Building trans-omics evidence: using imaging and 'omics' to characterize cancer profiles

Arunima Srivastava
Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210
Email: srivatava.1@osu.edu

Chaitanya Kulkarni
Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210
Email: kulkarni.132@osu.edu

Parag Mallick*
Canary Center for Cancer Early Detection, Stanford University, 3155 Porter Dr., Palo Alto, CA, 94305
Email: paragm@stanford.edu

Kun Huang*
Department of Medicine, Indiana University School of Medicine, 340 W 10th St #6200, Indianapolis, IN 46202
Email: kunhuang@iu.edu

Raghu Machiraju*
Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210
Email: Machiraju.1@osu.edu

** Corresponding authors*

Utilization of single modality data to build predictive models in cancer results in a rather narrow view of most patient profiles. Some clinical facets relate strongly to histology image features, e.g. tumor stages, whereas others are associated with genomic and proteomic variations (e.g. cancer subtypes and disease aggression biomarkers). We hypothesize that there are coherent "trans-omics" features that characterize varied clinical cohorts across multiple sources of data leading to more descriptive and robust disease characterization. In this work, for 105 breast cancer patients from the TCGA (The Cancer Genome Atlas), we consider four clinical attributes (AJCC Stage, Tumor Stage, ER-Status and PAM50 mRNA Subtypes), and build predictive models using three different modalities of data (histopathological images, transcriptomics and proteomics). Following which, we identify critical multi-level features that drive successful classification of patients for the various different cohorts. To build predictors for each data type, we employ widely used "best practice" techniques including CNN-based (convolutional neural network) classifiers for histopathological images and regression models for proteogenomic data. While, as expected, histology images outperformed molecular features while predicting cancer stages, and transcriptomics held superior discriminatory power for ER-Status and PAM50 subtypes, there exist a few cases where all data modalities exhibited comparable performance. Further, we also identified sets of key genes and proteins whose expression and abundance correlate across each clinical cohort including (i) tumor severity and progression (incl. GABARAP), (ii) ER-status (incl.

ESR1) and (iii) disease subtypes (incl. FOXC1). Thus, we quantitatively assess the efficacy of different data types to predict critical breast cancer patient attributes and improve disease characterization.

1. Introduction

Recent advances in whole slide imaging digitization and compilation in the form of the TCGA compendium[1], alongside matching high throughput profiling data has opened up many avenues for modeling different facets of an experiment. Histopathological images have been very successful in predicting clinical outcomes in the context of various TCGA cancers[2]. Similarly, transcriptomics and proteomics profiling has showcased distinct discriminatory power when modeling cancer subtypes for the purpose of targeted drug therapies and biomarker excavation[3]. The purpose of this work is to comprehensively compare the characterizing abilities of these three modalities of data across varying types of clinical cohorts, when traditionally, each are analyzed in the context of specific attributes only (**Figure 1**). The overarching goal is to (a) qualitatively compare the predictive power of each type of data modality while modeling different patient attributes and (b) find coherent biological signatures and features (e.g. driver genes and subtype-specific protein biomarkers) that provide a framework for evidence based depiction of each attribute cohort.

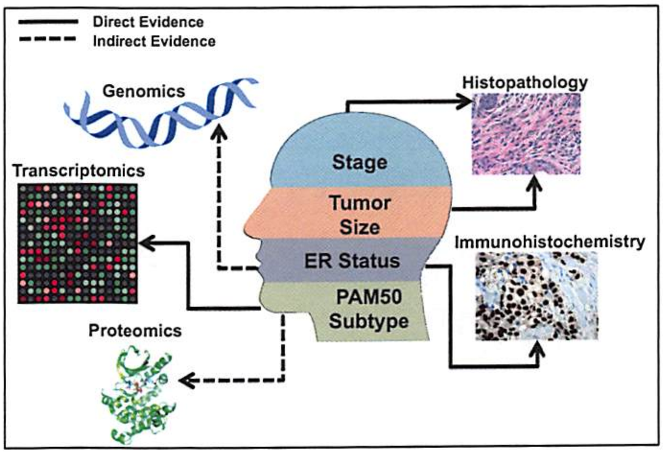


Figure 1. Various patient attributes and the data modalities that present validation or evidence for them.

To achieve the above goals, we utilize 105 patients from the TCGA-BRCA (Breast Cancer) dataset, which contain a clinically diverse set of patients and multiple levels of data. Namely, histopathological (H-hematoxylin and E-eosin stained) tissue images of the tumor region, transcriptomics (RNA-Seq) data and proteomics (Isobaric tag for relative and absolute quantitation –iTRAQ) data are available for all the patients categorized in this study. The data modalities were used to model prediction for the four (4) clinical attributes, namely, AJCC (American Joint Committee on Cancer) staging, tumor staging, ER-status and lastly PAM50 panel breast cancer subtypes.

