used in signal processing and has been applied to cancer genomics with notable success (7–9). This approach decomposes the molecular profiles into linear combinations of non-Gaussian independent sources or components, each of which is comprised of weighted contributions from individual genes. Therefore, ICA reduces the dimensionality of original data by representing the molecular profile of each sample as weighted sum of several “meta-genes” or “meta-proteins,” and the weight of specific meta-gene/protein (mixing scores) in one sample reflects the “activity” of that component in the sample. Different from the more conventional dimension reduction method of principal component analysis (PCA), which seeks to find uncorrelated factors that explain the variance among the data, and works the best when the underlying components are normally distributed, ICA are able to discover more informative representations of high-dimensional biological signals, which are usu...
ally super-Gaussian and contain more close-to-zero values than a normally-distributed sample (10). As clinical features are also available for the CPTAC samples, molecular signatures can be constructed from clusters of meta-genes/proteins that show activity patterns correlated with these clinical features. Further, taking advantage of a specific clinical feature as an “anchor,” this method may help extract patterns at different biological levels and across different cohorts, which may originate from the same cellular functionality (Fig. 1). The signatures extracted from different data sets were filtered based on their intrinsic stability and association with known clinical features (see Experimental Procedures below) and grouped into modules that showed similar correlation patterns to clinical features. Subsequent gene set enrichment analysis revealed the biological relevance of these modules to pathways such as HER2 signaling, mitosis, and histone modification. Our analysis has demonstrated that ICA was able to blindly extract information in the form of weighted gene combinations, which may be biologically meaningful at pathway level. With input from clinical features or other sample subgrouping indices, these signatures may be further integrated into multilevel models that provide insights into the molecular mechanisms of breast cancer.

**ExPERIMENTAL PROCEDURES**

**Data Sets**—Transcriptome and proteome data of 77 breast invasive carcinoma (BRCA) patients and 84 ovarian serous cystadenocarcinoma (OV) patients were acquired by the TCGA and CPTAC projects: the transcriptome was characterized using the Agilent mRNA expression array platform (Santa Clara, CA) MS and global protein abundance data were obtained using iTRAQ 4-plex LC-MS/MS (3, 11–13). For OV data, Preprocessing procedure collapsed peptides mapped to the same gene by taking the mean value. Expression levels of genes and abundance of proteins were included in the transcriptome and proteome matrices. Clinical and demographic features were associated with each tumor sample.

**Independent Component Analysis**—Data were presented in a $p \times n$ matrix $X$ with genes in rows and samples in columns. The goal of ICA is to decompose the $p \times n$ data matrix into the product of a source matrix $S = (p \times k)$ and a mixing matrix $A = (k \times n)$. The $i$th column of the source matrix represents coefficients of each of the $p$ genes for the $i$th independent component. The coefficient vector of each component could be considered as $p$ random samples that revealed probability distribution of a specific random variable. The mutual information between any pair of such variables is minimized, and the components are therefore statistically independent. Genes with coefficients of positive or negative values in the same component indicated that their levels may be enhanced or suppressed in the same biological process. The $i$th row of the mixing matrix represents contributions of the $i$th component in the source matrix to profiles of each of the $n$ samples. Rows of the mixing matrix therefore record the activity of each components (or meta-genes/meta-proteins) across $n$ samples.

The stochastic nature of most ICA algorithms entails that different randomly initiated runs would give rise to different results, and there is no guarantee that the true structure of the data could be correctly estimated from any single run (10, 14). To assess the statistical reliability of ICA results, stable components were filtered based on an approach adapted from Engreitz et al., which cluster components

---

1 The abbreviations used are: CPTAC, Clinical Proteomic Tumor Analysis Consortium; ICA, independent component analysis; PCA, principal component analysis; LC-MS/MS, liquid chromatography tandem mass spectrometry; FDR, false discovery rate; TCGA, The Cancer Genome Atlas; BRCA, breast invasive carcinoma; OV, ovarian serous cystadenocarcinoma.
obtained from different runs with K-medoids method (8). In brief, ICA was run with \( k = n \) for 50 times to extract as much information as possible. As the sign of the same component may be flipped during different randomly-initiated runs, we followed the algorithm of Engreitz et al. and took advantage of the fact that the extracted components usually had skewed distributions (different sizes of tails on positive and negative ends) to align flipped components (8). For each individual component, if the larger tail was on the negative side, the component would be flipped to ensure that all larger tails were positive. All \( 50 \times n \) components were then considered as data points in \( p \) dimensional space and subjected to K-medoids clustering with Spearman correlation as the dissimilarity measure. For each cluster, the number of different runs that its members were extracted from was documented as a measure for cluster consistency, alongside silhouette width. In general, clusters that with members appeared in more than 50% of all runs (25 out of 50) were considered as likely to contain true biological signals (see Signature Annotation section below).

