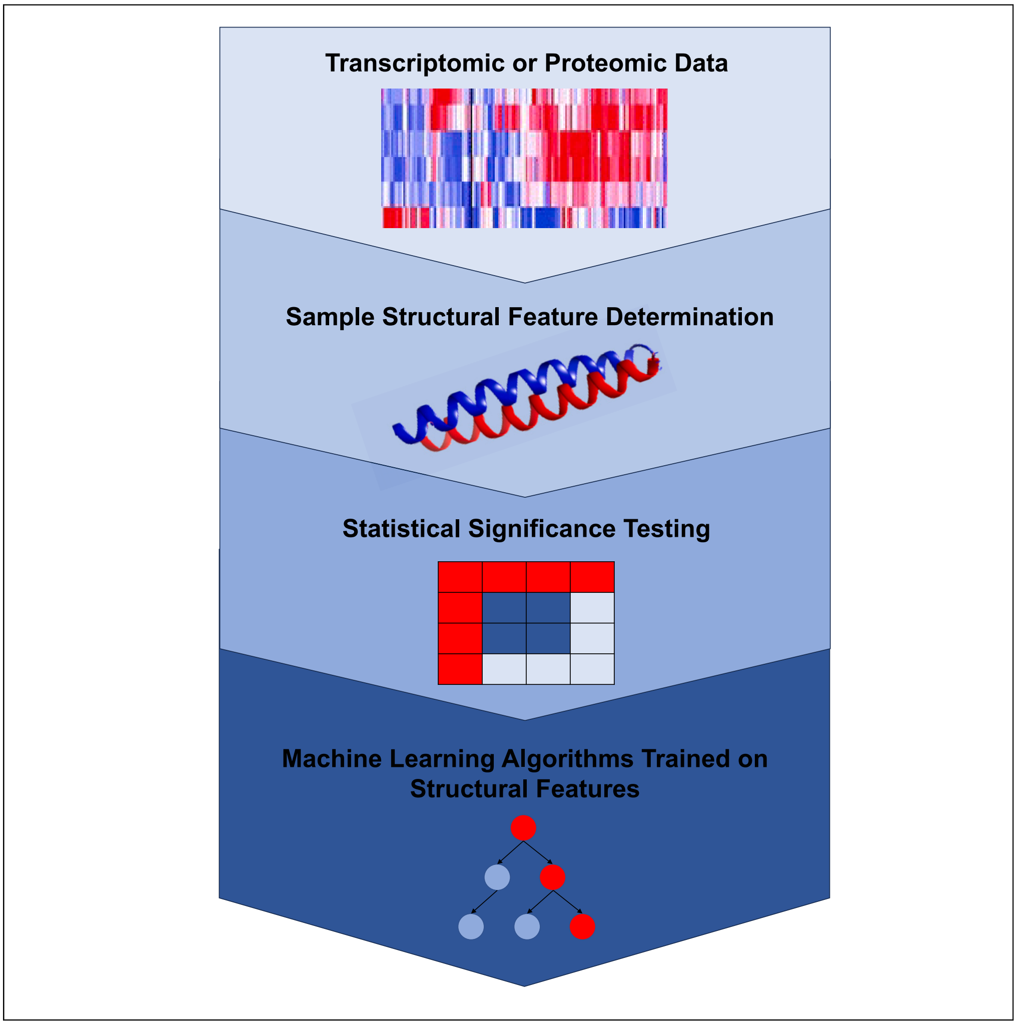
iScience

CellPress
OPEN ACCESS

Article

Structural analysis of genomic and proteomic signatures reveal dynamic expression of intrinsically disordered regions in breast cancer


Transcriptomic or Proteomic Data
Sample Structural Feature Determination
Statistical Significance Testing
Machine Learning Algorithms Trained on Structural Features

Nicole Zatorski, Yifei Sun, Abdulkadir Elmas, ..., Kuan-Lin Huang, Martin Walsh, Avner Schlessinger

nicole.zatorski@icahn.mssm.edu

Highlights

SAGES describes genomic and proteomic expression via underlying structural features

Feature selection trained on SAGES reveals unique protein features of breast tissue

SAGES of breast cancer reveal dynamic expression of intrinsically disordered regions

SAGES captures disease and drug trends, making it a potential tool for drug repurposing

Zatorski et al., iScience *27*, 110640
September 20, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.110640

iScience

CellPress
OPEN ACCESS

Article

Structural analysis of genomic and proteomic signatures reveal dynamic expression of intrinsically disordered regions in breast cancer

Nicole Zatorski,[1,5,*] Yifei Sun,[1] Abdulkadir Elmas,[2] Christian Dallago,[3,4] Timothy Karl,[4] David Stein,[1] Burkhard Rost,[4] Kuan-Lin Huang,[2] Martin Walsh,[1] and Avner Schlessinger[1]

SUMMARY

Structural features of proteins capture underlying information about protein evolution and function, which enhances the analysis of proteomic and transcriptomic data. Here, we develop Structural Analysis of Gene and protein Expression Signatures (SAGES), a method that describes expression data using features calculated from sequence-based prediction methods and 3D structural models. We used SAGES, along with machine learning, to characterize tissues from healthy individuals and those with breast cancer. We analyzed gene expression data from 23 breast cancer patients and genetic mutation data from the Catalog of Somatic Mutations In Cancer database as well as 17 breast tumor protein expression profiles. We identified prominent expression of intrinsically disordered regions in breast cancer proteins as well as relationships between drug perturbation signatures and breast cancer disease signatures. Our results suggest that SAGES is generally applicable to describe diverse biological phenomena including disease states and drug effects.

INTRODUCTION

With the advent of improved sequencing technology there has been a great emphasis placed on using proteomic and transcriptomic data to understand underlying disease etiologies and characteristics.[1] This has led to successful identification of biomarkers that have advanced precision medicine[2] in fields such as oncology.[3] For example, single cell RNA sequencing on primary breast cancer tumors has explored the heterogeneity of gene expression in tumor tissue and the preponderance of immune cell response to disease.[4] Analysis of gene sets from RNA expression can be based on a comparison of the gene names found to be differentially expressed in a sample population as compared to the control population. The alternative, well-established method for analyzing sequencing data are through the use of a gene ontology (GO) enrichment of the differentially expressed genes, which provides a standardized vocabulary and relationship for genes.[5] This gives a slightly more nuanced view of the signature; in that it provides annotations about the function of proteins encoded by genes. In proteomic analysis, protein names or fragments of protein sequences found using mass spectroscopy are analyzed for their relative abundances in a sample.[6] These gene expression and protein expression signatures can be further combined into interaction networks that reveal insights into gene associations when compared between different biological states.[7] Despite the developments in transcriptomic and proteomic analysis tools however, the existing methods described above do not capture functional similarities of genes based on their encoded protein structures. Because protein function and structure are so intrinsically linked, a new approach, which takes into account protein structural features, is need.

Structural features of proteins encompass a wide variety of characteristics, which provide insight into underlying functional and evolutionary relationships between proteins. Structural features range from sequence based descriptors of secondary structure[8] to hierarchical structure based classifications[9] to 3D model based descriptions of structure.[10] Accurate classifiers predicting structural features are available,[11–15] yielding 1D or string representations for input proteins. Moreover, the 3D structure prediction method AlphaFold2[16] allows genome-scale 3D structure prediction, and representations from full 3D coordinates to distance and contact maps. These descriptors can be applied to many types of data including genomic and proteomic expression as well as proteins assessed in chemical assays such as the IC50.

Because structural features are sequence and three-dimensional model based, they provide a more generalizable orthogonal representation that may capture underlying mechanistic information.[17] Thus, structural features of proteins can robustly describe biological function in

[1]Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, One Gustave Levey Pl, New York, NY 10029, USA
[2]Department of Genetic and Genomic Sciences, Icahn School of Medicine at Mount Sinai, One Gustave Levey Pl, New York, NY 10029, USA
[3]NVIDIA DE GmbH, Einsteinstraße 172, 81677 München, Germany
[4]Faculty of Informatics, Bioinformatics & Computational Biology, Technical University Munich (TUM), 85748 Garching, Germany
[5]Lead contact
*Correspondence: nicole.zatorski@icahn.mssm.edu

https://doi.org/10.1016/j.isci.2024.110640

a complementary but different approach to gene and protein expression data. Structural features have formed the basis of a genetic disease protein network that encompasses 3,453 disorders and reveals nearly 300 new links between genes and disease.[18] Structural features, along with machine learning approaches, have also predicted protein-protein interactions.[19] This demonstrates that network-based approaches and structural features can be applied together to reveal previously unexpected insights. Besides their general use to construct vast interaction networks, protein structural features have also uncovered patterns in specific diseases and biological perturbations. One set of structural features in particular, i.e., the structural fold, family, and superfamily compositions of gene sets from transcriptomics data, captured similarities and differences among tissues and drug treated cell lines when combined with machine learning techniques.[17] Furthermore, structural features reveal the evolution of human lung adenocarcinoma identity over the course of disease progression when applied to single cell RNA sequencing data.[20] Use of structural features has also generated new insights into drug perturbations including the relationship between the toxicities of kinase inhibitors on cardiomyocyte-like cell lines.[17] Indeed, structural information can detect off target effects of drugs when structurally similar proteins were targeted.[21] This holds true for structure-derived representations of proteins, such as secondary structure, solvent accessibility, and intrinsically disordered regions (IDRs).[11]

Here we tested whether representation of the gene or protein set with a variety of structural features can describe phenotypes. We first developed Structural Analysis of Gene and protein Expression Signatures (SAGES) (Figure 1), a method of generating structural features from gene and protein sets. The novelty of this approach stems from the large number of new features, primarily those derived from 3D representations of proteins. Furthermore, the implementation of SAGES has been streamlined compared to other feature generation software[17] to include a database updating pipeline and fewer dependencies, all while generating significantly more features. We applied SAGES, in combination with random forest models and recursive feature elimination (RFE), to GTEx,[22] a dataset of RNA expression data taken from human tissue samples. We also analyzed SAGES of breast cancer gene and protein expression data from new patient cohorts, as well as breast cancer data from the Catalog of Somatic Mutations In Cancer (COSMIC)[23] database of mutated and overexpressed genes. Furthermore, we applied SAGES drug perturbation signatures from the Connectivity map[24] to investigate the similarity between existing breast cancer drugs and the breast cancer SAGES signature.