Data modalities and biological models

Histology Images - There has been a recent upsurge in publications describing the utilization of H/E whole slide images (WSIs) to predict clinical outcomes. Many of these present the use of a deep learning approach on tiles of histology images, as opposed to semantic image features, to build classification models[2] (**supplementary material**). There are many commonalities between the inferences that can be drawn from morphological features identified by CNNs (atypical shape of cells, disintegration of tissue architecture and visible stromal invasiveness) and the hallmarks of cancer (cells resisting apoptosis, metastatic tendencies and limitless replicative potential). We chose GoogLeNet[4] inspired CNNs with histology data to classify the dataset into clinical cohorts as listed above.

Transcriptomics - From the four (4) patient stratifications that we aimed to assess, the molecular subtypes (ER-Status and PAM50 subtypes) are natively associated with transcript level variations across patients. Histology images remain the preferred and proven method for largely predicting clinical status, while a few analyses using transcriptomics, modeling staging and similar attributes have been performed with some success[5]. Transcript level expression has been primarily utilized to subtype patients, extract biomarkers and understand signaling and regulation using data from microarrays and RNA-Seq[6] using traditional methods of analysis that are well understood and widely implemented[7].

Proteomics - High throughput proteomics experiments have recently gained popularity as they enable the study of regulatory mechanisms in cancer. Combining proteomics and transcriptomics, models disease more effectively than using these datasets independently and thus proteogenomics analysis has been utilized for gauging better subtypes of disease, associating genomic variations with signaling and isolating disease driver genes and proteins[3,8].

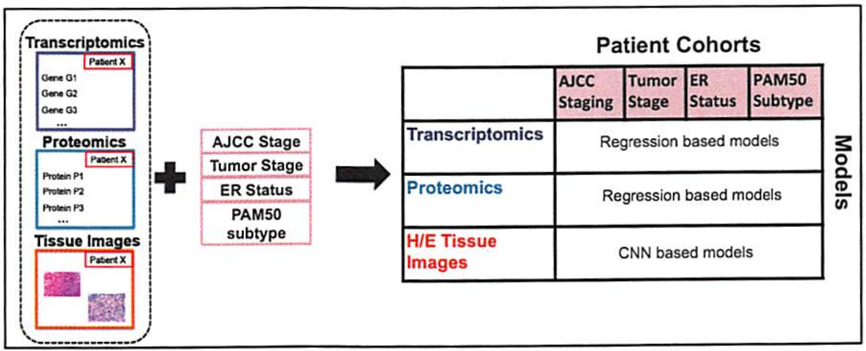


Figure 2. Showcases the construction of the multiple models across different data modalities. Comparing the performance of these models quantifies each data level's utility for identification of characteristic features for each of the four clinical attributes.

Our aim was to build models with "best practice" methods for each data modality, and for the same set of patients, train, test and predict the different clinical attributes. The goal was to compare utility of data types, using traditional, widely used methods for each data type. For the best performing models for each data modality, we compare the corresponding results for all four clinical cohorts (**Figure 2**). As expected, histology based models are ideal for predicting clinical outcomes and genomic data outperforms all other data types when predicting molecular subtypes. However, we also observe comparable performance metrics amongst non-traditional data models in some cases, which allow us to reinforce the characterization of cohorts across the scales. For instance, our results show that there exist salient genes and proteins that are able to distinguish between AJCC stages and tumor stages relatively successfully. These included genes that are verifiably associated with disease severity and tumor progression – e.g. GABARAP, CKS2 and CDH1. A more comprehensive list of molecular markers is included in subsequent sections. While image-based classifiers outperform them, they can still help build trans-omic evidence and reinforce disease profiles. In summary, this work explores the likelihood of finding evidence in data types that include trans-omic indicators. Additionally it forms a critical foundational layer for future integrative imaging genomics models by highlighting utility of data types and extracting meaningful features.

2. Methods and Materials

Below we describe, a workflow for (a) building data driven predictive models (Section 2.2), (b) evaluating performance of all the models for comparison (Section 2.3) and (c) extracting key driving features from successful models (Section 2.4). Scripts used to build and evaluate models as well as find key features are included in the **supplementary material**.