All computations were carried out on the R platform. Package “fastICA” which implements the iterative FastICA algorithm (15) was used to extract non-Gaussian independent components with logcosh contrast function. Components were subsequently assigned to clusters using the “cluster” package. Clusters were visualized with 2d t-SNE using the R package “tsne.” The number of clusters was determined as equal to number of components extracted at each run of ICA. When the number of samples is small comparing to the number of features, which is usually the case for biological data, it is convenient to retrieve as many as independent signal sources as possible, and the number of components extracted is equal to sample size.

The package “pcaMethods” was used to calculate principal components for comparison with independent components. In permutation tests, for each gene, protein abundance measurements of all samples were randomly shuffled without replacement. One hundred permuted data sets were created to assess the specificity of the ICA method.

**Signature Annotation**—Component clusters were annotated with GO terms by running Gene Set Enrichment Analysis against centroid coefficients as the pre-ranked gene lists (16, 17). Enrichment map of components were visualized with Cytoscape 3 (18). Each cluster was also associated with clinical features as following: First, 22 clinical features were recoded into ordinal variables (supplemental Table S1). Second, ordinary linear regression models were built with corresponding mixing scores for members in a component cluster to predict each of the ordinal responses. Counts of significant associations between components and clinical features (p value for the slope coefficient < 5.9 × 10\(^{-7}\), Bonferroni correction at \( \alpha = 0.01 \) level) were tabulated. Hierarchical clustering with complete linkage was applied to the clinical associations of independent components clusters extracted from both transcriptome and proteome data.

**RESULTS**

**Stable Molecular Signatures Extracted from Proteome and Transcriptome Data**—For both the proteome and transcriptome data sets we identified 77 clusters of meta-proteins or meta-genes from independent components obtained from multiple randomly initiated runs. In this study, we chose to run the procedure for 50 times with considerations of both computation time and numeric stability, as further increase the number of runs give rise to similar results (supplemental Fig. S3). Centers of the stable meta-gene and meta-protein clusters, which could be found by averaging gene coefficients within each cluster, could represent pathway-level signatures. The stability of these signatures could be inspected by visualizing all meta-genes and meta-proteins with t-distributed stochastic neighbor embedding (t-SNE), a widely used dimensionality reduction technique (19) (Fig. 2A, 2C, supplemental Fig. S1). A large portion of the meta-gene and meta-protein
clusters are compact, whereas the rest formed a visually indistinguishable mixture, consistent with the observation that the average silhouette widths of all clusters followed a bimodal distribution (Fig. 2B, 2D). Another metric that could help evaluate the stability of extracted signatures is the number of different runs that the members of each cluster originated from. It has been proposed that reliable independent components should be highly recurrent, therefore, a cluster the members of which were extracted from different runs would be considered as more stable. Indeed, this metric largely agrees with silhouette width, meaning that compact clusters were most often comprised of meta-proteins or meta-genes extracted from all 50 runs.

**Mixing Scores Reveal Clinical Relevance of Identified Components**—In addition to their numerical stability, the clusters of components representing groups of meta-proteins or meta-genes identified from different ICA runs can also be evaluated with their relevance to known characteristics of breast cancer. For each ICA decomposition, the rows of the mixing score matrices represented the “activity level” of the corresponding independent components in all of the samples. Therefore, it is possible to establish associations between the activity patterns of meta-genes and meta-proteins and clinical features available for TCGA samples, which would in turn reveal the functional relevance of the signatures. We recoded 22 clinical features into ordinal factors and used linear regression to assess their correlation with activity scores of meta-genes or meta-proteins in each signature cluster. To select the most significant associations, Bonferroni correction was applied to control for multiple comparisons at $\alpha = 0.01$ level, such that a nominal $p$ value of $5.9 \times 10^{-7}$ was set as the significance threshold (corrected $\alpha = 0.01/(77 \times 22)$). Within each cluster, the percentage of meta-proteins or meta-genes with activity levels that showed significant linear relation with the 22 clinical features was documented. Large portion of significant association between a signature cluster and a clinical feature indicates that the signature may contain pathway-level information about molecular mechanisms underlying the clinical feature. Thirty-three out of the 77 proteome IC clusters showed significant association with 10 clinical features, whereas 60 transcriptome clusters were correlated with 12 clinical features (Fig. 3). Six proteomic signatures and 7 tran-
scriptomic signatures showed strong clinical associations, such that the corresponding mixing scores were significantly correlated with a clinical variable in more than 50% of all runs (counts >25). In contrast, when the same procedure applied to 100 sets of randomly permuted proteome data, no component clusters showed strong clinical associations, as the count of significant linear correlations between mixing scores of a given signature cluster and any of the tested clinical variables was no larger than 2 in all 7700 clusters of the 100 permutation settings.