RESULTS

Structural features are predictive of normal tissue type

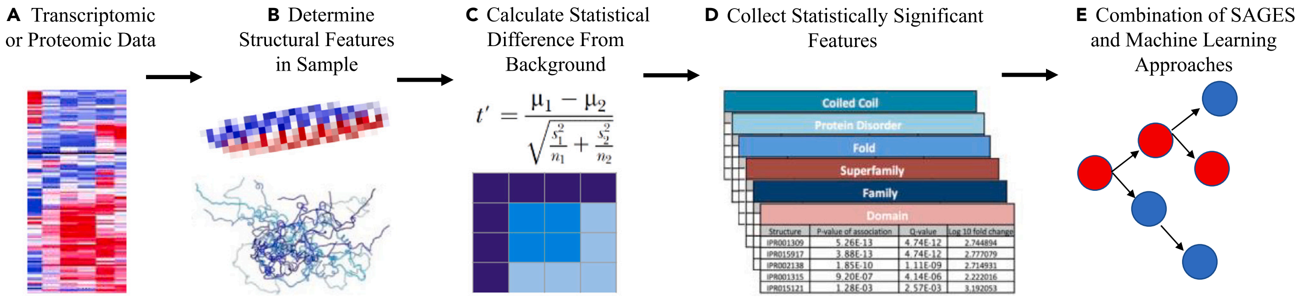
To test whether or not structural features can capture meaningful biological patterns, we evaluated the ability of SAGES to predict normal tissue from GTEx using only structural features and no gene name information. SAGES takes a set of genes or proteins and, using protein sequences and 3D models, determines the structural features of protein regions representative of the set (Figure 1). The use of 3D models sets SAGES apart from other feature generation methods.[17] SAGES was applied to the GTEx database to generate a structural feature set for each tissue sample. For each of the 30 tissue types in GTEx, we trained three random forest models using different input features, including structural features, gene names, and a combination of structural features with gene names (Figures S1 and S2). To reduce redundancy in protein similarity in the training and test sets, only the representative protein from each CD-HIT[25] cluster with a threshold of 40% was included in the features. We measured average performance for all tissues following 10-fold cross validation (Table 1). Notably, accuracy and AUROC are 97% for all groups, including those trained on a combination of structural features and gene names.

Interestingly, because performance of the model trained on genes was already high, a comparable performance with structural features alone that does not include any specific information on the genes is notable. For example, in the model trained on normal breast tissue SAGES the AUROC and F score were 0.949 ± 0.012 and 0.951 ± 0.011, respectively (Figure 2), while the gene name trained model obtained similar performance (0.949 ± 0.014 and 0.951 ± 0.013, respectively) (Table S2). This demonstrates that, while gene names contain adequate information for recapitulating tissue type, remarkably, structural features alone also contain sufficient information for recapitulating biological identity of tissues. The breast tissue prediction model trained on both types of features preformed similarly to the structural features trained model with AUROC and F-score of 0.951 ± 0.016 and 0.953 ± 0.015 (Figure 2). This is unsurprising considering that it uses the same structural features in addition to the gene names. SAGES was further applied to independent normal human tissue data from the ARCHS4[26] human and mouse tissue database. On average, predictive performance on the external dataset demonstrates that models trained on GTEx SAGES data outperform those trained on gene names or gene names and SAGES data combined. Predictor performance of normal tissue type using structural features compared to gene names alone demonstrates that the orthogonal information based on underlying protein characteristics is comparable to the information used in gene expression analysis.

Structural features reveal characteristics of normal breast tissue

To interrogate the contribution of different features to the model, we used RFE[27] to analyze the normal breast tissue predictor trained on a combination of structural features and genes from GTEx samples (Table 2). Using all features as input to RFE allows us to explore the relative importance of genes and structural features when used in tandem. We observe that prominent genes are differentially expressed in normal breast tissue, including APOD, a component of HDL; FABP4, a fatty acid binding protein; SAA1, serum amyloid; and TIMP3, a metallopeptidase inhibitor. The identification of genes such as APOD and FABP4, which are known to be associated with breast tissue, increases our confidence in the result.[28] Other genes such as TIMP3 and SAA1 are known more for their breast cancer prognostic value[29,30] and further support the robustness of the feature selection process.

Notably, features related to IDRs and conserved regions contributed greatly to the breast tissue predictor, even among all features, including gene names as well as other structural features (Table 2). These include the composition of glutamine, serine, and glutamate in


A Transcriptomic or Proteomic Data
B Determine Structural Features in Sample
C Calculate Statistical Difference From Background
$t' = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
D Collect Statistically Significant Features
Coiled Coil
Protein Disorder
Fold
Superfamily
Family
Domain
E Combination of SAGES and Machine Learning Approaches

Structure	P-value of association	Q-value	Log 10 fold change
IPR001309	5.26E-13	4.74E-12	2.744894
IPR015917	3.88E-13	4.74E-12	2.777079
IPR002138	1.85E-10	1.11E-09	2.714931
IPR001315	9.20E-07	4.14E-06	2.222016
IPR015121	1.28E-03	2.57E-03	3.192053

Figure 1. SAGES workflow

(A) SAGES takes as input a set of genes or proteins.

(B) For each protein, SAGES determines structural features of protein regions from protein sequences and 3D models.

(C) SAGES normalizes features by the magnitude of their corresponding protein's expression level.

(D) Sample structural features are then compared to a background using *p* values from a Fisher's exact test or Welch's t test. Bonferroni and false discovery rate corrected significance levels are provided based on the found feature size.

(E) The output is a vectorized depiction of the structural features of a transcriptomics or proteomics sample that can be used as the input for biological analysis of samples and machine learning applications.

the protein list; threonine and phenylalanine in conserved regions of proteins; glutamate residues in beta sheet regions; and histidine residues in the IDRs. In particular, this reveals that certain amino acids are differentially expressed in proteins that define breast tissue. For example, glutamine can potentially contribute to protein structural instability when found in IDRs,[32] and histidine, is known to be disorder promoting.[33] In normal breast tissue, many of the important genes selected through RFE are contributed to this abundance of IDRs such as FABP4,[34,35] SAA1,[36,37] and TIMP3.[38] Taken together, this demonstrates a trend of enrichment of proteins with IDRs and disorder related features in breast tissue.

Breast cancer structural features differ from those of normal breast tissue

Transcriptomic data reveal an overexpression of proteins with IDRs and intrinsically disordered binding regions: Experimentally derived samples

Human breast tumor samples and normal breast tissue controls were collected from 23 individuals during their surgical treatment procedures and sequenced. We compared SAGES of these biopsied breast tumors to those of normal tissue samples from the same individual to identify changes in the type of protein features overexpressed in the diseased tissue. The SAGES analysis of all 23 patients, as well as the following subsets: HER2 negative patients, progesterone receptor (PR) negative patients, estrogen receptor (ER) negative patients, Taxotere, Carboplatin, Herceptin, Perjeta (TCHP) treated patients, and AC-T treated patients, all demonstrated statistically significantly different protein features from background. Interestingly, the most highly represented features in overexpressed genes in breast cancer patients are long IDRs and IDRs that interact with other proteins (intrinsically disordered binding regions) (Figure 3A). This trend is observed both in the agglomeration of all 23 samples and specifically in HER2 negative patients (Figure 3B).

The important role of IDRs in tumorigenesis of a wide variety of cancers has previously been highlighted.[39] For example, many of the corresponding proteins with IDRs associated with cancer have been implicated in cell signaling pathways.[40] Consistent with these observations, in breast cancer in particular, many known signaling exhibit enrichment of IDRs such as HER2,[41] SIPA1,[42] and BRCA1.[43] An examination of domains, folds, superfamilies, and families that are expressed in the breast cancer samples but not their corresponding normal breast tissues reveals cell signaling related substructures such as the PDZ domain[44] and the tyrosine-protein kinase catalytic domain.[45] Additional features of note include zinc finger domains and immunoglobulin substructures, which are both implicated in breast cancer progression.[46,47] The immunoglobulin V-set domain is highly overexpressed in the set of breast cancer with PR negative disease and does not appear in HER2 or ER negative patients (Figure 3B). Interestingly, patients treated with TCHP had some features that were overexpressed compared to

Table 1. Performance of tissue type predictors with different input features

Features Trained on	Tissue Type	Accuracy	AUROC	F score	Precision	Recall
Structural Features and Gene Names	Average of all tissues	0.971 ± 0.064	0.971 ± 0.064	0.972 ± 0.056	0.969 ± 0.074	0.979 ± 0.05
Gene Names	Average of all tissues	0.968 ± 0.065	0.968 ± 0.065	0.97 ± 0.058	0.966 ± 0.076	0.976 ± 0.052
Structural Features	Average of all tissues	0.967 ± 0.063	0.967 ± 0.063	0.97 ± 0.054	0.965 ± 0.077	0.979 ± 0.049

Averaged performance of random forest tissue type predictors on all thirty GTEx tissues following 10-fold cross validation. The Structural Features rows mark the performance obtained when the model was trained using only structural features. The gene name row denotes the performance of the model trained on one hot encoded gene name symbols. The structural features and gene names row represent the performance of the model trained on both structural features and gene names.