2.1. *TCGA Breast Cancer multi-level dataset*

The TCGA breast cancer study[9] is a well-characterized and thoroughly comprehensive experimental study of breast invasive carcinoma[5,10]. A total of 105 patients, for whom all three (3) types of data were available were utilized in this work. The three data modalities include RNA-Seq mRNA expression, selected parts of H/E stained whole slide tissue images and proteomic abundance from mass spectrometry (LC-MS/MS –Liquid Chromatography-Mass Spectrometry) analysis. Further, each of these 105 patients have associated clinical profiles that detail their AJCC Stage (Stage I, II, III and IV), Tumor Stage (T1, T2, T3 and T4), ER status (+/-) and finally PAM50 mRNA subtypes of the cancer (Luminal A, Luminal B, HER2 enriched and Basal like).

2.2. *Different data modalities and subsequent model construction*

2.2.1. *Histology images and CNN based classification models*

As mentioned previously, we utilized the GoogLeNet (2014) Convolutional Neural Network to construct a classifier using the tiles of histology images that were acquired from the Genomic Data Commons (GDC)[11]. As is typical with neural networks, multiple parameters (structure of input, number of network layers, pre-training data, etc.) are instrumental to the eventual performance. To this end, we constructed classifiers using five (5) different versions of the standard GoogLeNet (with and without pre-training data, larger input tile sizes, spatially invariant input tiles and only epithelial regions as training data) and evaluated the performance for each across all clinical cohorts. See **supplementary material** for more details including key transformative features of the different CNN models.

2.2.2. *Transcriptomics data and modeling*

Transcript expression RNA-Seq data (percentile-normalized version) was accessed from the UCSC Xena (http://xena.ucsc.edu/)[12]

Project. Superfluous gene names or transcript measurements (tagged as NA) were removed and Z-transform normalization of the resulting data was performed. The final data matrix contained the normalized transcript measures for 20501 unique genes across 105 patients, and the corresponding four (4) dimensional clinical attribute vectors. To build predictive models for multi-class labels, we employed multinomial log-linear regression (using function "*mutinom*" from R-package "*nnet*"[13]) on different subsets of the transcriptomics data. To build these regression models from the transcriptomics dataset, we find subsets of the genome wide dataset that potentially represented a variety of cohorts (histology and molecular profiling based). **Table 1** presents the gene lists[11,14–17] which helped extract those data subsets and a summation of what patient profiles they help characterize (**Figure 3**).

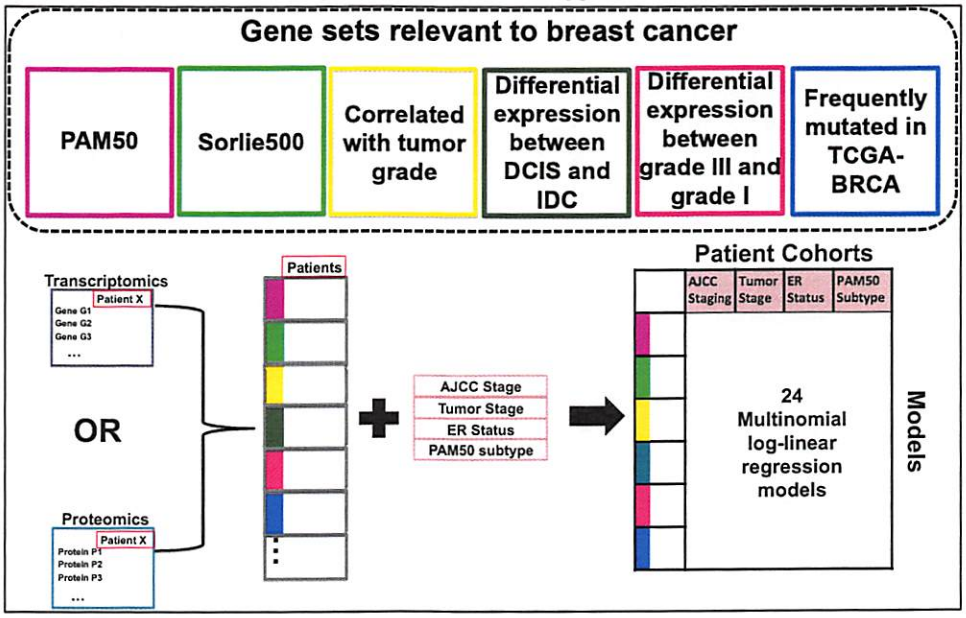

Gene sets relevant to breast cancer
PAM50
Sorlie500
Correlated with tumor grade
Differential expression between DCIS and IDC
Differential expression between grade III and grade I
Frequently mutated in TCGA-BRCA
Transcriptomics
Patient X
Gene G1
Gene G2
Gene G3
...
OR
Proteomics
Patient X
Protein P1
Protein P2
Protein P3
...
Patients
AJCC Stage
Tumor Stage
ER Status
PAM50 subtype
Patient Cohorts
AJCC Staging
Tumor Stage
ER Status
PAM50 Subtype
24 Multinomial log-linear regression models
Models

Figure 3. Workflow detailing the construction of regression models from transcriptomics and proteomics data. 6 different models are constructed for each clinical cohort (AJCC Staging, Tumor Staging, ER Status and PAM50 Subtypes). These 6 different models are constructed by extracting subsets of the proteogenomic data for gene sets relevant to various different attributes of breast cancer. Using these data subsets, and clinical cohort labels, we construct multinomial log-linear regression predictive models.