To further assess the adequacy and efficacy of the ICA method on the CPTAC breast cancer proteomics data, we compared the results obtained with ICA to that extracted by principal component analysis (PCA), a “standard” and more widely used method of dimension reduction (Fig. 4). With PCA, the BRCA data could be represented by 77 principal components, each of which is a 77-element vector representing scores for the 77 samples. In the context of ICA, because of the stochastic nature of the method, a similar representation was constructed by averaging the activity scores corresponding to the independent component in each of the 77 clusters obtained from the 50 repetitive runs. As shown in Fig. 4, PCA gave rise to signatures that contained mixed information, as the significant associations with clinical features concentrated on Principal Component 2. A total of 6 PCA-extracted signatures showed significant correlation ($p < 0.001$, linear regression) with one or more clinical variables, whereas 17 ICA-extracted signatures were significant associated with at least one clinical variables. This comparison, consistent with the fact that ICA was designed for “de-mixing” signals of different sources, further demonstrated that ICA could extract more specific components.
TABLE I

Average coefficients of selected BRCA proteome signatures

|       | pr_02  |       | pr_04  |       | pr_29  |       |
|-------|--------|-------|--------|-------|--------|-------|
| pr_02 | 11.13  | HBZ   | −4.59  |       |        |       |
| pr_04 | 10.68  | CKB   | −4.06  |       |        |       |
| pr_29 | 10.10  | PIP   | −3.96  |       |        |       |
|       | 8.87   | GP2   | −3.91  |       |        |       |
|       | 8.46   | CKM   | −3.56  |       |        |       |
|       | 8.11   | TNFSF12| −3.31 |       |        |       |
|       | 6.55   | MAPT  | −3.25  |       |        |       |
|       | 6.39   | CALB1 | −3.03  |       |        |       |
|       | 6.02   |       |        |       |        |       |
|       | 5.93   |       |        |       |        |       |
|       | 4.91   |       |        |       |        |       |
|       | 4.88   |       |        |       |        |       |
|       | 4.71   |       |        |       |        |       |
|       | 4.64   |       |        |       |        |       |
|       | 4.64   |       |        |       |        |       |
|       | 3.96   |       |        |       |        |       |
|       | 3.91   |       |        |       |        |       |
|       | 3.85   |       |        |       |        |       |
|       | 3.72   |       |        |       |        |       |
|       | 3.70   |       |        |       |        |       |
|       | 3.65   |       |        |       |        |       |
|       | 3.55   |       |        |       |        |       |
|       | 3.53   |       |        |       |        |       |
|       | 3.44   |       |        |       |        |       |
|       | 3.26   |       |        |       |        |       |
|       | 3.25   |       |        |       |        |       |
|       | 3.25   |       |        |       |        |       |
|       | 3.14   |       |        |       |        |       |
|       | 3.10   |       |        |       |        |       |

Independent Component Analysis of Breast Cancer Proteogenomic Data
information about biological processes underlying high dimensional data.

The biological relevance of the clusters identified with ICA could be further validated by inspecting the average gene coefficients (the centroids of the clusters). For example, the BRCA proteome signature cluster that contains 29 members that were significantly correlated with the HER2 receptor immuno status and 22 members associated with the Her2 subtype index (pr_02), was heavily weighted on ERBB4 and ERBB2, two tyrosine receptor kinases that mediate the Her2 signaling, and the two proteins were assigned the two largest coefficients (11.13 and 10.68, Table I). Another signature (pr_29) correlated with estrogen receptor (ER) and progesterone receptor (PR) status also exhibited high average scores for PGR (7.73) and ESR1 (4.42) proteins. Interestingly, the signature pr_04, which displayed strong correlation with the Luminal A subtype, is enriched with a group of mitosis-related proteins, which showed low levels of abundance in the Luminal A subtype (Fig. 5). This finding is consistent with the low proliferation level of Luminal A subtype (20). Protein abundance of genes with large weights in the same signatures also showed strong intra-module correlation (supplemental Fig. S2). The aforementioned associations between proteome signatures and known biological processes demonstrated that ICA was able to rediscover clinically relevant information from proteome data in an unsupervised manner.

Large value coefficients of known marker genes revealed the biological relevance of a subset of the molecular signatures, but more biological information could be exploited on the pathway level. The coefficient vector for each signature is a preranked gene list that can be subjected to gene set enrichment analysis (GSEA) on curated gene sets and GO terms (16), and the collective effect of genes with coefficients of smaller absolute values could contribute to the enrichment metric. GSEA results showed that the clinically related proteome IC clusters retrieved a lot of the breast cancer related gene sets determined by experimental manipulations, as well as gene sets that characterized basic biological processes (Table II). For example, the proteome signatures pr_02 and pr_29, which were strongly associated with Her2 and ER status, also exhibited enrichment of the corresponding gene sets (Table II). The transcriptome signature tx_07, which were correlated with Basal subtype, ER and PR status, showed negative enrichment of targets of LSD1, a histone demethylase of H3K4 sites, which had known links to be poor breast cancer prognostics and ER negative status (21, 22). Interestingly, the same signature cluster also exhibited positive enrichment of genes with H3K27Me3 sites, indicating that epigenetic regulation is highly orchestrated in breast cancer.

