Data Analytics to Predict Metastatic Breast Cancer Subtype: Moving Towards Precision Medicine and Personalized Care

Joshua Burd, Donovan Guttieres, Liang Li, Ege Ozgirin, and Sitara Persad

IDS.131 Graduate Biology
December 13, 2017

Contents

1. Background and Motivation

Cancers are predominantly characterized by mutations within the genome. Almost no tumors have the exact same mutation profile, giving rise to multiple cancer subtypes with diverse clinical outcomes. This makes cancer both complex and heterogeneous, with added difficulty to determine which treatments patients will be most responsive to and the eventual outcome of the pathology. In an effort to better understand the molecular processes driving tumor progression, multiple initiatives such as the The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have sought to gather multi-omic (genomic, proteomic, transcriptomic, metabolomic, etc.) data to build profiles for thousands of tumors and enable stratification based on clinically and biologically meaningful subtypes (1). While this has been limited by the availability to extract relevant information from data, efforts have been made to link mutation profiles with gene interaction networks. This allows to use variable selection methods, clustering, network smoothing and other methods to analyze mutation profiles in the context of prior knowledge on molecular pathways in order to label tumors into specific subtypes in ways that inform both understanding of the underlying biologic network leading to disease progression and clinically relevant treatment options.

According to a research from the National Cancer Institute, the number of breast cancer will increase by 50% by 2030. With the increased prevalence of breast cancer in the US, as well as in other parts of the world, there is increased effort to curb both mortality and morbidity. When breast cancer is first detected, tests are done to study cancer cells in ways that can inform the rate of growth, likelihood the cancers cells will spread to other parts of the body (metastasis), how well certain treatments might work, and how likely the cancer is to recur. Additional information can be gleaned from the presence of hormone receptors, size of tumor, family history, and additional clinical information. The prognosis, or evolution of the breast cancer, will be heavily influenced by all the factors mentioned above and thus inform which therapeis by be most appropraite. Various multi-omic techniques have been employed in an effort to predict cancer subtype, thus personalizing treatment options to a patient's unique tumor characteristics. There are two frequently used classifications for cancer subtype: molecular subtype classes (basal-like, luminal A, luminal B, HER2-enriched, and normal-like) with unique genetic signatures and phenotypic subtypes (breast carcinoma, invasive ductal carcinoma, invasive lobular carcinoma, and mixed ductal/lobular carcinoma) deduced from histology. In both cases, knowledge on cancer subtype classification serves as an important step towards more patient-centric care. For the purposes of this project, we focus on phenotypic subtypes since that is the main onco-classifier provided in the dataset.

A wide range of data types and variables may be useful in predicting cancer subtype. These include both non-genetic and genetic data. This project is aimed at determining the relative value for these different data types on predicting metastatic breast cancer subtype. Non-genetic data includes a wide range of diagnostic (e.g. hormone levels, tumor size, number of lymph nodes, location of metastasis) and patient/demographic (e.g. age, race, etc.) information. The genetic data provided is based on copy number alterations (CNA), that provides gene-level data on the occurence and type of genetic mutation. Sequencing of dignostic biopsies from cancer tumors can be easily done and the applicability of CNA for stratifying tumor subtypes provides a rich source of data for analysis (2). While thousands of genes are involved in underlying molecular and signaling pathways, with many potentially mutated, emerging throughput sequencing techniques allow to systematically analyze differences and similarities between tumors on the basis of their genetic trait (3). It will be especially relevant to see whether certain genes act as influential tumor suppressor genes and oncogenes, which can potentially explain phenotypic variation

2. Data Description

This project is a trial analysis of a new dataset released by the Metastatic Breast Cancer (MBC) Project - a collaboration between the Broad Institute of MIT and Harvard, the Dana-farber Cancer Institute, and a large coalition of nonprofit and patient partners (4). This initiative combines a crowd sourced, social media-driven effort to engage women with metastatic breast cancer with high trust clinical partnerships - allowing the creation of a large population of patient and tumor-specific genomic and phenotypical information. As of December 2017, more than 3,900 patients have self-reported information about their breast cancer diagnoses, treatment regimes and outcomes, and of that population more than 2,400 have consented to sharing their clinical records and tumor samples, while more than 1,400 of those have provided saliva samples for whole genome sequencing.