Table 1. Gene lists proven to be relevant to different breast cancer patient attributes

Gene Lists	Patient Attribute Characterization
PAM50	50 genes, defining expression based intrinsic subtypes of breast cancer
Sorlie500	500 intrinsic unique genes, defining refined subtypes of breast cancer
Top 200 genes related to tumor grade	200 genes, highly and significantly correlated with tumor grade
Most frequently mutated in TCGA-BRCA	Top 20 most frequently mutated genes in the TCGA-BRCA project patients, as reported by data analysis performed by the GDC portal
Differentially expressed between DCIS-S and IDC-S	305 genes classifying tumor invasion, presenting significant differential expression between stroma of ductal carcinoma *in situ* and invasive ductal carcinoma
Differentially expressed between grade III and grade I	620 genes, found differentially expressed between stroma of tumor grade III and tumor grade I patients, classifying the extent of abnormality of the tumor

2.2.3. *Proteomics data and modeling*

Normalized proteomics abundance (iTRAQ - Isobaric tag for relative and absolute quantitation) for the 105 patients considered for this work were extracted from supplementary data provided in the work of Mertins Et al.[3] (Supplementary Table 03 – Global Proteome G1). The details of data normalization and pre-processing are described in the supplementary information of the publication cited above. Redundant profiles or samples were removed from the normalized dataset mimicking the technique utilized previously for transcriptomics data. The final data matrix of normalized protein abundances contained 6386 unique proteins across the relevant 105 patients. Similar to the analysis technique engaged for the transcriptomics dataset (regression modeling with biologically driven subsets of data), we built models (**Figure 3**) with subsets of the proteomic dataset in the context of the four (4) selected clinical attributes.

2.3. *Metrics of performance for each model*

Standard performance metrics of precision, recall and F-score for all versions of the models constructed from each data type are used. As described above, multiple variations of models from each data modality were constructed, and the ones that performed best, within a single data level, were chosen for performance comparison across data levels. For proteogenomics datasets, we calculated the performance measures using cross-validation and partitioning 70% of the data for training, using function "*prediction*" from the R-package "*ROCR*"[18]. For the CNN based models, suitable programmatic metrics[19] were utilized for the same. Specifics of model evaluation, including the details regarding division of data to training and testing sets, are further detailed in the **supplementary material**.

2.4. *Isolating the key biological signature features characterizing each cohort*

We aim to identify key features from all data modalities, which project the highest discriminatory power for each clinical attribute.

Transcriptomics and Proteomics - We select the model that predominantly outperforms all other proteogenomic models and extract genes and proteins that contain high predictive power for each of the patient attributes. We utilize the RFE (Recursive Feature Elimination)[20,21] workflow (see **supplementary material**) to find subsets of features for each model that are critical for guiding the classification (using function "*rfe*" provided by the R-package "*caret*"[20]). Due to selection bias, we execute the algorithm multiple times, and extract the features that are consistently deemed important for the model.

Histology Images –While CNNs produce weighted feature maps used to drive classification, it is challenging to map them to verifiable histology image features. For the purposes of this work, we assess genomic features only, but propose to expand this study in the future to utilize CNN saliency maps in an effort to extract relevant feature images from CNNs.

3. Results and Discussion

In this section we report performance metrics for models generated from all data types, and the key genes and proteins that drive successful classification. All other relevant information (gene lists, modeling parameters, R data frames) is detailed in the **supplementary material**.

3.1. *Classification of patient attributes using histopathological tissue images using CNNs*

We compared the performance metrics of precision, recall and F-score across all the different versions of the image classification CNNs. We observed that the best predictive model for classifying patients to AJCC stages was borne from the "*Inception (v3 2015) with pre-training version of the CNN*" (0.52 F-score), whereas for tumor stages the "*Inception (v3 2015) with pre-training and boosted data*" (0.50 F-score) outperformed the other models. For clinical cohorts (ER-Status and PAM50 subtypes) traditionally related to genomic data, the best predictive models compared to all other CNN based models were again the "*Inception (v3 2015) with pre-training and boosted data"* (0.64 F-score) and "*Inception (v3 2015) with no pre-training"* (0.35 F-score) respectively. To summarize, it is not surprising that tissue images predict the AJCC and tumor staging as well as ER-status drastically better than PAM50 subtypes. Additionally, the varying parameters and boosting data techniques had little to no discernible effect to the quality of the models produced.