Although the most clinically relevant clusters (total count of significant association >25) contained proteome and transcriptome signatures that are highly recurrent, these clusters were not very homogeneous (supplemental Table S2). On the other hand, out of the 20 most consistent clusters with average silhouette widths larger than 0.9 and are recurrent in every run for both proteome and transcriptome, only 3 signatures (pr_04, tx_46, tx_63) showed direct clinical relevance. The meta-proteins and meta-genes in the clusters of pr_04 and tx_46 were both significantly associated with Luminal A subtype index and showed enrichment of genes involved in estrogen response (Table II). The signature tx_63 was associated with the Basal subtype and ER status, and its coefficients exhibited significant negative enrichment of genes in 16q24 loci ($p < 0.001$, FDR < 0.001), a breast cancer
risk factor, as well as positive enrichment of genes involved in endocrine therapy response ($p < 0.001$, FDR < 0.001). The large extent of disagreement between clinical relevance and cluster consistency suggest that ICA may serve as a promising approach to knowledge discovery, as stable clusters that did not show clinical relevance may provide new insights into the molecular mechanisms of breast cancer.

**Integrative Pathway Analysis of Proteomic and Transcriptomic Components Guided by Clinical Features**—The correlation between mixing scores and clinical features provided a valuable opportunity to find the link between independently extracted components from different data types. For example, in both BRCA proteome and transcriptome analyses, there was a highly consistent signature (pr_04 and tx_46), such that meta-proteins and meta-genes in the two clusters were recurrent in all 50 randomly initiated runs and the corresponding mixing scores all showed significant correlation with index of Luminal A subtype (Figs. 2 and 3). To integrate the proteomic and transcriptomic information with guidance from clinical relevance, we applied the hierarchical clustering algorithm to the vectors of clinical association counts for the most clinically relevant meta-protein and meta-gene clusters. Proteome and transcriptome signatures could therefore be grouped based on their similarity in functional indications, as the metric of direct correlation between gene coefficients of different signatures was limited by noise introduced by high dimensionality (supplemental Fig. S4). Combined network visualization of GO terms enriched in proteome and transcriptome signatures allowed us to examine complementary functional modules on different biological levels. For example, the pr_04 and tx_46 clusters may characterize Luminal A subtype specific processes from protein and RNA perspectives (Fig. 6B). Although majority of the gene set networks consisted of nodes and edges common between proteome and transcriptome, specific networks could be identified for both data modes, with tx_46 showed specific enrichment in a DNA
| Clinically relevant signatures                                                                 | Top enriched GO terms                                                                 | Top enriched curated sets                                                                 |
|-----------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|
| pr_02 GO_MYOSIN_II_COMPLEX                                                                      | GO_ERBB2_SIGNALING_PATHWAY                                                            | REACTOME_RESPIRATORY_ELECTRON_TRANSPORT                                                   |
| pr_04 GO_BRANCHED_CHAIN_AMINO_ACID_METABOLIC_PROCESS                                             | GO_DNA_REPLICATION                                                                    | WOO_LIVER_CANCER_RECURRANCE_UP                                                            |
| pr_18 GO_CILIARY_TIP                                                                            | GO_OXOREDUCTASE_ACTIVITY_acting_on_the_CH_OH_Group_of_donors_NAD_or_NADP_as_acceptor  | DUTERTRE_ESTRADIOL_RESPONSE_24HR_UP                                                       |
| pr_29 GO_MITOCHONDRIAL_TRANSLATION                                                              | GO_ANCHORED_COMPONENT_OF_MEMBRANE                                                      | REACTOME_LIPOPROTEIN_METABOLISM                                                          |
| pr_48 GO_ACTIN_MYOSIN_FILAMENT_SLIDING                                                           | GO_CELLULAR_MONOVALENT_INORGANIC_CATION_HOMEOSTASIS                                   | DOANE_BREAST_CANCER_ESR1_UP                                                               |
| pr_49 GO_MICROBODY_LUMEN                                                                        | GO_RBOSOME_BIogenesis                                                                | SMID_BREAST_CANCER_BASAL_UP                                                               |
| tx_07 GO_CYTOSOLIC_LARGE_RIBOSOMAL_SUBUNIT                                                        | GO_EMBRYONIC_HINDLIMB_MORPHOGENESIS                                                   | CAIRO_HEPATOBlastoma_CLASSES_UP                                                           |
| tx_19 GO_NECROPTOTIC_PROCESS                                                                     | GO_Oligodendrocyte_differentiation                                                     | MIKKELSEN_NPC_HCP_WITH_H3K27ME3                                                            |
| tx_21 GO_RESPONSE_TO_PH                                                                          | GO_STEROID_BINDING                                                                   | KOHOUTEK_CONTI_TARGETS                                                                   |
| tx_31 GO_VASODILATION                                                                           | GO_MHC_PROTEIN_COMPLEX                                                                | NIKOLSKY_BREAST_CANCER_17Q11_Q21_AMPLICON                                               |
| tx_34 GO_VOLTAGE_GATED_CALCIUM_CHANNEL_COMPLEX                                                   | GO_MUSCLE_CELL_MIGRATION                                                             | DOANE_BREAST_CANCER_ESR1_UP                                                               |
| tx_40 GO_REGULATION_OF_ENDORCINE_PROCESS                                                         | GO_EPIDERMIS_DEVELOPMENT                                                             | ICIBHA_GRAFT_VERSUS                                                                       |
| tx_46 GO_NEGATIVE_REGULATION_OF_CALCIUM_ION_TRANSPORT                                           | GO_CHROMOSOMESEGREGATION_1 VS 3 DN                                                   | HOST_DISEASE_D7_DN                                                                        |
| tx_63 GO_CALCIUM_DEPENDENT_CELL_CELL_ADHESION_VIA_PLASMA_MEMBRANE_CELL_Adhesionוביל מולקוללי | GO_CILIMUM_ORGANIZATION                                                               | LIN_SILENCED_BY_TUMOR_MICROENVIRONMENT                                                    |
| tx_64 GO_INNER_EAR_RECEPTOR_CELL_DEVELOPMENT                                                      | GO_POSITIVE_REGULATION_OF_INTERLEUKIN_2_PRODUCTION                                    | TURA_SHALI_BREAST_DUCTAL_CARCINOMA_VS_DUCTAL_NORMAL_DN                                   |
| tx_72 GO_PROTEIN_LIPID_COMPLEX                                                                   | GO_NEUROEPITHELIAL_CELL_DIFFERENTATION                                                | TSUNODA_CISPLATIN_RESISTANCE_DN                                                           |
| Consistent signatures                                                                           |                                                                                      |                                                                                           |
| pr_33 GO_NUCLEAR_TRANSCRIBED_MRNA_CATABOLIC_PROCESS_NONSENSE_MEDIATED_DECAY                      | GO_IRON_ION_BINDING                                                                  | JAATINENHEMEPOETOICSTEM_CELL_DN                                                           |
| pr_45 GO_TRANSITIONAL_TERMINATION                                                                | GO_LYMPHOCYTE_ACTIVATION                                                              | WALLACE_PROSTATE_CANCER_RAGE_UP                                                           |
| pr_15 GO_VOLTAGE_GATED_ION_CHANNEL_ACTIVITY                                                      | GO_BLOOD_MICROPARTICLE                                                                | HSIAO_LIVER_SPECIFIC_GENES                                                                |
| pr_25 GO_DNA_PACKAGECOMPLEX                                                                      | GO_CONTRACTILE_FIBER                                                                 | REACTOME_MUSCLE_CONTRACTION                                                               |
| pr_51 GO_TUBEFORMATION                                                                           | GO_LIPID_PARTICLE                                                                    | NAKAYAMA_SOFT_TISSUE_TUMORS_PC2_DN                                                        |
| pr_08 GO_SRNA_METABOLIC_PROCESS                                                                   | GO_NUCLEAR_OUTER_MEMBRANE_ENDOPLASM_SMETULLUM_MEMBRANE_NETWORK                        | MILU_PSEUDOPODIA_HAPTOTAXIS_DN                                                            |