On October 30, 2017 the MBC Project released their first publicly available data set. It includes:

- Whole exome sequence (WES) data from 103 tumor samples obtained from 76 patients, in the form of mutations, insertions, and deletions in genes' copy number, along with diagnostic, detailed pathology (PATH) and treatment history information from medical records (MedR) of each patient. PATH contains the phenotypic cancer subtype diagnosis determined by examining biopsy cells and tissues under a microscope.
- For the same 76 patients: self-reported demographic, treatment response information collected from surveys (PRD)

The intention of this project is to test clustering and regression analysis methods on this small public data release in preparation for future larger data releases. For this project, our group accessed this data set through the cBioPortal, an open source web application hosted by the Center for Molecular Oncology at at Memorial Sloan Kettering Cancer Center. Three data sets labeled "data_CNA.txt" (genetic data), "data_clinical_patient" (patient specific survey data), and "data_clinical_sample" (tumor specific clinical record data) were downloaded from cBioPortal web interface.

2.1. Genetic Data. This data represents the presence of mutations for each tumor sample derived from Whole Exome Sequencing (WES). The data form was a matrix of mutation types of 22k+ genes for 103 tumor samples. A python dictionary was used to map different types of mutation (homozygous deletion, hemizygous deletion, no mutation, gain, high level amplification) to the integer values on the interval [-2,2].

2.2. Patient and Sample Data. Information was provided for 103 total tumor samples from 76 patients. Patient specific and tumor specific data sets were merged by copying patient information to the tumor samples taken from those patient.

The data set drawn from medical records was very sparse because of different data entry practices in different clinical settings. This support ongoing efforts to push for industry standardization and interoperability of electronic health records. Any variable for which >40% samples did not report a value was deleted.

The remaining data contained information on: cancer subtype classification, successful therapies, age at diagnosis, time between diagnosis of cancer and diagnosis of metastatic breast cancer, cancer classified as HR+, ER+ and/or PR+ (hormone receptor status), HER2 classification, racial demographic, cancer stage at diagnosis, tumor sample quality, biopsy procedure type, reaction to radiation therapy, location of metastasis, and several other specific variables representing the presence of known oncogene mutations.

This data was in the form of heterogeneous text entries. Python dictionaries mapping the key terms in these text entries to integer values were created to methodically create data compatible with regression and clustering methods. Predictive mean matching (implemented in python using the mice.impute.pmm() function) was used as a general purpose imputation method. This algorithm restricts imputations to observed values, which worked well for our integer codings of clinical terms.

3. Method

In order to explore the questions set out in the previous section, we employed several analysis approaches, consisting of logistic regression with regularization and clustering.

3.1. Regression Analysis. We used logistic regression to answer the following questions relating to predicting **cancer subtype** (BREAST, IDC, MLDC, IDC).

1. Can we predict cancer subtype from the copy number alteration data?
2. Can we predict cancer subtype from the sample clinical and patient reported data?
3. How well does a model combining these features perform?
4. Which mutated genes and clinical features are most influential indicators for predicting subtype?

We also explored similar questions relating to predicting **effective therapy** as well as **tumour location** after metastasis.

For each question, we performed logistic regression, matching the data to the cancer subtype label by the tumor sample ID. We generated a training and hold-out validation set comprising 40% and 15% of the original dataset, respectively. We ensured that the distribution over categories was similar for the training, validation and test set, so that we should achieve reasonable generalization.

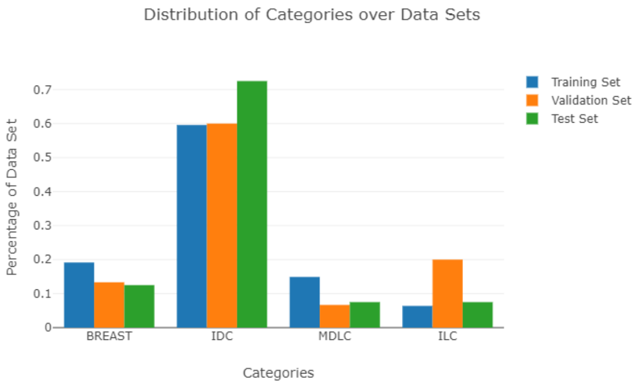


Fig. 1. The distribution of categories is consistent throughout each dataset.