3.2. *Modeling and classification using proteogenomics*

We now describe the results and the attained efficacy of the multinomial log-linear regression models built with transcriptomics and proteomics measures. Examples of the genes and proteins identified as critical to the predictive model (using RFE analysis) are listed in **Figure**

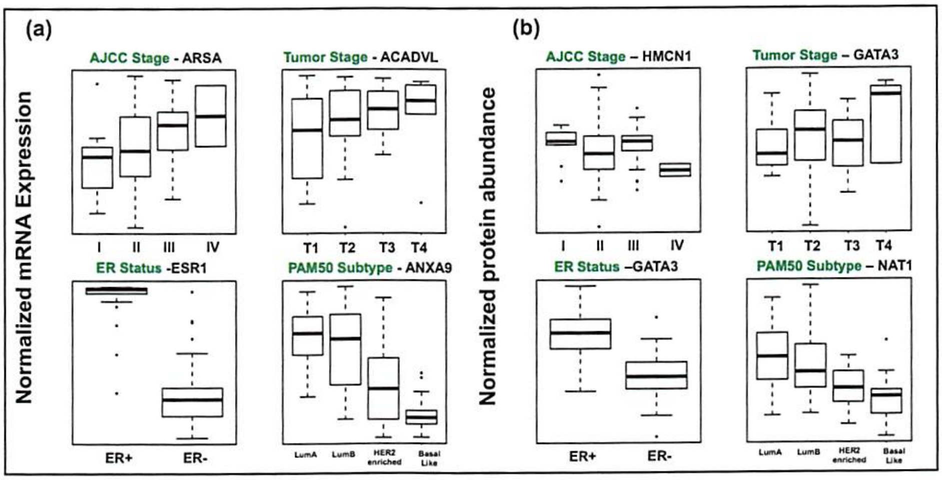


Figure 4. (a) Boxplots presenting gene expression (percentile normalized) of example top ranking critical features (genes) derived from "best" models for each clinical cohort using transcriptomics data. (b) Boxplots presenting protein abundance (mixture model-based normalization) of example top ranking critical features (proteins) derived from "best" models for each clinical cohort using proteomics data.

4, and examples with corresponding literature evidence are listed below. Performance metrics for all the models and literature evidence for key genes and proteins as identified by RFE analysis is listed in **supplementary material**.

3.2.1. *Transcriptomics model results*

AJCC and Tumor Stage-The performance metric indicates that the transcriptomics model generated with *Top 200 genes related to tumor grade* outperform all other models when predicting both AJCC Stage and Tumor Stage (0.43 and 0.4 F-score respectively). The RFE approach isolates 11 genes that are critical to the model predicting AJCC Stage and 10 genes for the tumor stage model. All 21 of the extracted key genes from the above two models are part of a verified breast cancer grading signature[22]. A large number of these genes are shown to be significantly associated with clinical outcome, a result that can be easily extended to stage and tumor size association (analyzed using the PRECOG database[23] (https://precog.stanford.edu), which associates gene expression to clinical outcome, tabulated in **supplementary materials**). These genes include CKS2, overexpression of which is known to be involved in tumor development by blocking cell cycle S-phase signaling which causes cells to abnormally proliferate in stress conditions[24] and ARPC1A, associated with poor prognosis in many cancers[25].

ER-Status-The classification between patients with differing ER-Status is performed perfectly by the *PAM50* set of intrinsic genes (0.925 F-score). Gene expression of ESR1 and other co-expressed genes, included within the *PAM50* set are widely known as having high discrimination for ER-positive status[26]. The RFE guided list of driver genes for this model lists 7 key genes including ESR1. Investigating these critical genes using Gene Set Enrichment Analysis (GSEA)[27], we observe that 6 out of the 7 identified genes are known to be up regulated specifically in ESR1 positive breast cancer tumors (false discovery rate of 2.2 e^{-11})[28]. This gene set includes driver genes such as FOXA1, which is involved with ESR1 for regulation[29]**,** NAT1, known to be commonly overexpressed in ER-alpha (+) tumors[30], and MDM2, which regulates ER-alpha and estrogen responsiveness in breast cancer cells[31].