TABLE II
GO terms and curated gene sets enrichment of selected BRCA signatures
| Top enriched GO terms | Top enriched curated sets |
|-----------------------|--------------------------|
| **Negative**          | **Positive**             |
| GO_FATTY_ACID_BINDING | GO_DEFENSE_RESPONSE_TO_VIRUS | REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA | TAKEDA_TARGETS_OF_NUP98_HOXA9_FUSION_3D_UP |
| GO_GOLGI_VESICLE_TRANSPORT | GO_TRANSFERASE_ACTIVITY_TRANSFERRING_ALKYL_ORARYL_OTHER_THAN_METHYL_GROUPS | LEE_LIVER_CANCER_CIPROFIBRATE_UP | KEGG_DRUG_METABOLISM_CYTOCHROME_P450 |
| GO_Branched_Chain_Amino_Acid_Metabolic_Process | GO_DNA_REPLICATION | WOO_LIVER_CANCER_RECURRENCE_DN | DUTERTRE_ESTRADIOL_RESPONSE_24HR_UP |
| GO_PHOSPHOLIPID_BIOSYNTHETIC_PROCESS | GO_DNA_PACKAGING | KEGG_DRUG_METABOLISM_OTHER_ENZYMES | WAMUNYOKOI_OVARIAN_CANCER_Grades_1_2_DN |
| GO_Organelar_Large_Ribosomal_Subunit | GO_NUCLEAR_TRANSCRIBED_MRNA_CATABOLIC_PROCESS_NONSENSE_MEDIATED_DECAY | ZHANG_RESPONSE_TO_IKK_INHIBITOR_AND_TNF_UP | REACTOME_PEPTIDE_CHAIN_ELONGATION |
| GO_Negative_Regular_of_Calcium_Ion_Transport | GO_CHROMOSOME_SEGREGATION | SOTRIOU_BREAST_CANCER_GRADE_1_VS_3_DN | DUTERTRE_ESTRADIOL_RESPONSE_24HR_UP |
| GO_Bicarbonate_Transport | GO_PROTEINACEOUS_EXTRACELLULAR_MATRIX | MOOTHA_HUMAN_MITO0DB_6_2002 | ANASTASSIOUl_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE |
| GO_Nucleotide_Phosphorylation | GO_ADAPTIVE_IMMUNE_RESPONSE | SMID_BREAST_CANCER_LUMINAL_B_UP | SMID_BREAST_CANCER_NORMALLIKE_UP |
| GO_Translational_Initiation | GO_SODIUM_CHANNEL_ACTIVITY | HSIOHOUSEKEEPINGGENES | GINESTER_BREAST_CANCER_2Q13_AMPLIFICATION_UP |
| GO_Cytosolic_Ribosome | GO_DEFENSE_RESPONSE_TO_VIRUS | REACTOME_PEPTIDE_CHAIN_ELONGATION | HECKER_IFNB1_Targets |
| GO_Ribosomal_Subunit | GO_HELAECASE_ACTIVITY | REACTOME_PEPTIDE_CHAIN_ELONGATION | GABRIELY_MIR1_TARGETS |
| GO_Regulation_of_Mesenchymal_Cell_Proliferation | GO_INFLAMMATORY_RESPONSE | SENQUPTA_NASOPHARYNGEAL_CARCINOMA_WITH_LMP1_DN | MCLACHLAN_DENTAL_CARIES_UP |
| GO_Calcium_Dependent_Cell_Cell_Adhesion_Via_Plasmatic_Membrane_Cell_Adhesion_Molecules | GO_CILIUM_ORGANIZATION | NIKOLSKY_BREAST_CANCER_16Q24 AMPLICON | CREIGHTON_ENDOCRINE_Therapy_RESISTANCE_2 |
| GO_Positive_Regulation_of_Mesenchymal_Cell_Proliferation | GO_LIPID_PARTICLE | MUELLER_METHYLATED_IN_GIOMBASTOMA | NAKEYAMA_SOFT_TISSUE_TUMORS_PCA2_DN |
replication-related network, and pr_04 showed specific enrichment in polysaccharide metabolism (Fig. 6C).