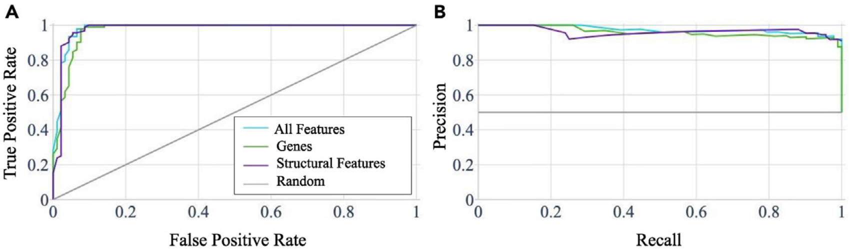


Figure 2. SAGES performance in predicting tissue from GTEx

Graphs depicting performance of the random forest predictor of normal breast tissue trained on structural features, gene names, and a combination of all features based on GTEx samples. The line labeled random represents performance of a model that has no skill and instead randomly selects a classification.

(A) The receiver operating characteristic curve (ROC) for the breast tissue prediction model trained on different features, which all perform comparably. The AUROC for this model is 94.9% for training on structural features, 95.1% for training on gene names, and 95.3% for training on all features.

(B) The precision recall curve for the breast tissue prediction model trained on different features. The precision and recall for each of the feature sets used to train the model are 92.8% and 97.5% for structural features, 92.6% and 98.0% for gene names, and 93.0% and 98.0% for all features combined.

background (Figure S3A), while those treated with AC-T only had features that were under expressed at a low magnitude compared to background (Figure S3B). These trends provide motivation for further exploration of SAGES features in drug treatment.

These experimentally determined breast cancer gene expression samples were further analyzed using GO with Bonferroni correction and the Benjamini-Hochberg[48] procedure. GO revealed no difference between diseased tissue and healthy breast tissue in the same patients. This differs from the SAGES analysis discussed previously, which demonstrated statistically significantly different features between diseased and non-diseased tissue. Use of GTEx breast tissue as the control for the breast cancer gene expression samples revealed overexpression of proteins involved in the cell cycle (0022402, 0007049) and metabolism (0090304, 0043170). As cancer is associated with cell cycle dysregulation and unchecked growth,[49] the abundance of gene expression with theses assignments is unsurprising. These genes correspond to proteins that are found throughout all parts of the cell including the nucleus (0005634), cytosol (0005829), membrane (0016020), and intracellular organelles (0043229). This wide spread localization is consistent with protein translocation during signaling processes.[50]

Transcriptomic data reveal an overexpression of proteins with IDR and intrinsically disordered binding regions: COSMIC database

SAGES captured breast cancer disease trends across datasets from multiple sources as seen in the analysis of breast cancer gene expression data from COSMIC,[23] an established cancer mutation database. Comparisons between mutated and unmutated but overexpressed breast cancer genes from COSMIC against normal breast tissue gene expression from GTEx demonstrated clear overexpression of IDRs and intrinsically disordered binding regions (Figure 4), in agreement with the analysis of the patient-derived breast cancer gene expression samples (Figure 3).

Further, SAGES of the COSMIC data also provide insight into the difference between breast cancer genes that are mutated and those that are overexpressed but have unchanged amino acid sequences. Particularly, the most common mutations include: missense substitutions, nonsense substitutions, frameshift insertions, and frameshift deletions.[23] Both sets (i.e., mutated genes and overexpressed genes) over express proteins with long IDRs and intrinsically disordered binding regions, and the mutated genes are also enriched with immunoglobulin subtype domains as compared to normal breast tissue. Overall, the mutated genes are more different, in terms of number of statistically significantly different features and log ratio of expression, from normal background breast tissue than the unmutated genes. Interestingly, both mutated and overexpressed gene sets are transmembrane helix depleted.

Addition of GO enrichment analysis further differentiates the roles associated with the mutated and unmutated but overexpressed breast cancer genes. Mutated breast cancer genes are associated with adhesion (such as 0007156, 0098742, 0031344), signaling (such as 0099177, 0050804, and 0007267), cellular differentiation (such as 0000902, 0007420, and 0048858), metabolism-including those of lipids, (such as 0019222, 0044237, and 0044238), signaling related to immune response (such as 0002757, 0002429, and 0002768), and regulation of cell death (0010941). Localization of these proteins primarily coincides with the plasma membrane (e.g., 0098590) and the extracellular region (e.g., 0005615). Notably, the mutated genes were implicated in processes related to adhesion. This is unsurprising due to the loss of cellular adhesion associated with most cancers.[51] Overexpressed breast cancer genes that have no amino acid sequence changes are associated with metabolism (such as 0008152, 0044238, and 0006629), development (such as 0030154), signaling (such as 0023052 and 0007165), regulation of cell death (0010942), and immune response (such as 0045087). Some of these proteins are associated with the extracellular space (0005615), however, many others are associated with the endoplasmic reticulum and ribosomes (e.g., 0005788 and 0005840). The localization to organelles related to protein synthesis is supported by the known increased metabolic activity of cancer cells.[52] Importantly, both of these RNA expression derived samples have processes related to signaling.

Proteomic data reveal an overrepresentation of transmembrane proteins

Proteomic data, which can reveal differential protein abundance when compared to transcriptomic data,[53] provides a unique view of expressed breast cancer protein features. This supplements our understanding of breast cancer proteins, which are revealed through gene

Table 2. Ranking and description of features found most informative in the breast tissue predictor

Rank[a]	Feature Name[b]	Feature Description[c]
1.3	FABP4	Fatty acid binding protein
1.7	APOD	Component of HDL
3.3	SAA1	Serum amyloid
6.6	Number of glutamines in protein	Structural feature
8.9	AZGP1	Zinc binding glycoprotein
9.8	Number of threonines in conserved regions	Structural feature
9.8	CD59	CD59 blood group
13.9	Number of phenylalanines in conserved regions	Structural feature
16	RPL23	Ribosomal protein
18.1	XBP1	X box binding protein
18.4	Number of serines in protein	Structural feature
18.7	MGP	Matrix Gla Protein
19.7	VIM	Vimentin
22.2	TIMP3	TIMP metallopeptidase inhibitor 3
22.8	TXNIP	thioredoxin interacting protein
23.3	C7	complement c7
25.3	IGFBP4	insulin like growth factor binding protein 4
26.5	SERPING1	serpin family g member 1
29.1	Number of glutamic acids in protein	Structural feature
32.1	PNPLA2	Patatin Like Phospholipase Domain Containing 2
32.2	C10orf10	DEPP1 Autophagy Regulator
32.5	Number of glutamic acids in sheet region	Structural feature
34.6	GSN	Gelsolin
34.6	RPLP2	Ribosomal Protein Lateral Stalk Subunit P2
37.4	Number of histidines in disordered regions	Structural feature

The twenty-five features that contributed most to the random forest model predicting normal breast tissue from structural features and gene names of GTEx samples.

[a] *Rank* marks the lowest value of rank corresponds to the feature that contributes most to the model as assigned using RFE with 10 random seeds.

[b] Feature Name represents contains the gene name symbol or structural feature name that corresponds with the rank in that same row.

[c] Feature Description corresponds to additional information about the function or name of the feature in that row. Gene feature descriptors were summarized from the Human Gene Database Gene Cards version 5.10.[31]

expression analysis. Proteomics data from 17 breast cancer tumors with normal tissues from the same patients was analyzed using SAGES.[54] The result showed the presence of proteins with transmembrane helix regions containing a larger number of amino acids on average and immunoglobulin domains (Figure 5). The proteomics data revealed different patterns in features from those calculated using gene expression signatures from the 23 different breast cancer patients sequenced for this study and the COSMIC gene expression data such as presence of the kinase regulatory SH3 domain,[55] with the exception of the immunoglobulin like domain. Interestingly, both domain types are often found in receptor tyrosine kinases (RTKs), which play a key role in signaling.[56] Notably, features from the proteomics breast cancer dataset had shorter IDRs and binding regions that are intrinsically disordered. Because these partially intrinsically disordered proteins are often involved in cell signaling, such as RTKs, they are tightly regulated and have short half-life.[57] The difference in structural features found in the analysis of gene and protein expression of breast cancer therefore can be attributed to the transient nature of signaling proteins, which contain IDRs.

This absence of transient signaling proteins in the proteomic sample is reflected in the GO analysis. On a biological process level, the proteins in breast cancer had functions related to metabolic processes (e.g., 0044281) and transcription (e.g., 0006351, 0097659). One of the areas where these proteins localize in the cell is the nucleus, specifically chromatin (0000785) and chromosomes (0005694). This aligns with the biological processes related to transcription. The proteins also localize to the extracellular space (0005615), which could be related to the increased presence of proteins with transmembrane helices, and the endoplasmic reticulum lumen (0031093). Notably absent from the overrepresented proteins found in the proteomic analysis of breast cancer are the GO terms