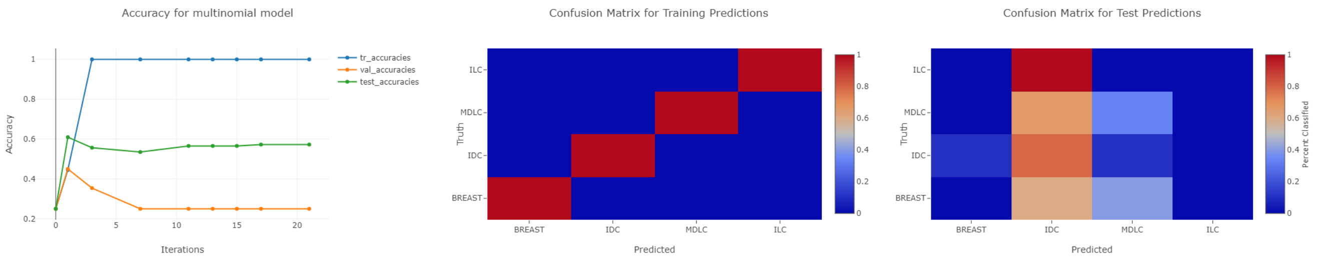
We tested the performance of the one-vs-rest and the multinomial approach, selecting the model with highest accuracy. We evaluated the performance of both L_1 and L_2 regularization.

3.2. Feature Selection. Performing L_1 regularization resulted in sparse co-efficients, which we then examined further. We considered all the non-zero weights which corresponded to clinical and patient reported data, as well as informative genes from the CNA data.

4. Cancer Subtype Prediction

4.1. From Copy Number Alteration (CNA) Data.

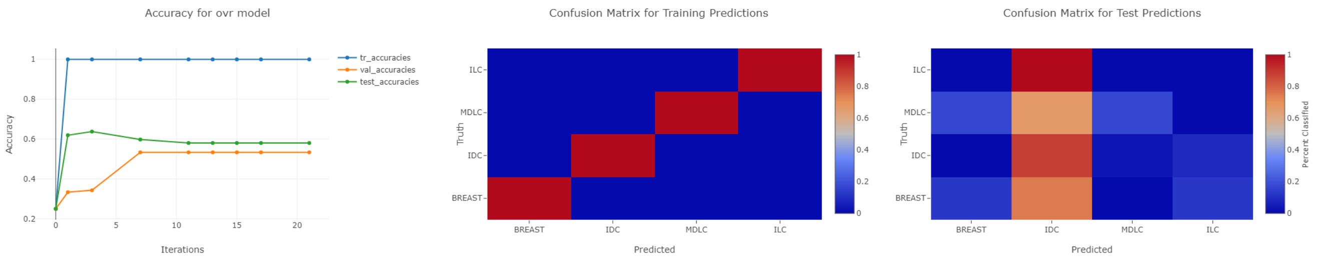
4.1.1. Ridge Regression. We observe that we are able to achieve perfect accuracy on the training data set using CNA information, but that this accuracy does not generalize to the test set, despite using a hold-out set. We hypothesize that this generalization error stems primarily from the data imbalance (60% of the data was of type IDC). Secondly, the dataset released is quite small, so the test set contained relatively few examples.



(a) F1 Accuracy on Training, Validation & Test Data **(b)** Confusion Matrix showing Predictions on Training Set **(c)** Confusion Matrix showing Predictions on Test Set

Fig. 2. Results using CNA Data with L_2 Regularization

4.1.2. LASSO Regression. We observe similar accuracy using L_1 regularization, and slightly better performance on the test data. The LASSO model enforces sparsity which leads to a more interpretable model, as we will discuss in the following sections.