To further validate the ICA workflow, we apply the method to proteomic data of the ovarian cancer CPTAC cohort (4). Signatures found in the data showed association with previously reported transcriptome-based subtypes (supplemental Fig. S5) (23). Moreover, there were strong correlations between ICA-extracted proteome signatures and protein modules derived with weighted correlation network analysis (supplemental Fig. SSD) (4, 24). To further highlight the potential of ICA workflow as an integrative method, we explored pan-cancer proteomic features by examining the correlation between ICA-extracted signatures obtained from breast cancer and ovarian cancer data sets. Between the 77 breast cancer proteome signatures (Fig. 2–3) and the 84 ovarian cancer proteome signatures (supplemental Fig. S5), four pairs of highly correlated signatures were identified (Spearman correlation larger than 0.2) (supplemental Fig. S6). Notably, the shared signature pairs include BRCA_IC_004/OV_IC_051, which was previously annotated to correlate with Luminal A subtype and show enrichments of mitosis-related genes (Fig. 5).
The share signature pair BRCA_IC_014/OV_IC_004, which was associated with the Mesenchymal subtype of ovarian cancer (supplemental Fig. S6), showed consistent enrichments in extracellular-matrix related genes. Taken together, our results demonstrated the ICA workflow could consistently extract biological relevant information from multi-omics data sets.

**DISCUSSION**

We have used independent component analysis to gain insights into the mechanisms of breast cancer and develop protein/gene modules as clinical signatures. Meta-proteins and meta-genes are extracted blindly from the data, and signatures are further selected based on consistency of meta-gene/protein clusters or the association between their activity scores and known clinical features. Gene set annotation revealed that several selected signatures contained biologically relevant information. A proteome signature (pr_02) characterizing strong activation of the Her2 pathway was recovered as a Her2-related meta-protein cluster. On the transcriptome level, a meta-gene enriched for genes in the 16q24 risk loci (tx_63) formed a stable signature that showed correlation with both ER status and Basal subtype. Stable signatures on both proteome and transcriptome levels (pr_04 and tx_46) were found to be heavily associated with the Luminal A index, suggesting that cell division and growth is specifically regulated in this subtype.

As an unsupervised blind source separation method, independent component analysis has been applied to multiple biological data types. Consistent with previous reports, our results demonstrated that as an unsupervised learning method, ICA can extract biological meaningful information solely based on the intrinsic structures of the transcriptome and proteome data. Because of the underlying assumptions, this method would only yield satisfactory results when the “true” signals in the data do not follow Gaussian distribution, and it would fail to separate any mixture of normally distributed sources. Successful application in the current use case suggested that ICA can capture some aspects of the mechanistic processes underlying proteome and transcriptome profiles. It is reasonable to assume that the observed RNA and protein levels are the sum of several up-regulation and down-regulation modules, in which the distribution of individual gene levels deviate radically from the normal distribution that characterizes a noisy background.