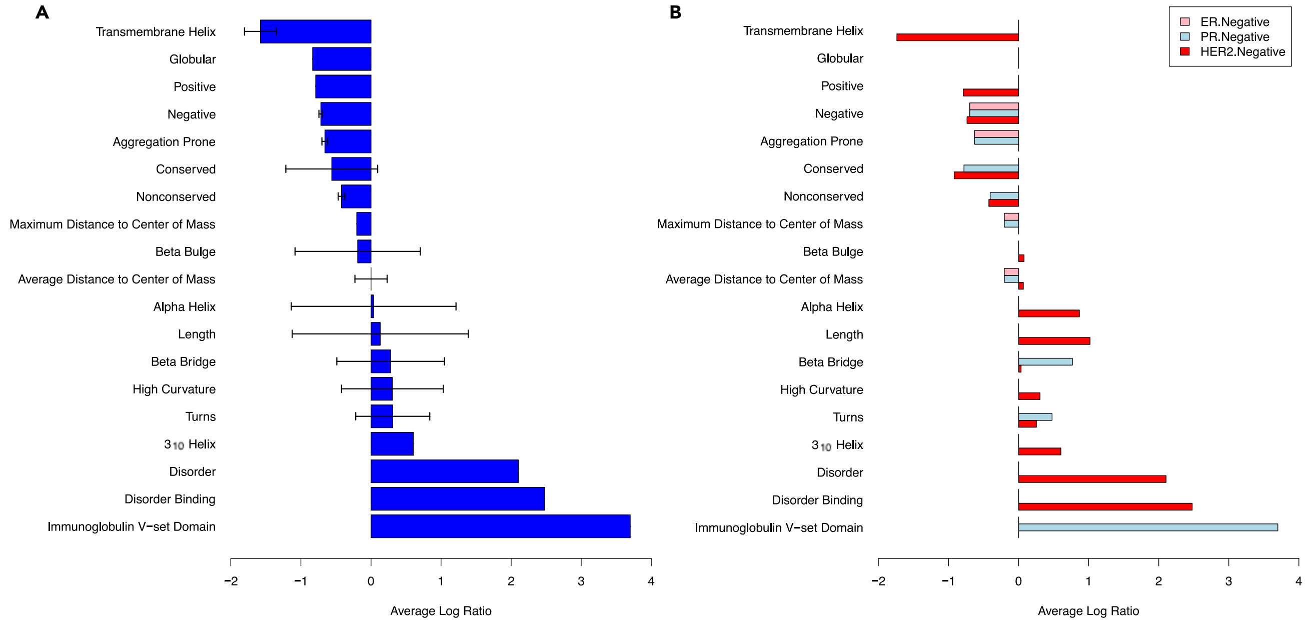


Figure 3. Structural features of normal and breast cancer tissue from newly generated human patient samples

Average log ratios of SAGES features that are statistically significantly different in breast cancer samples as compared to normal breast tissue. The log ratio of sample divided by background for each feature was averaged if the calculated *p* value for that feature was less than the Bonferroni adjusted significance level. Error bars represent the standard deviation of the values used to calculate the average. If there was only one instance of feature found to be statistically significant, the error bars were set to zero. Features that are not families, folds, superfamilies, domains, or frequency counts of a secondary structure consist of the number of amino acids within the protein that make up the secondary structure.

(A) The average log ratio of SAGES features for gene expression from 23 breast cancer tumors compared to normal breast tissue from the same individuals.

(B) The average log ratio of SAGES features for gene expression from 23 breast cancer patients separated by negative receptor status.

related to any cellular signaling, as well as cellular adhesion. Taken together, our results suggest that combining analyzing using enrichments of both GO and structural features generated with SAGES supports the correlation between lack of signaling molecules and IDRs in the proteomic samples.

SAGES captures similarity between perturbation signatures of existing breast cancer therapies and disease

It has been proposed that gene signatures can be applied for drug repurposing.[58] In particular, the signature reversion principle (SRP) is based on the assumption that drug perturbation gene expression signature has a negative correlation with a disease gene expression signature.[58] To investigate the extension of the SRP, we applied SAGES to gene and protein expression changes induced by breast-cancer drugs and breast cancer. The number of statistically significantly different SAGES features between the drug perturbation and both genomic and proteomic breast cancer analyses were calculated and averaged for breast cancer treatments and all other therapeutics. Comparisons between drug perturbation with proteomic and genomic breast cancer backgrounds show that breast cancer drugs produce a perturbation in cell lines with SAGES features that are more similar to those obtained from breast cancer than those obtained from other drugs (Figure 5A). With the proteomic background, the breast cancer drugs had a lower average number of SAGES features with statistically significant differences (314) than other drugs (321) (p = 0.00070). With the genomic background, the breast cancer drugs had a lower average number of features with statistically significant differences (323) than other drugs (327) (p = 0.050).

This increased similarity between drug perturbation and breast cancer is captured in a gene-protein level comparison between the perturbations and backgrounds but not in a gene-gene level comparison (Figure 6B). When the Jaccard coefficients of the breast cancer (0.0060) and other (0.0055) drug genes are compared to the proteomic breast cancer background, there is smore similarity to the background in the breast cancer drug genes (p = 0.00102). However, this trend of SAGES features and proteins of breast cancer drug perturbations being more similar to the targeted disease is not recapitulated in the comparison of gene expression of drug perturbations and breast cancer for both drug groups. Breast cancer drugs have an average Jaccard coefficient of 0.00575 compared to other drugs with a Jaccard coefficient of 0.00572 (p = 0.77). Because drug repurposing using perturbation signatures is classically done at a gene level,[59] the emergence of this association at the protein and structural level demonstrates the power of the orthogonal information provided with SAGES and raises a potential application of SAGES in the signature based repurposing sphere.

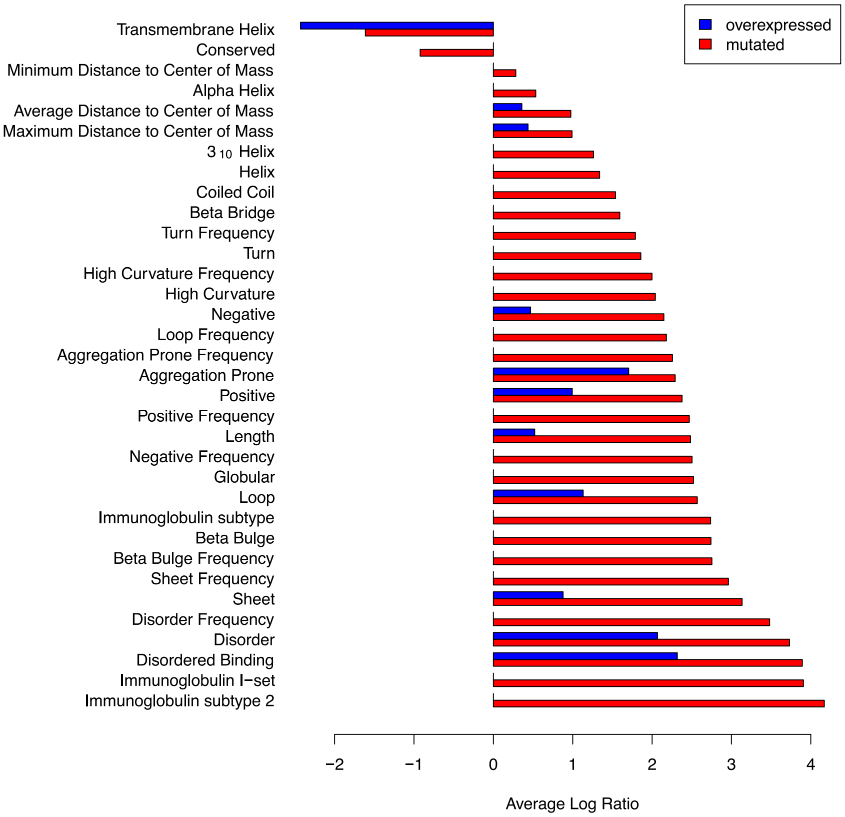


Figure 4. Structural features of normal and breast cancer tissue from the COSMIC database

The average log ratio of SAGES features for gene expression of mutated and unmutated but over expressed genes from COSMIC breast cancer tumors compared to normal breast tissue from GTEx. Features are displayed by size of average log ratio level for mutated genes, where positive values correspond to features with increased length or frequency in breast cancer compared to normal breast tissue. Due to the fact that the COSMIC data were based on a set of samples combined before analysis, there are no error bars.

DISCUSSION

Genomic and proteomic data contains valuable information about biological states. Currently, gene expression is assessed on a gene name similarity basis or through the use of enrichment of GO annotations in gene lists, however, these existing methods do not capture underlying structure based functional information.

Application of features derived from different levels of protein structure can greatly enhance transcriptomic analysis. SAGES is distinct from existing methods in that it provides both sequence and 3D structure based orthogonal data, which supplement transcriptomic biological insight (Figure 1). Furthermore, SAGES generates a large number of these features. These detailed, structural level, descriptions of proteins encoded by the genes in the input protein dataset provide orthogonal information to gene names and GO functional descriptions.

The analysis of normal tissue transcriptomic data from GTEx revealed the generalizability of structural features to all tissue types with the development of tissue type random forest predictors (Table 1). Use of features alone captured just as much information about biological state as the traditionally used gene names. Feature selection applied to the normal breast tissue samples further revealed that, interestingly, amino acid composition of IDRs was shown to play an informative role in breast tissue prediction (Table 2).