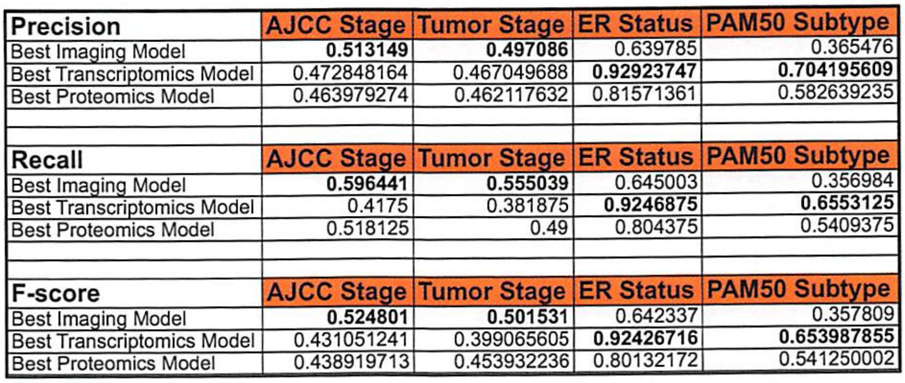
PAM50 mRNA Subtype-While one would expect the data subset from *PAM50* gene set to perform ideally when predicting the PAM50 subtype status, *Sorlie500* transcriptomics subset in fact presents a superior classifier when performing classification to PAM50 mRNA subtypes (0.65 F-score versus 0.59 for the *PAM50* transcriptomics subset). A total of 13 genes from the *PAM50* gene set are included in the *Sorlie500* gene set, along with other intrinsic subtyping genes. Acquisition, normalization, pre-processing and composition of data all are known to fluctuate the results of PAM50 mRNA based subtyping[32]. We hypothesize that while highly relevant (known biomarkers) genes like ESR1, ERBB2, FOXA1, FOXC1[33] etc. (included in both *PAM50* and *Sorlie500* gene sets, and identified as highly important for this model) drive the subtype classification, additional intrinsic genes encompassed in the *Sorlie500* set further help stratification. The hypothesis is confirmed by the RFE analysis, which selected 23 highly important genes to the model, 7 of which also belonged to the *PAM50* gene set and included known discriminants such as ESR1 and ERBB2[34].

3.2.2. *Proteomics model results*

AJCC, Tumor Stage and ER-Status-All three of these cohorts are best predicted by proteomics data subset of the *Most frequently mutated genes*. The proteomics data subsets presented F-score measures of 0.43, 0.45 and 0.80 for the AJCC stage, tumor stage and ER-status classification respectively, compared to 0.43, 0.39 and 0.92 F-score observed for the best models from the transcriptomics dataset. The RFE analysis further isolates a list of two (2) (e.g. DST, found to be breast cancer tumor progression suppressor[35]), three (3) (e.g. disease severity biomarker CDH1[36]) and four (4) (e.g. GATA3, strongly associated with ER-Status[37]) proteins respectively that are the main drivers of these models.

PAM50 mRNA Subtype - The subset of *PAM50* proteins and their corresponding normalized abundances outperform all other models when attempting to classify the PAM50 mRNA subtype of the patients using the proteomics data. All performance measures showcase that this model is still lacking in predictive power as compared to the best *Sorlie500* model derived from the transcriptomics data (0.65 F-score for the transcriptomics *Sorlie500* model versus 0.54 for proteomics *PAM50* model). This may be caused, potentially due to a disconnect between various transcripts and corresponding proteins due to post-transcriptional regulation. The RFE analysis outlines 10 proteins that drive the classification for this model and they include well known breast cancer relevant proteins such as ESR1, ERBB2 and RRM2 (associated with basal proliferative tumors[38]). Four (4) of the key genes from the corresponding list derived from the transcriptomics model (ESR1, FOXA1, ERBB2 and NAT1) are included in the key proteins list as well.

3.3. *Comparison of all data modalities*



Precision	**AJCC Stage**	**Tumor Stage**	**ER Status**	**PAM50 Subtype**
Best Imaging Model	**0.513149**	**0.497086**	0.639785	0.365476
Best Transcriptomics Model	0.472848164	0.467049688	**0.92923747**	**0.704195609**
Best Proteomics Model	0.463979274	0.462117632	0.81571361	0.582639235
Recall	**AJCC Stage**	**Tumor Stage**	**ER Status**	**PAM50 Subtype**
Best Imaging Model	**0.596441**	**0.555039**	0.645003	0.356984
Best Transcriptomics Model	0.4175	0.381875	**0.9246875**	**0.6553125**
Best Proteomics Model	0.518125	0.49	0.804375	0.5409375
F-score	**AJCC Stage**	**Tumor Stage**	**ER Status**	**PAM50 Subtype**
Best Imaging Model	**0.524801**	**0.501531**	0.642337	0.357809
Best Transcriptomics Model	0.431051241	0.399065605	**0.92426716**	**0.653987855**
Best Proteomics Model	0.438919713	0.453932236	0.80132172	0.541250002

Figure 5. Precision, recall and F-score across all clinical cohorts for the best performing models for each data modality respectively. Standard deviations for measures calculated using cross-validation are included in supplementary materials.