In the current use case, we have designated the number of components to be equal to the number of samples, as is usually assumed in blind source separation applications. It remains an open question whether all extracted signatures are reflective of some “true” biological processes. Moreover, the stochastic nature of the algorithm entails that, although all components were found simultaneously in each run without any component “privileged” over others (25), there is no guarantee that all solutions found at local optima are practically useful features. Indeed, the fact that a lot of the extracted components were neither consistent nor recurrent in multiple runs suggested that they may arise from noise instead of stable signals. However, previous applications of ICA on biological data have showed that overestimating the number of independent sources gave rise to signatures driven by smaller gene groups without affecting the rediscovery of the most stable components, while underestimating the number of sources compromised signal detection (26). It is possible that assuming the number of independent sources to be equal to the number of samples may result in noisy signals that are “unnecessary.” However, in an exploratory analysis setting where the number of samples is relatively small, this problem could be remedied by inspecting the stability and biological relevance of extracted signatures post hoc and extracting more components may lead to better separation of signals (supplemental Fig. S3). In the case of large sample size, the number of components will have to be determined by heuristics, and different choices may be compared in downstream analyses.

Using clinical association (total significant associations >25 across all tested clinical variables) or signature stability (cluster silhouette >0.9) as selection criterion gave rise to two lists of potential signatures (Table II) with only a few overlapping members (pr_04, tx_46, tx_63), suggesting that the most clinically relevant signatures are not very stable under the current method. Other clustering method such as density-based clustering may be used to improve the estimation of stable signatures (14, 27). On the other hand, it is possible that the most stable signatures described the housekeeping processes common in all samples, but they may also help reveal novel molecular mechanisms of breast cancer that are not previously linked to any phenotype. As a feature construction procedure, ICA could also facilitate knowledge integration from multiple data types. In addition to the integration of proteome and transcriptome signatures as demonstrated in this work, future studies could further apply ICA on omics data sets of multiple cancer types. Extracted molecular signatures may be grouped in another round of clustering procedure to reveal pan-cancer and cancer type specific mechanisms.

**DATA AVAILABILITY**

Proteome data of BRCA and OV are available at the CPTAC data portal (https://cptac-data-portal.georgetown.edu/cptac/s/S015; https://cptac-data-portal.georgetown.edu/cptac/s/S020). Transcriptome data are available at the NCI Genomics Data Commons (https://portal.gdc.cancer.gov) under project name TCGA-BRCA and TCGA-OV. Data tables are also available at the publication page (https://www.nature.com/articles/nature18003; https://www.nature.com/articles/nature10166).

* We would like to acknowledge funding by the National Cancer Institute (NCI) through CPTAC award U24 CA210972 and a contract 13XS068 from Leidos Biomedical Research, Inc., and by a grant from the Shifrin-Myers Breast Cancer Discovery Fund.

S This article contains supplemental Figures and Tables.
REFERENCES

1. Narod, S. A., Igual, J., and Miller, A. B. (2015) Why have breast cancer mortality rates declined? J. Cancer Policy 5, 8–17.

2. Mendes, D., Alves, C., Alonso, N., Cardoso, P., Passos-Coelho, J. L., Costa, L., Andrade, S., and Batel-Marques, F. (2015) The benefit of HER2-targeted therapies on overall survival of patients with metastatic HER2-positive breast cancer – a systematic review. Breast Cancer Res. 17, 140.

3. Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, J., M. J., Carr, S. A., and NCI CPTAC. (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 534, 56–62.

4. Zhang, H., Tiu, T., Zhang, S. H., B., McDermott, J. E., Zhou, J.-Y., Petyuk, V. A., Chen, L., Ray, D., Sun, S., Yang, F., Chen, L., Wang, W., Shah, P., Ch., S. A., and Batel-Marques, F. (2015) The benefit of HER2-targeted therapies on overall survival of patients with metastatic HER2-positive breast cancer – a systematic review. Breast Cancer Res. 17, 140.

5. Zhang, H., Tiu, T., Zhang, S. H., B., McDermott, J. E., Zhou, J.-Y., Petyuk, V. A., Chen, L., Ray, D., Sun, S., Yang, F., Chen, L., Wang, W., Shah, P., Ch., S. A., and Batel-Marques, F. (2015) The benefit of HER2-targeted therapies on overall survival of patients with metastatic HER2-positive breast cancer – a systematic review. Breast Cancer Res. 17, 140.