We applied SAGES to multiple breast cancer datasets to demonstrate how this method can be used to inform our understanding of breast tissue and tumor biology. SAGES, applied to newly collected samples excised from 23 women during their breast cancer treatment surgeries, revealed structural and functional differences between tumor and normal tissue. Notably, proteins with IDRs and intrinsically disordered binding regions are overexpressed in diseased breast cancer tissue (Figure 3). This was supported by SAGES analysis of breast cancer genes that are mutated and overexpressed in breast cancer according to COSMIC (Figure 4). GO analysis of these samples reveals that the proteins that are overexpressed in breast cancer and contain these IDR features are associated with cell signaling, metabolism, and immune response. Uniquely, mutated breast cancer proteins also were associated with cellular adhesion. Proteomic analysis of human breast cancer samples revealed a different feature landscape, which was correlated with a loss of representation of proteins related to cell signaling (Figure 5). This is thought to be due to the transient nature of these cell signaling proteins. This SAGES difference can be traced to the underlying experimental gene expression and protein expression results (Figure S4; Table S4). Analysis with Jaccard similarity[60] of breast tissue genomic and

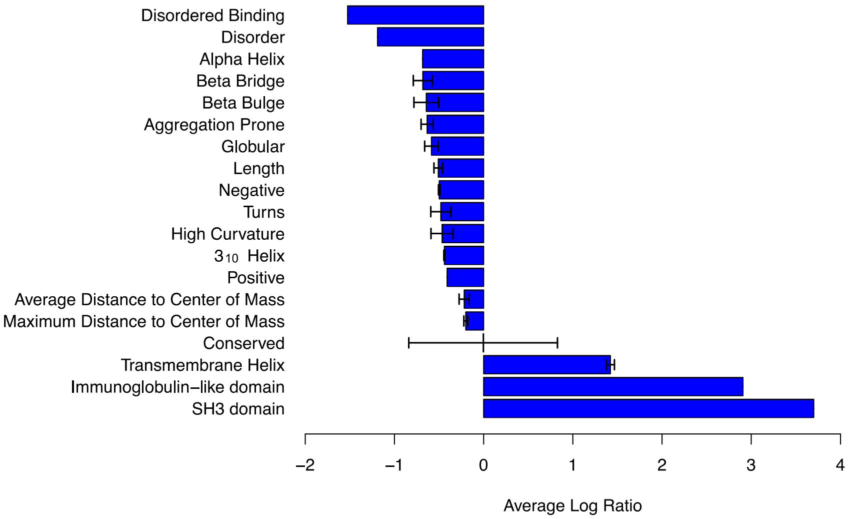


Figure 5. Structural features of primary breast tumors compared to matched normal tissues from a global proteomics dataset quantified using mass spectrometry

The average log ratio of SAGES features for proteomics of 17 breast cancer tumors extracted from surgical patients compared to normal breast tissue from the same patients. Negative values correspond to features with decreased length or frequency in breast cancer compared to normal breast tissue.
Error bars represent the standard deviation of the values used to calculate the average. In the case of a single instance of a feature having statistical signficance, the error bars are set to zero.

proteomic data demonstrates that there is very little overlap between the top 250 overexpressed genes (Figure S4A). Remarkably, application of SAGES improves similarity of breast cancer signatures across data sources (Figure S4B), once again demonstrating the usefulness of this tool in biological analysis.

SAGES was used to interrogate drug perturbation signatures from the Connectivity Map and the resulting analysis showed breast cancer drug perturbation signatures are more similar to breast cancer expression signatures on the feature and protein level than other drugs (Figure 6). This, along with SAGES analysis of notable features in breast cancer patients before receiving AC-T or TCHP treatment (Figure S3), has interesting implications for SRP. This further demonstrates the capacity for structural features to capture both established and novel underlying biological function. The scope of SAGES is wide and we anticipate that it has the potential for a vast number of future applications.

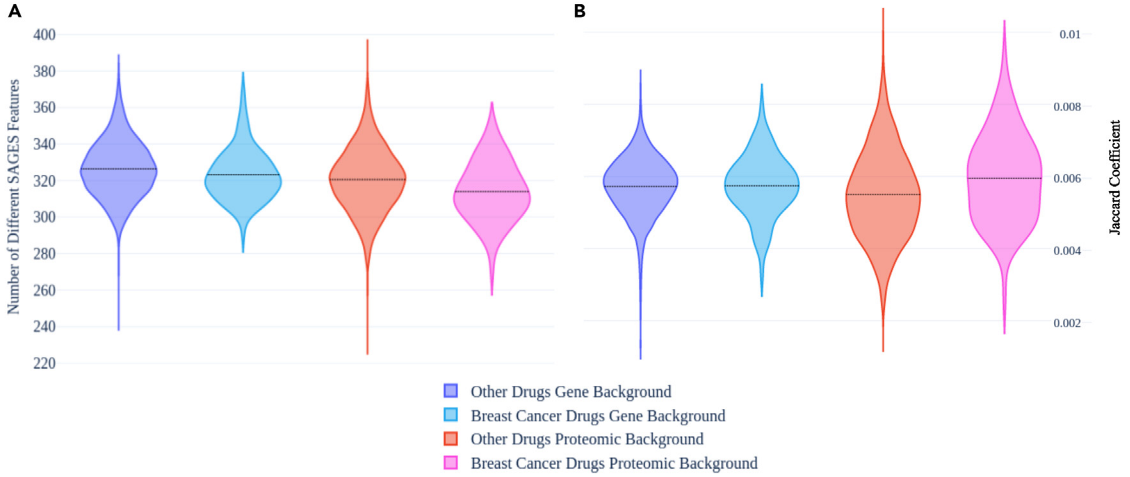


Figure 6. Perturbation signatures of existing breast cancer drugs and drugs for all other indications compared to breast cancer signatures

Difference between connectivity map breast cancer and other drug perturbation signatures compared to signatures derived from breast cancer COSMIC gene expression and experimental proteomics expression.
(A) Counts of statistically significant different features between SAGES of breast cancer and drug perturbation.
(B) Jaccard coefficient representing similarity of drug perturbation signatures to gene and protein breast cancer signature backgrounds.

Limitations of the study

This work investigates the structural features of genomic and proteomic signatures of tissues with an emphasis on breast cancer and tissue. Original experimental data for genomic and proteomic samples originates from different groups of 23 and 17 individuals, respectively. In order to increase the sample size and therefore the generalizability of this study, additional open-source genomic and proteomic datasets were utilized.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Structural features
  - ○ Breast cancer proteomics and transcriptomics
- METHOD DETAILS
  - ○ Structural feature generation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Structural feature statistical analysis
  - ○ Prediction of normal tissues
  - ○ Breast cancer analysis
  - ○ Breast cancer drug perturbation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110640.

ACKNOWLEDGMENTS

We would like to thank Michael Bernhofer for his assistance with manuscript review and Nolan Caile for his assistance with database curation. This work is supported in part by the National Institutes of Health grant F30 HL160179 to N.Z. and R01 HD107528 as well as U01 CA271318 to A.S.

AUTHOR CONTRIBUTIONS

N.Z. developed and used SAGES to analyze the datasets. N.Z. and A.S. wrote the manuscript. Y.S. and M.W. generated the breast cancer gene expression data. A.E. and K.L.H. provided the processed proteomic breast cancer data. N.Z., C.D., M.B., T.K., B.R., and D.S. contributed to feature generation.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 25, 2023
Revised: May 5, 2024
Accepted: July 30, 2024
Published: August 7, 2024

REFERENCES

1. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science *313*, 1929–1935. https://doi.org/10.1126/science.1132939.
2. Heidecker, B., and Hare, J.M. (2007). The use of transcriptomic biomarkers for personalized medicine. Heart Fail. Rev. *12*, 1–11. https://doi.org/10.1007/s10741-007-9004-7.
3. Merry, E., Thway, K., Jones, R.L., and Huang, P.H. (2021). Predictive and prognostic transcriptomic biomarkers in soft tissue sarcomas. npj Precis. Oncol. *5*, 17. https://doi.org/10.1038/s41698-021-00157-4.
4. Chung, W., Eum, H.H., Lee, H.O., Lee, K.M., Lee, H.B., Kim, K.T., Ryu, H.S., Kim, S., Lee, J.E., Park, Y.H., et al. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat. Commun. *8*, 15081. https://doi.org/10.1038/ncomms15081.
5. Liu, W., Liu, J., and Rajapakse, J.C. (2018). Gene Ontology Enrichment Improves Performances of Functional Similarity of Genes. Sci. Rep. *8*, 12100. https://doi.org/10.1038/s41598-018-30455-0.