(a) F-1 Accuracy for Training, Validation and Test Data **(b)** Confusion Matrix showing Predictions on Training Set **(c)** Confusion Matrix showing Predictions on Test Set

Fig. 3. Results using CNA Data with L_1 Regularization

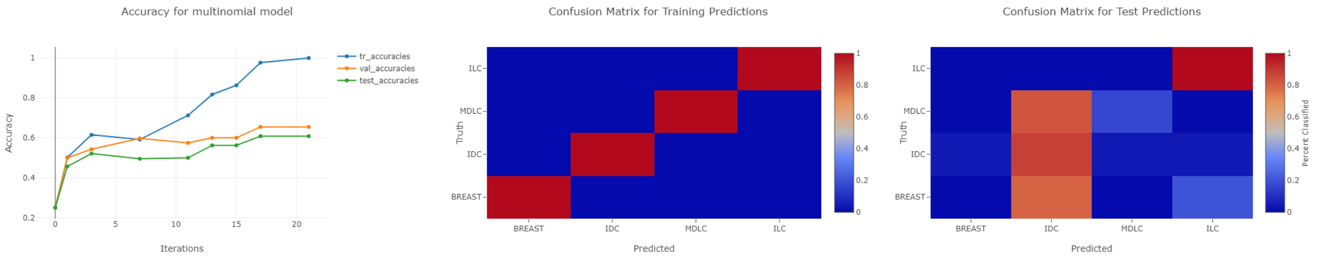
4.2. From Clinical and Patient Reported Data. Overall, the clinical and patient reported data gave very good prediction performance, which was a surprise to us, as we expected the genetic data (CNA) to have a strong predictive effect on cancer subtype. However, this discrepancy may be due to the fact that the cancer subtypes are phenotypic labels, which may not correspond with the underlying molecular mechanisms.

4.2.1. Ridge Regression. We observe a slightly better performance on the test data using sample data, with less of a tendency to simply predict IDC based on the data imbalance.

4.2.2. LASSO Regression. Using L_1 penalty, we observe worse performance on the training set, but much better generalization performance on the test set, which does quite well classifying more tumor subtypes.

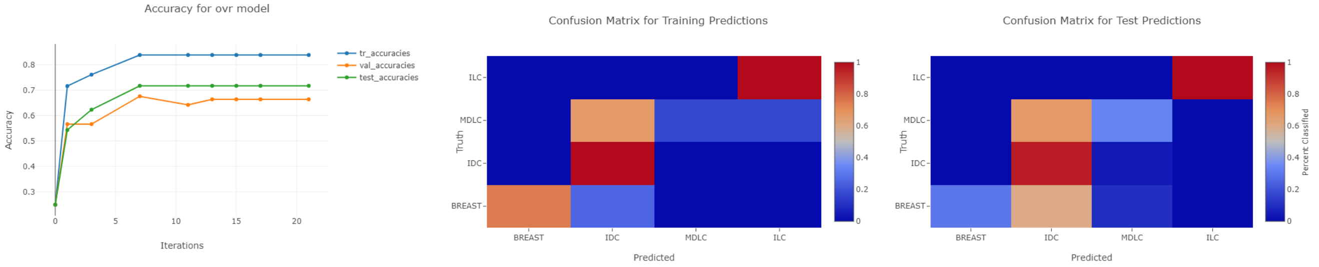
4.3. From CNA and Clinical/Patient Reported Data. We sought to train the best possible model, incorporating data from both clinical/patient reported data and CNA sequencing data.

4.3.1. Ridge Regression. We observe very high weighted accuracy on both the test and training data.



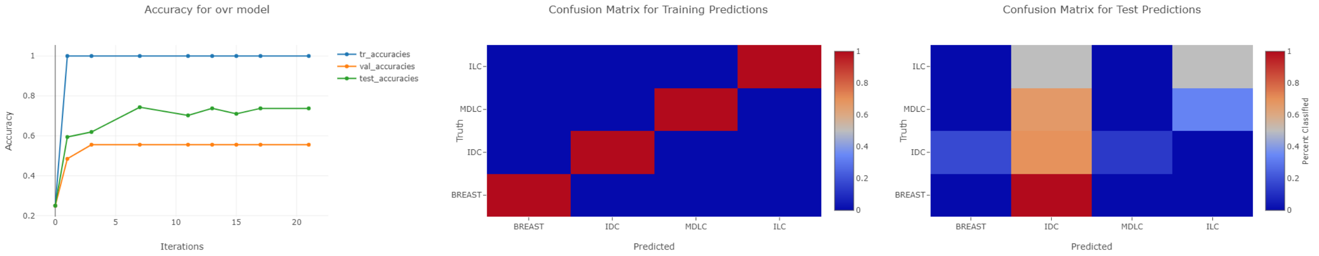
(a) F-1 Score for Training, Validation and Test Data **(b)** Confusion Matrix showing Predictions on Training Set **(c)** Confusion Matrix showing Predictions on Test Set