6. Zhang, H., Tiu, T., Zhang, S. H., B., McDermott, J. E., Zhou, J.-Y., Petyuk, V. A., Chen, L., Ray, D., Sun, S., Yang, F., Chen, L., Wang, W., Shah, P., Ch., S. A., and Batel-Marques, F. (2015) The benefit of HER2-targeted therapies on overall survival of patients with metastatic HER2-positive breast cancer – a systematic review. Breast Cancer Res. 17, 140.
12. Zhang, H., Liu, T., Payne, S. H., Zhang, B., McDermott, J. E., Zhou, J.-Y., Petyuk, V. A., Chen, L., Ray, D., Sun, S., Yang, F., Chen, L., Wang, J., Shah, P., Cha, S. W., Aiyetan, P., Woo, S., Tian, Y., Gritsenko, M. A., Clauss, T. R., Choi, C., Monroe, M. E., Thomas, S., Nie, S., Wu, C., Moore, R. J., Yu, K.-H., Tabb, D. L., Fenyb, D., Bafna, V., Wang, Y., Rodriguez, H., Ho, E. S., Hitkate, T., Rivers, R. C., Sokoll, L., Zhu, H., Shih, I.-M., Cope, L., Pandey, A., Zhang, B., Snyder, M. P., Levine, D. A., Smith, R. D., Chan, D. W., Rodland, K. D., Carr, S. A., Gillette, M. A., Klauser, K. R., Kuhn, E., Mani, D. R., Mertins, P., Ketchum, K. A., Thangdu, R., Cai, S., Oberti, M., Paulovich, A. G., Whiteaker, J. R., Edwards, N. J., Mcgarvey, P. B., Madhavan, S., Wang, P., Chan, D. W., Pandey, A., Shih, I.-M., Zhang, H., Zhang, Z., Zhu, H., Cope, L., Whiteley, N., Skates, S. J., White, F. M., Levine, D. A., Boja, E. S., Kininger, C. R., Hitkate, T., Mesri, M., Rivers, R. C., Rodriguez, H., Shaw, K. M., Stein, S. E., Fenyo, C., Liu, T., McDermott, J. E., Payne, S. H., Rodland, K. D., Smith, R. D., Rudnick, P., Snyder, M., Zhao, Y., Chen, X., Sansosti, T., Hoochm, A. N., Liebler, D. C., Sanders, M. E., Shi, Z., Siebenn, R. J. C., Tabb, D. L., Zhang, B., Zimmerman, L. J., Wang, Y., Davies, S. R., Ding, L., Ellis, M. J. C., and Townsend, R. R. (2016) Integrated proteogenomic characterization of high-grade serous ovarian carcinoma. Cell 166, 755–765

13. Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dhir, R., Slebos, R. J. C., Tabb, D. L., Zhang, B., Zimmerman, L. J., Wang, Y., Meier, C., Shi, Z., Skates, S. J., White, F. M., Levine, D. A., Boja, E. S., Kininger, C. R., Hitkate, T., Mesri, M., Rivers, R. C., Rodriguez, H., Shaw, K. M., Stein, S. E., Fenyo, C., Liu, T., McDermott, J. E., Payne, S. H., Rodland, K. D., Smith, R. D., Rudnick, P., Snyder, M., Zhao, Y., Chen, X., Sansosti, T., Hoochm, A. N., Liebler, D. C., Sanders, M. E., Shi, Z., Siebenn, R. J. C., Tabb, D. L., Zhang, B., Zimmerman, L. J., Wang, Y., Davies, S. R., Ding, L., Ellis, M. J. C., and Townsend, R. R. (2016) Integrated proteogenomic characterization of high-grade serous ovarian carcinoma. Cell 166, 755–765

14. Himberg, J., Hyvärinen, A., and Esposito, F. (2004) Validating the independent component analysis of reproducible transcriptomic data analysis. BMC Bioinformatics 5, 1214–1227

15. Hyvärinen, A., and Oja, E. (2000) Independent component analysis: algorithms and applications. Neural Networks 13, 411–430

16. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lande, E. S., and Mesirov, J. P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A. 102, 15545–15550

17. Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehár, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Housnis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lande, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. 34, 267–273

18. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwildtis, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504

19. Van Der Maaten, L., and Hinton, G. (2008) Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605

20. Coates, A. S., Winer, E. P., Goldhirsh, A., Gelber, R. D., Gnant, M., Piccart-Gebhart, M., Thürlimann, B., Senn, H.-J., and Panel Members. (2015) Tailoring therapies—improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. 26, 1533–1546

21. Lim, S., Janzer, A., Becker, A., Zimmer, A., Schule, R., Buettner, R., and Kiefel, J. (2010) Lysine-specific demethylase 1 (LSD1) is highly expressed in ER-negative breast cancers and a biomarker predictive aggressor. Carcinogenesis 31, 512–520

22. Nagasawa, S., Sedukhina, A. S., Nakagawa, Y., Maeda, I., Kubota, M., Ohnuma, S., Tsagawa, K., Otaka, T., Roche, S., Bannert, M. J., Nairn, K., Jeyasekharan, A. D., and Sato, K. (2015) LSD1 overexpression is associated with poor prognosis in Basal-Like Breast Cancer, and Sensitivity to PARP Inhibition. PLoS ONE 10, e0118002

23. Verhaak, R. G. W., Tamayo, P., Yang, J.-Y., Hubbard, D., Zhang, H., Creighton, C. J., Fereday, S., Lawrence, M., Carter, S. L., Mermel, C. H., Amin, N., Schwikowski, B., and Ideker, T. (2008) Components for reproducible transcriptomic data analysis. BMC Bioinformatics 9, 559

24. Hyvärinen, A. (1999) Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Networks 10, 626–634

25. Jiang, B., Heiser, L. M., and Zinovyev, A. (2017) Determining the optimal number of independent components for reproducible transcriptomic data analysis. BMC Genomics 28, 2579–2605

26. Langfelder, P., and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559

27. Jahirabadkar, S., and Kulkarni, P. (2013) Clustering for high dimensional data. Aplic. Mat. 63, 29–35