As previously mentioned, the best performing models, for each data modality, were ultimately compared across all three data levels for each cohort (**Figure 5**). Histology image based models outdid the transcriptomics and proteomics models while predicting the AJCC and tumor stage. Both ER-Status and PAM50 subtypes were ideally characterized by transcriptomics data, with proteomics data models performing second best. While this was in accordance with our expectations, there were a few points of interest within these comparative results. For instance, while imaging models were the most precise in classifying patient staging, it is important to note that both transcriptomics and proteomics models were not drastically less precise (0.51 vs ~0.47). This indicates that there exist features within genomic variations, which have the discriminatory power to effectively characterize histology-based staging. These comparisons not only quantified the utility of each data type in modeling various clinical subtypes, they also identified previously unexplored associations between data types and patient profiles.

4. Conclusions and Future Work

In this study we showcased the different data modalities and how they are utilized for modeling different facets of cancer. We employed "best practice" modeling techniques for three different data modalities to predict four (4) varied attributes of breast cancer patients in the TCGA compendium. We quantified the predictive power of data modalities for different aspects of patient profiles. Finally, key genomic features critical to all different clinical attributes were identified and validated using existing literature. We wish to expand this work by exploring the various pathways (e.g. FOXA1/ESR1/GATA3 interacting pathway) that include the identified key genes and proteins, in the context of various different cohorts. Additionally, we wish to perform analysis to explain what causes the differences between predictive powers across data modalities (e.g. transcriptomics and proteomics models). Further, we wish to utilize more sophisticated methods, including more robust regression to account for the structure of data, for each data modality to perform better stratification of patient subtypes and construct robust patient similarity frameworks using these trans-omic evidences.

Supplementary Materials - https://github.com/arunima2/PSB_2018

References

1. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkol.* 2015;1A:A68-A77. doi:10.5114/wo.2014.47136.
2. Araújo T, Aresta G, Castro E, et al. Classification of breast cancer histology images using Convolutional Neural Networks. Sapino A, ed. *PLoS One.* 2017;12(6):e0177544. doi:10.1371/journal.pone.0177544.
3. Mertins P, Mani DR, Ruggles K V., et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature.* 2016;534(7605):55-62. doi:10.1038/nature18003\rhttp://www.nature.com/nature/journal/v534/n7605/abs/nature18003.html#supplementary-information.
4. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* Vol 07-12-June. ; 2015:1-9. doi:10.1109/CVPR.2015.7298594.
5. Yao F, Zhang C, Du W, Liu C, Xu Y. Identification of gene-expression signatures and protein markers for breast cancer grading and staging. *PLoS One.* 2015;10(9). doi:10.1371/journal.pone.0138213.
6. Bertucci F, Finetti P, Rougemont J, et al. Gene expression profiling identifies molecular

subtypes of inflammatory breast cancer. *Cancer Res.* 2005;65(6):2170-2178. doi:10.1158/0008-5472.CAN-04-4115.

7. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 2016;17:13. doi:10.1186/s13059-016-0881-8.
8. Beretov J, Wasinger VC, Millar EKA, Schwartz P, Graham PH, Li Y. Proteomic Analysis of Urine to Identify Breast Cancer Biomarker Candidates Using a Label-Free LC-MS/MS Approach. Aboussekhra A, ed. *PLoS One.* 2015;10(11):e0141876. doi:10.1371/journal.pone.0141876.
9. Koboldt DC, Fulton RS, McLellan MD, et al. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012;490(7418):61-70. doi:10.1038/nature11412.
10. Kim D, Li R, Dudek SM, Ritchie MD. Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer. *J Biomed Inform.* 2015;56:220-228. doi:10.1016/j.jbi.2015.05.019.
11. Grossman RL, Heath AP, Ferretti V, et al. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med.* 2016;375:1109-1112. doi:10.1056/NEJMp1002530.
12. Goldman M, Craft B, Swatloski T, et al. The UCSC cancer genomics browser: Update 2015. *Nucleic Acids Res.* 2015;43(D1):D812-D817. doi:10.1093/nar/gku1073.
13. Venables WN, Ripley BD. Modern Applied Statistics with S. *Issues of Accuracy and Scale.* 2002;(March):868. doi:10.1198/tech.2003.s33.
14. Bernard PS, Parker JS, Mullins M, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.* 2009;27(8):1160-1167. doi:10.1200/JCO.2008.18.1370.
15. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A.* 2003;100(14):8418-8423. doi:10.1073/pnas.0932692100.
16. Ma X-J, Dahiya S, Richardson E, Erlander M, Sgroi DC. Gene expression profiling of the tumor microenvironment during breast cancer progression. *Breast Cancer Res.* 2009;11(1):R7. doi:10.1186/bcr2222.
17. Ma X-J, Salunga R, Tuggle JT, et al. Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci U S A.* 2003;100(10):5974-5979. doi:10.1073/pnas.0931261100.
18. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics.* 2005;21(20):7881. http://rocr.bioinf.mpi-sb.mpg.de.
19. Bowles M. Machine learning in Python. *Igarss 2014.* 2014;(1):1-5. doi:10.1007/s13398-014-0173-7.2.
20. from Jed Wing MKC, Weston S, Williams A, et al. caret: Classification and Regression Training. 2015. http://cran.r-project.org/package=caret.
21. Zeng X, Chen Y-W, Tao C, Alphen D Van. Feature Selection Using Recursive Feature Elimination for Handwritten Digit Recognition. *2009 Fifth Int Conf Intell Inf Hiding Multimed Signal Process.* 2009:1205-1208. doi:10.1109/IIH-MSP.2009.145.
22. Ma XJ, Sgroi DC, Erlander MG. Grading of breast cancer. 2017. https://www.google.com/patents/US20170073767.
23. Gentles AJ, Newman AM, Liu CL, et al. The prognostic landscape of genes and infiltrating