6. Zeidan, B.A., Townsend, P.A., Garbis, S.D., Copson, E., and Cutress, R.I. (2015). Clinical proteomics and breast cancer. Surgeon *13*, 271–278. https://doi.org/10.1016/j.surge.2014.12.003.
7. Grimes, T., Potter, S.S., and Datta, S. (2019). Integrating gene regulatory pathways into differential network analysis of gene expression data. Sci. Rep. *9*, 5479. https://doi.org/10.1038/s41598-019-41918-3.
8. Bernhofer, M., Dallago, C., Karl, T., Satagopam, V., Heinzinger, M., Littmann, M., Olenyi, T., Qiu, J., Schütze, K., Yachdav, G., et al. (2021). PredictProtein - Predicting Protein Structure and Function for 29 Years. Nucleic Acids Res. *49*, W535–W540. https://doi.org/10.1093/nar/gkab354.
9. Fox, N.K., Brenner, S.E., and Chandonia, J.M. (2014). SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. *42*, D304–D309. https://doi.org/10.1093/nar/gkt1240.
10. Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637. https://doi.org/10.1002/bip.360221211.
11. Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. Nat. Rev. Mol. Cell Biol. *8*, 995–1005. https://doi.org/10.1038/nrm2281.
12. Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O., and Ofran, Y. (2003). Automatic prediction of protein function. Cell. Mol. Life Sci. *60*, 2637–2650. https://doi.org/10.1007/s00018-003-3114-8.
13. Gerstein, M., and Levitt, M. (1997). A structural census of the current population of protein sequences. Proc. Natl. Acad. Sci. USA *94*, 11911–11916. https://doi.org/10.1073/pnas.94.22.11911.
14. Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., et al. (2013). A large-scale evaluation of computational protein function prediction. Nat. Methods *10*, 221–227. https://doi.org/10.1038/nmeth.2340.
15. Rost, B., Radivojac, P., and Bromberg, Y. (2016). Protein function in precision medicine: deep understanding with machine learning. FEBS Lett. *590*, 2327–2341. https://doi.org/10.1002/1873-3468.12307.
16. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
17. Rahman, R., Zatorski, N., Hansen, J., Xiong, Y., van Hasselt, J.G.C., Sobie, E.A., Birtwistle, M.R., Azeloglu, E.U., Iyengar, R., and Schlessinger, A. (2021). Protein structure-based gene expression signatures. Proc. Natl. Acad. Sci. USA *118*, e2014866118. https://doi.org/10.1073/pnas.2014866118.
18. Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat. Biotechnol. *30*, 159–164. https://doi.org/10.1038/nbt.2106.
19. Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., et al. (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature *490*, 556–560. https://doi.org/10.1038/nature11503.
20. Zatorski, N., Stein, D., Rahman, R., Iyengar, R., and Schlessinger, A. (2022). Structural signatures: a web server for exploring a database of and generating protein structural features from human cell lines and tissues. Database *2022*, baac053. https://doi.org/10.1093/database/baac053.
21. MacDonald, M.L., Lamerdin, J., Owens, S., Keon, B.H., Bilter, G.K., Shang, Z., Huang, Z., Yu, H., Dias, J., Minami, T., et al. (2006). Identifying off-target effects and hidden phenotypes of drugs in human cells. Nat. Chem. Biol. *2*, 329–337. https://doi.org/10.1038/nchembio790.
22. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.
23. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. *47*, D941–D947. https://doi.org/10.1093/nar/gky1015.
24. Lamb, J. (2007). The Connectivity Map: a new tool for biomedical research. Nat. Rev. Cancer *7*, 54–60. https://doi.org/10.1038/nrc2044.
25. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.
26. Lachmann, A., Torre, D., Keenan, A.B., Jagodnik, K.M., Lee, H.J., Wang, L., Silverstein, M.C., and Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. Nat. Commun. *9*, 1366. https://doi.org/10.1038/s41467-018-03751-6.
27. Darst, B.F., Malecki, K.C., and Engelman, C.D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet. *19*, 65. https://doi.org/10.1186/s12863-018-0633-8.
28. Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N., Adori, C., Oksvold, P., Edfors, F., Limiszewska, A., Hikmet, F., et al. (2020). An atlas of the protein-coding genes in the human, pig, and mouse brain. Science *367*, eaay5947. https://doi.org/10.1126/science.aay5947.
29. Danforth, D.N., Jr. (2016). Genomic Changes in Normal Breast Tissue in Women at Normal Risk or at High Risk for Breast Cancer. Breast Cancer *10*, 109–146. https://doi.org/10.4137/BCBCR.S39384.
30. Ignacio, R.M.C., Gibbs, C.R., Kim, S., Lee, E.S., Adunyah, S.E., and Son, D.S. (2019). Serum amyloid A predisposes inflammatory tumor microenvironment in triple negative breast cancer. Oncotarget *10*, 511–526. https://doi.org/10.18632/oncotarget.26566.
31. Barshir, R., Fishilevich, S., Iny-Stein, T., Zelig, O., Mazor, Y., Guan-Golan, Y., Safran, M., and Lancet, D. (2021). GeneCaRNA: A Comprehensive Gene-centric Database of Human Non-coding RNAs in the GeneCards Suite. J. Mol. Biol. *433*, 166913. https://doi.org/10.1016/j.jmb.2021.166913.
32. Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett. *579*, 3346–3354. https://doi.org/10.1016/j.febslet.2005.03.072.
33. Zhao, J., Zou, L., Li, Y., Liu, X., Zeng, C., Xu, C., Jiang, B., Guo, X., and Song, X. (2021). HisPhosSite: A comprehensive database of histidine phosphorylated proteins and sites. J. Proteomics *243*, 104262. https://doi.org/10.1016/j.jprot.2021.104262.
34. Gillilan, R.E., Ayers, S.D., and Noy, N. (2007). Structural basis for activation of fatty acid-binding protein 4. J. Mol. Biol. *372*, 1246–1260. https://doi.org/10.1016/j.jmb.2007.07.040.
35. Wang, J., Tang, J., Wang, B., Song, J., Liu, J., Wei, Z., Zhang, F., Ma, X., and Cao, Y. (2009). FABP4: a novel candidate gene for polycystic ovary syndrome. Endocrine *36*, 392–396. https://doi.org/10.1007/s12020-009-9228-5.
36. Ghelichkhani, F., Gonzalez, F.A., Kapitonova, M.A., Schaefer-Ramadan, S., Liu, J., Cheng, R., and Rozovsky, S. (2022). Selenoprotein S: A versatile disordered protein. Arch. Biochem. Biophys. *731*, 109427. https://doi.org/10.1016/j.abb.2022.109427.
37. Stevens, F.J. (2004). Hypothetical structure of human serum amyloid A protein. Amyloid *11*, 71–80. https://doi.org/10.1080/13506120412331272296.
38. Lee, K.E., Procopio, R., Pulido, J.S., and Gunton, K.B. (2023). Initial Investigations of Intrinsically Disordered Regions in Inherited Retinal Diseases. Int. J. Mol. Sci. *24*, 1060. https://doi.org/10.3390/ijms24021060.
39. Mészáros, B., Hajdu-Soltesz, B., Zeke, A., and Dosztanyi, Z. (2021). Mutations of Intrinsically Disordered Protein Regions Can Drive Cancer but Lack Therapeutic Strategies. Biomolecules *11*, 381. https://doi.org/10.3390/biom11030381.
40. Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradović, Z., and Dunker, A.K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. J. Mol. Biol. *323*, 573–584. https://doi.org/10.1016/s0022-2836(02)00969-5.
41. Wright, P.E., and Dyson, H.J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. Nat. Rev. Mol. Cell Biol. *16*, 18–29. https://doi.org/10.1038/nrm3920.
42. Ma, Y., Weng, J., Wang, N., Zhang, Y., Minato, N., and Su, L. (2021). A novel nuclear localization region in SIPA1 determines protein nuclear distribution and epirubicin-sensitivity of breast cancer cells. Int. J. Biol. Macromol. *180*, 718–728. https://doi.org/10.1016/j.ijbiomac.2021.03.101.
43. Mark, W.Y., Liao, J.C.C., Lu, Y., Ayed, A., Laister, R., Szymczyna, B., Chakrabartty, A., and Arrowsmith, C.H. (2005). Characterization of segments from the central region of BRCA1: an intrinsically disordered scaffold for multiple protein-protein and protein-DNA interactions? J. Mol. Biol. *345*, 275–287. https://doi.org/10.1016/j.jmb.2004.10.045.
44. Jeleń, F., Oleksy, A., Smietana, K., and Otlewski, J. (2003). PDZ domains - common players in the cell signaling. Acta Biochim. Pol. *50*, 985–1017.
45. Li, E., and Hristova, K. (2006). Role of receptor tyrosine kinase transmembrane domains in cell signaling and human pathologies. Biochemistry *45*, 6241–6251. https://doi.org/10.1021/bi060609y.
46. Jen, J., and Wang, Y.C. (2016). Zinc finger proteins in cancer progression. J. Biomed. Sci. *23*, 53. https://doi.org/10.1186/s12929-016-0269-9.

47. Whiteside, T.L., and Ferrone, S. (2012). For breast cancer prognosis, immunoglobulin kappa chain surfaces to the top. Clin. Cancer Res. *18*, 2417–2419. https://doi.org/10.1158/1078-0432.CCR-12-0566.
48. Hu, J.X., Zhao, H., and Zhou, H.H. (2010). False Discovery Rate Control With Groups. J. Am. Stat. Assoc. *105*, 1215–1227. https://doi.org/10.1198/jasa.2010.tm09329.
49. Zafonte, B.T., Hulit, J., Amanatullah, D.F., Albanese, C., Wang, C., Rosen, E., Reutens, A., Sparano, J.A., Lisanti, M.P., and Pestell, R.G. (2000). Cell-cycle dysregulation in breast cancer: breast cancer therapies targeting the cell cycle. Front. Biosci. *5*, D938–D961. https://doi.org/10.2741/zafonte.
50. Teruel, M.N., and Meyer, T. (2000). Translocation and reversible localization of signaling proteins: a dynamic future for signal transduction. Cell *103*, 181–184. https://doi.org/10.1016/s0092-8674(00)00109-4.
51. Cavallaro, U., and Christofori, G. (2004). Cell adhesion and signalling by cadherins and Ig-CAMs in cancer. Nat. Rev. Cancer *4*, 118–132. https://doi.org/10.1038/nrc1276.
52. DeBerardinis, R.J., and Chandel, N.S. (2016). Fundamentals of cancer metabolism. Sci. Adv. *2*, e1600200. https://doi.org/10.1126/sciadv.1600200.
53. Pascal, L.E., True, L.D., Campbell, D.S., Deutsch, E.W., Risk, M., Coleman, I.M., Eichner, L.J., Nelson, P.S., and Liu, A.Y. (2008). Correlation of mRNA and protein levels: cell type-specific gene expression of cluster designation antigens in the prostate. BMC Genom. *9*, 246. https://doi.org/10.1186/1471-2164-9-246.
54. Elmas, A., Tharakan, S., Jaladanki, S., Galsky, M.D., Liu, T., and Huang, K.L. (2021). Pan-cancer proteogenomic investigations identify post-transcriptional kinase targets. Commun. Biol. *4*, 1112. https://doi.org/10.1038/s42003-021-02636-7.
55. Kurochkina, N., and Guha, U. (2013). SH3 domains: modules of protein-protein interactions. Biophys. Rev. *5*, 29–39. https://doi.org/10.1007/s12551-012-0081-z.
56. Lemmon, M.A., and Schlessinger, J. (2010). Cell signaling by receptor tyrosine kinases. Cell *141*, 1117–1134. https://doi.org/10.1016/j.cell.2010.06.011.
57. Gsponer, J., Futschik, M.E., Teichmann, S.A., and Babu, M.M. (2008). Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science *322*, 1365–1368. https://doi.org/10.1126/science.1163581.
58. Pushpakom, S., Iorio, F., Eyers, P.A., Escott, K.J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., et al. (2019). Drug repurposing: progress, challenges and recommendations. Nat. Rev. Drug Discov. *18*, 41–58. https://doi.org/10.1038/nrd.2018.168.
59. Hodos, R.A., Kidd, B.A., Shameer, K., Readhead, B.P., and Dudley, J.T. (2016). In silico methods for drug repurposing and pharmacology. Wiley Interdiscip. Rev. Syst. Biol. Med. *8*, 186–210. https://doi.org/10.1002/wsbm.1337.
60. Tang, M., Kaymaz, Y., Logeman, B.L., Eichhorn, S., Liang, Z.S., Dulac, C., and Sackton, T.B. (2021). Evaluating single-cell cluster stability using the Jaccard similarity index. Bioinformatics *37*, 2212–2214. https://doi.org/10.1093/bioinformatics/btaa956.
61. Mészáros, B., Erdos, G., and Dosztanyi, Z. (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. *46*, W329–W337. https://doi.org/10.1093/nar/gky384.
62. Mészáros, B., Simon, I., and Dosztanyi, Z. (2009). Prediction of protein binding regions in disordered proteins. PLoS Comput. Biol. *5*, e1000376. https://doi.org/10.1371/journal.pcbi.1000376.
63. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. *305*, 567–580. https://doi.org/10.1006/jmbi.2000.4315.
64. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.
65. Gabler, F., Nam, S.Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas, A.N., and Alva, V. (2020). Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr. Protoc. Bioinf. *72*, e108. https://doi.org/10.1002/cpbi.108.
66. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J.L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.
67. Kuriata, A., Iglesias, V., Pujols, J., Kurcinski, M., Kmiecik, S., and Ventura, S. (2019). Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. Nucleic Acids Res. *47*, W300–W307. https://doi.org/10.1093/nar/gkz321.
68. Rengasamy, M., Zhang, F., Vashisht, A., Song, W.M., Aguilo, F., Sun, Y., Li, S., Zhang, W., Zhang, B., Wohlschlegel, J.A., and Walsh, M.J. (2017). The PRMT5/WDR77 complex regulates alternative splicing through ZNF326 in breast cancer. Nucleic Acids Res. *45*, 11106–11120. https://doi.org/10.1093/nar/gkx727.
69. Krug, K., Jaehnig, E.J., Satpathy, S., Blumenberg, L., Karpova, A., Anurag, M., Miles, G., Mertins, P., Geffen, Y., Tang, L.C., et al. (2020). Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. Cell *183*, 1436–1456.e31. https://doi.org/10.1016/j.cell.2020.10.036.
70. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.
71. Rosner, B. (2011). Fundamentals of Biostatistics, 7th Edition (Brooks/Cole, Cengage Learning).
72. Armstrong, R.A. (2014). When to use the Bonferroni correction. Ophthalmic Physiol. Opt. *34*, 502–508. https://doi.org/10.1111/opo.12131.
73. van Iterson, M., Boer, J.M., and Menezes, R.X. (2010). Filtering, FDR and power. BMC Bioinf. *11*, 450. https://doi.org/10.1186/1471-2105-11-450.
74. Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., and Thomas, P.D. (2019). Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat. Protoc. *14*, 703–721. https://doi.org/10.1038/s41596-019-0128-8.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Breast cancer gene expression	Author generated data	NCBI GEO: GSE227679
Software and algorithms		
SAGES	Presented here and GitHub:https://github.com/schlessinger-lab/structural_features	1.0
Normal tissue predictor	Presented here and GitHub: https://github.com/schlessinger-lab/Structural-Analysis-of-Genomic-and-Proteomic-Breast-Cancer-Signatures	1.0
Tissue feature selection	Presented here and GitHub: https://github.com/schlessinger-lab/Structural-Analysis-of-Genomic-and-Proteomic-Breast-Cancer-Signatures	1.0
Data cleaning code	Presented here and GitHub: https://github.com/schlessinger-lab/Structural-Analysis-of-Genomic-and-Proteomic-Breast-Cancer-Signatures	1.0
Paper figure generation code	Presented here and GitHub: https://github.com/schlessinger-lab/Structural-Analysis-of-Genomic-and-Proteomic-Breast-Cancer-Signatures	1.0
Disordered regions	IUPRED	2a; RRID: SCR_014632
Disordered binding regions	ANCHOR	1
Transmembrane helical regions	TMHMM	2.0; RRID: SCR_014935
Globular regions	PredictProtein	2021
Helical regions	PredictProtein	2021
Positive amino acid containing regions	PredictProtein	2021
Negative amino acid containing regions	PredictProtein	2021
Coiled-coil regions	PredictProtein	2021
Sheet regions	PredictProtein	2021
Loop regions	PredictProtein	2021
Conserved regions	PredictProtein	2021
Non-conserved regions	PredictProtein	2021
310 helical regions	DSSP	2; RRID: SCR_002725
Alpha helical regions	DSSP	2; RRID: SCR_002725
Beta bridge regions	DSSP	2; RRID: SCR_002725
Beta bulge regions	DSSP	2; RRID: SCR_002725
Turn regions	DSSP	2; RRID: SCR_002725
High curvature regions	DSSP	2; RRID: SCR_002725
Distances from the center of mass	biopython's pdb parser	1.79; RRID: SCR_007173
Aggregation prone regions	Aggrescan3d	2.0
Gene ontology	Panther classification system	18.0; RRID: SCR_002811
Other		
Normal tissue RNA expression data	The Genotype-Tissue Expression project (GTEx)	8.0; RRID: SCR_013942
Independent normal tissue RNA expression data for machine learning predictor	RNA-seq and ChIP-seq sample and signature search (ARCHS[4])	Downloaded July 2, 2023; RRID: SCR_015683

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Catalogue of Somatic Mutations In Cancer	COSMIC	V97; RRID: SCR_002260
Connectivity Map	LINCS	NCBI GEO: GSE92743, GSE70138, GSE92743; RRID: SCR_002639
Breast cancer proteomics	GDC data portal: https://portal.gdc.cancer.gov/	RNA-seq data for CPTAC cohorts; RRID: SCR_014514

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Nicole Zatorski (Nicole.zatorski@icahn.mssm.edu).

Materials availability

- RNA-seq data for CPTAC cohorts are available at the GDC data portal: https://portal.gdc.cancer.gov/.
- All other previously publicly available data sets that the authors have used are described in the key resources table.

Data and code availability

- All original code has been deposited at GitHub: https://github.com/schlessinger-lab/Structural-Analysis-of-Genomic-and-Proteomic-Breast-Cancer-Signatures- and https://github.com/schlessinger-lab/structural_features. The code is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Structural features

SAGES calculates sequence-based features including length, frequency and amino acid composition from a variety of sources. These include sequence predictions such as: IDRs predicted with IUPRED2a[61]; protein binding regions in intrinsically disordered peptides predicted with ANCHOR[62]; transmembrane helical regions predicted with TMHMM 2.0[63]; and globular, helical, positive amino acid containing regions, negative amino acid containing regions, coiled-coil, sheet, loop, conserved, and non-conserved regions predicted with PredictProtein.[8] SAGES also assigns SCOPe[9] and UniProt[64] folds, families, superfamilies, and domains using HHpred,[65] with optimized parameters determined in previous work.[17] SAGES applies the same amino acid frequency calculation to features derived from structural models generated using AlphaFold2[16] (default parameters): 3_{10} helix, alpha helix, beta bridge, beta bulge, turns, and high curvature regions from the Dictionary of Protein Secondary Structure (DSSP)[10]; distances from the center of mass from biopython's pdb parser[66]; and aggregation propensity from Aggrescan3d.[67] Structures were determined locally using the AlphaFold2 algorithm. Features of each protein are compiled in a novel database, which SAGES accesses. This database can be updated with scripts found in the SAGES GitHub folder GitHub: https://github.com/schlessinger-lab/structural_features, titled update_alphafold.db.py and udate_databases.py.

Breast cancer proteomics and transcriptomics

Breast tumor tissue and normal control breast tissue was excised from 23 breast cancer patients during surgery. The human specimen collection procedure was approved by the Institutional Review Board of Icahn School of Medicine at Mount Sinai / Mt. Sinai Health Care System under the study ID HSM#-00135. Samples were then prepared and sequenced according to standard protocol.[68] The women enrolled had the following demographics: 13 were post-menopausal, 4 were Estrogen Receptor (ER) negative, 16 were Human Epidermal Growth Factor Receptor (HER2) negative, 5 were Progesterone Receptor (PR) negative, 16 had invasive ductal carcinoma, and 2 had internal mammary chain involvement. Five women received treatment prior to sample collection. One woman was administered Taxotere, Carboplatin, Herceptin, Perjeta (TCHP). Three women were administered Adriamycin, cyclophosphamide, Taxol (AC-T). One woman was administered Anastrozole. Global proteomics data from primary breast tumors and matched normal tissues were generated through mass-spectrometry (MS)[69] and pre-processed as previously described.[54] Data can be found at GSE227679. Data pertaining to mutated and overexpressed genes in human breast cancer was also sourced from the Catalogue of Somatic Mutations In Cancer (COSMIC).[23] This was divided into two sets, one containing only the most commonly mutated genes in breast cancer and the other containing only unmutated, overexpressed genes in breast cancer. For weighting purposes, the number of times a gene was mutated compared to the total number of times that gene was expressed served as the weight for the first set and the number of times the gene was overexpressed was used as the weight for the second set. Breast cancer proteomic data can be found at GDC: https://portal.gdc.cancer.gov/.