immune cells across human cancers. *Nat Med.* 2015;21(8):938-945. doi:10.1038/nm.3909.

24. Liberal V, Martinsson-Ahlzén H-S, Liberal J, et al. Cyclin-dependent kinase subunit (Cks) 1 or Cks2 overexpression overrides the DNA damage response barrier triggered by activated oncoproteins. *Proc Natl Acad Sci U S A.* 2012;109(8):2754-2759. doi:10.1073/pnas.1102434108.
25. Lomakina ME, Lallemand F, Vacher S, et al. Arpin downregulation in breast cancer is associated with poor prognosis. *Br J Cancer.* 2016;114(5):545-553. doi:10.1038/bjc.2016.18.
26. Iwamoto T, Booser D, Valero V, et al. Estrogen Receptor (ER) mRNA and ER-related gene expression in breast cancers that are 1% to 10% ER-positive by immunohistochemistry. *J Clin Oncol.* 2012;30(7):729-734. doi:10.1200/JCO.2011.36.2574.
27. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci.* 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102.
28. Doane a S, Danso M, Lal P, et al. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. *Oncogene.* 2006;25(28):3994-4008. doi:10.1038/sj.onc.1209415.
29. Chaudhary S, Krishna B, Mishra S. A novel FOXA1/ESR1 interacting pathway: A study of Oncomine^{TM} breast cancer microarrays. *Oncol Lett.* June 2017. doi:10.3892/ol.2017.6329.
30. Abba MC, Hu Y, Sun H, et al. Gene expression signature of estrogen receptor alpha status in breast cancer. *BMC Genomics.* 2005;6(1):37. doi:10.1186/1471-2164-6-37.
31. Kim K, Burghardt R, Barhoumi R, Lee S ook, Liu X, Safe S. MDM2 regulates estrogen receptor α and estrogen responsiveness in breast cancer cells. *J Mol Endocrinol.* 2011;46(2):67-79. doi:10.1677/JME-10-0110.
32. Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. *J Natl Cancer Inst.* 2015;107(1):357. doi:10.1093/jnci/dju357.
33. Han B, Bhowmick N, Qu Y, Chung S, Giuliano AE, Cui X. FOXC1: an emerging marker and therapeutic target for cancer. *Oncogene.* 2017;(February):1-7. doi:10.1038/onc.2017.48.
34. Kao J, Salari K, Bocanegra M, et al. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS One.* 2009;4(7). doi:10.1371/journal.pone.0006146.
35. Lee S, Stewart S, Nagtegaal I, et al. Differentially expressed genes regulating the progression of ductal carcinoma in situ to invasive breast cancer. *Cancer Res.* 2012;72(17):4574-4586. doi:10.1158/0008-5472.CAN-12-0636.
36. Memni H, Macherki Y, Klayech Z, et al. E-cadherin genetic variants predict survival outcome in breast cancer patients. *J Transl Med.* 2016;14(1). doi:10.1186/s12967-016-1077-4.
37. Voduc D, Cheang M, Nielsen T. GATA-3 expression in breast cancer has a strong association with estrogen receptor but lacks independent prognostic value. *Cancer Epidemiol Biomarkers Prev.* 2008;17(February):365-373. doi:10.1158/1055-9965.EPI-06-1090.
38. Bertucci F, Finetti P, Cervera N, et al. How different are luminal A and basal breast cancers? *Int J Cancer.* 2009;124(6):1338-1348. doi:10.1002/ijc.24055.