METHOD DETAILS

Structural feature generation

Transcriptomic and proteomic signatures can be represented by the structural and functional features of the proteins transcribed in the gene set (Figure 1). These structural features can be derived from protein sequence or from three-dimensional, resolved structure. SAGES calculates sequence based features including length, frequency and amino acid composition from sequence predictions from sources described in the key resources table. SAGES applies the same amino acid frequency calculation to features derived from structural models generated using AlphaFold2[16] with default parameters). These models were created locally for all sequences using the AlphaFold2 algorithm. SAGES determines the number of amino acids (length), the number of separate instances of a feature type (frequency), and the number of each type of amino acid (composition) for secondary features. For all secondary features SAGES uses a cutoff of 50% predicted probability according to the various feature prediction tools listed above when determining amino acid content of a region. Unlike the other features, aggregation propensity is denoted with a range of values that can be negative or positive rather than a percentage of predicted probability so all amino acids with positive values are included in the analysis. Additionally, SAGES tabulates the total number of contacts as well as the minimum, maximum, and average amino acid distance from the protein's center of mass. (Figure 1B) All features are listed in Table S1.

SAGES normalizes the features using the expression level of the input genes or proteins which correspond to each feature. This weighted average normalization ensures feature frequency adequately reflects prominence within the sample. SAGES further normalizes the output feature values using the number of genes or proteins input in the sample to ensure consistency between different sized samples. Family, fold, superfamily, and domain were not normalized due to the underabundance of these features within each input set.

QUANTIFICATION AND STATISTICAL ANALYSIS

Structural feature statistical analysis

The statistical tests included in the method compare the current sample to a background sample (Figure 1C). The preloaded background consists of the entire human proteome from all GTEx V8 samples.[22] This background was replaced with a control background matched to each experimental sample in the breast cancer feature analysis. The frequencies of categorical variables are compared to the background using the scipy[70] statistical package's Fisher's exact test to determine a p value. A two tailed, type two, t test was used to determine the t value for averages of lengths of secondary structures compared to background Equation 1. Because variance between sample and background was not assumed to be equal, a Welch's t-test was used.[71]

$$\textit{t value} = \frac{\textit{background mean} - \textit{sample}\ \text{mean}}{\sqrt{\dfrac{\text{standard deviation background}^2}{\textit{background size}} - \dfrac{\text{standard deviation sample}^2}{\textit{sample size}}}} \qquad \text{(Equation 1)}$$

The scipy statistical package's t.sf function with degrees of freedom equal to sample size minus two was used to determine the p value from the t value.

The p value cutoff was set at 0.05 and Bonferroni corrected[72] according to the number of found features in that set of feature types. For families, folds, superfamilies, and domains, this means the number of each of those types of features found. For this correction, number of amino acids and secondary structure elements were grouped together and always summed to 219 leading to an adjusted p value of 0.000228. The p value cutoff for the averages was corrected by 13 resulting in an adjusted value of 0.00385. The statsmodels stats.multitest package's false discovery rate correction,[73] was also calculated to provide an additional metric for determining significance for large samples. Base 2 log ratios of feature frequency in samples over background was also provided for categorical variables (Figure 1C). The code for structural features can be found on the following GitHub: https://github.com/schlessinger-lab/structural_features.

Prediction of normal tissues

Structural features were used to predict tissue type from RNA expression of normal tissues. Furthermore, the breast tissue model specifically was also used to highlight important structural characteristics of proteins that were highly predictive in this tissue type (Figure 1E). The top 250 most highly expressed genes for each GTEx sample were extracted from GTEx V8[22] along with their level of overexpression, which is defined as the log fold change greater than zero. As seen in previous work, this is a sufficiently sized number of genes for capturing underlying tissue specificity.[17] Structural features (section structural features are predictive of normal tissue type) for each sample were determined. There were 3,532 structural features. In order to reduce dataset redundancy, CD-HIT[25] analysis was conducted with a similarity 40%, the lowest possible allowable threshold. CD-Hit identified 3647 unique clusters of proteins corresponding to the genes in the tissues out of the initial 5,376 genes. Non-representative proteins were excluded from the training features. Sample features were min-max normalized (Equation 2) in preparation for the tissue type prediction where x_i equals the value of a feature for a sample and x equals the set of values for all samples corresponding to that feature.

$$\textit{normalized value} = \frac{x_i - \min(x)}{\max(x) - \min(x)} \qquad \text{(Equation 2)}$$

For each of the 30 tissue types, the samples were split into 90/10 training and test sets with equal number of positive and negative labels. The scikit learn version 0.24.1 random forest classifier was used to predict whether a sample could be classified as that type of tissue. The random forest classifier had the following parameter values.

Argument Name	Argument Input Value
criterion to measure split quality	gini impurity
maximum depth	none
minimum sample split	2
minimum samples leaves	1
minimum weight fraction leaf	0
maximum features	square root of the number of features
maximum leaf nodes	none
minimum impurity decrease	0
samples	bootstrap
generalization score	no out of bag samples used for estimation
number of parallel jobs	none
verbosity	none
warm start	false
class weights	none
pruning complexity parameter	0
maximum samples	none

This classifier fits 100 decision trees on data subsets and averages them into a meta predictor to improve performance. Following 10-fold cross validation with random seed 0-9, performance was measured using: the area under the receiver operating characteristic curve (AUROC), accuracy Equation 3 where y_i is the value of the ith sample and y_i' is the corresponding predicted value, precision Equation 4 where TP is true positives and FP is false positives, recall Equation 5 where FN is false negatives, and F-score Equation 6. Standard deviations for all metrics were computed using python.statistics stdev.

$$accuracy(y, y') = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(y_i' = y_i) \quad \text{(Equation 3)}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad \text{(Equation 4)}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad \text{(Equation 5)}$$

$$F1 = 2\frac{\mathrm{P\ R}}{\mathrm{P+R}} \quad \text{(Equation 6)}$$

Performance of the classifiers trained on GTEx was also validated on an external dataset - ARCHS4,[26] a human and mouse tissue dataset. Predictors for each tissue type were evaluated in a 10-fold cross validation fashion using AUROC and Equations 3, 4, 5, and 6 on the data type originally used to train the predictor.

Additionally, the most predictive features for normal breast tissue were determined using sklearn recursive feature elimination (RFE) with random forest as the estimator.[27] Weights were assigned to all features and then the less important features were eliminated recursively until a single feature remained. To allow for full interrogation of the expressed proteins, no proteins were excluded on the basis of structural similarity. This generated a ranking of features based on contribution to the model. RFE was conducted 10 times with random seeds 0-9 and rankings were summed. The 25 features with the lowest score, and therefore best rank are reported in Table 2.

Breast cancer analysis

The 250 most highly expressed genes were extracted from each genomic and proteomic sample and SAGES was applied (Figure 1). Each patient breast tumor sample had a matching normal breast tissue control from the same patient, which was used to generate the background for the statistical analysis described as part of the SAGES method (Figure 1C). The SAGES of the top 250 genes from all GTEx normal breast tissue samples served as the background controls for the COSMIC samples (Figure 1C). For each sample, if a feature was statistically significant (p value less than the Bonferroni corrected significance level of 0.0011 determined by dividing 0.05 by the number of non-amino acid specific features) compared to the background, the log ratio was included in the averages visualized in Figure 3. This statistical significance based selection of features corresponds to Figure 1D in the SAGES methodology. The standard deviation was determined for all features that came from more than one sample. Due to the large number of amino acid related features, these were excluded from the visualization. Folds, superfamilies, and domains were present only in the breast cancer samples and not in the background were excluded from this analysis.

Gene Ontology (GO) analysis was also performed using the top 250 most overexpressed genes for each dataset with the panther classification system and the Bonferroni adjusted significance level.[74] Results from the Benjamini-Hochberg[48] procedure were also examined. The following sample-background pairs were compared using this approach: breast cancer transcriptomic samples-normal breast transcriptomic samples, breast cancer proteomic samples-normal breast proteomic samples, breast cancer proteomic samples- breast cancer transcriptomic samples, normal breast tissue proteomic samples-normal breast tissue transcriptomic samples, COSMIC breast cancer samples with and without mutations-GTEx normal breast tissue.

All data sets investigated were compared using Jaccard similarity to provide context for the results observed in SAGES and GO analysis. These data sets were the same as those used in GO analysis, described above. Data sets were compared pairwise according to Equation 7. A larger Jaccard coefficient indicated larger similarity.

$$Jaccard = \frac{\text{length}(\text{set }1 \cap \text{set }2)}{length\ (set1 \cup set2)} \qquad \text{(Equation 7)}$$

Breast cancer drug perturbation

Perturbation signatures of cell lines treated with drugs from the Connectivity Map[1] were compared to various breast cancer backgrounds to investigate how drug induced gene expression relates to disease signature. Unweighted SAGES for all overexpressed genes for each drug were calculated and compared to unweighted SAGES of all proteins from the breast cancer proteomics samples with log ratio expression greater than one and to unweighted SAGES of all genes from the COSMIC mutated breast cancer gene dataset. Unweighted SAGES were used to ensure that all overexpressed proteins contributed equally to the analysis. For each sample, the number of significantly different features from background were counted. The significance level of 0.05 was selected. Multiple hypothesis testing was not employed because the aim was ultimately to assess similarity to background rather than difference from background and using this technique would increase the type II error. The samples were divided into breast cancer treatment drugs (doxorubicin, fulvestrant, letrozole, megestrol, methotrexate, paclitaxel, raloxifene, tamoxifen, and vinblastine) according to a list published by the National Cancer Institute, and all other drugs in the Connectivity Map database. The average number of statistically significantly different features for each group were calculated and a two-sided, type 2, student t test was used to determine the p value.

Gene expression of the perturbation samples and the expression signatures used to calculate the SAGES of the two backgrounds were also directly compared. The Jaccard coefficient (Equation 7)was used to determine signature similarity. Samples were divided into breast cancer and other drugs and compared with a two-sided, type 2, student t test. Additionally, the average Jaccard coefficient for both groups was determined.