Fig. 4. Results using Clinical and Patient Data with L_2 Regularization



(a) F-1 Accuracy Score for Training, Validation and Test Data **(b)** Confusion Matrix showing Predictions on Training Set **(c)** Confusion Matrix showing Predictions on Test Set

Fig. 5. Results using Clinical and Patient Data with L_1 Regularization



(a) F-1 Score for Training, Validation and Test Data **(b)** Confusion Matrix showing Predictions on Training Set **(c)** Confusion Matrix showing Predictions on Test Set

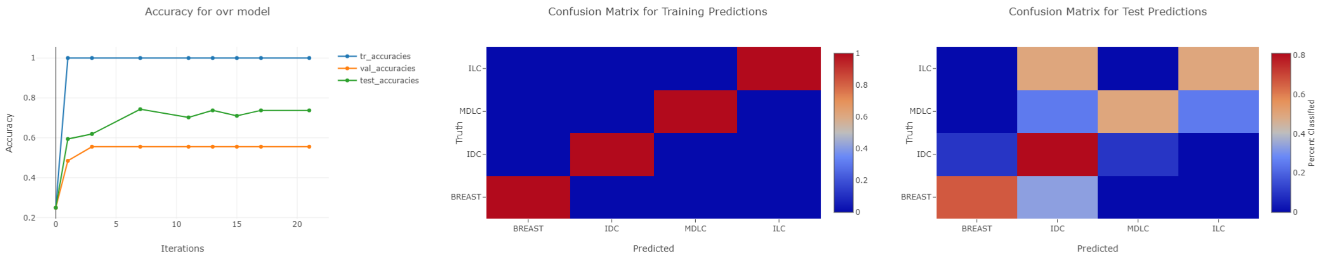
Fig. 6. Results using CNA and Clinical and Patient Data with L_2 Regularization

4.3.2. LASSO Regression. This model gives the best performance on the test set. We observe that many predictions are within the correct category. Furthermore, the misclassifications themselves are informative. For example, note that MDLC (which is Mixed Ductal and Lobular Carcinoma) is often classified as Ductal or Lobular. Secondly, predictions on ILC tend to be either ILC or IDL, suggesting possible similarities between these cancer subtypes at the sequence level.

5. Model Interpretation

Using the best predictive model obtained using L1 regularization, we look at the non-zero weighted variables to understand the factors most important in predicting cancer subtype.These are a mixture of clinical and patient data, as listed in Table 1 below.

We also extract the most influential mutated genes in predicting cancer subtype and visualize these on Figure 8 below, highlighting the location of these genes along the chromosomes. The mapping shows the following



(a) F-1 Score for Training, Validation and Test Data **(b)** Confusion Matrix showing Predictions on Training Set **(c)** Confusion Matrix showing Predictions on Test Set

Fig. 7. Results using CNA and Clinical and Patient Data with L_1 Regularization

Cancer Sub Type	Important Clinical/ Patient Features
BREAST	Age at diagnosis, Time between diagnosis and metastasis, Time between diagnosis and sample collection, Biopsy Location,
IDC	Time between diagnosis and metastasis, Size of Tumor, Time between diagnosis and sample collection, Location where biopsy was performed
MDLC	Time between diagnosis and metastasis, Size of Tumor, Number of Lymph Nodes Spread To, Time between diagnosis and sample collection, Location where biopsy was performed, Quality of biopsy
ILC	Time between diagnosis and metastasis, , Time between diagnosis and sample collection,Location where biopsy was performed, Quality of biopsy

Table 1. Important clinical and patient variables for predicting cancer subtype

number of genes for each cancer subtype: BREAST (45), IDC (72), MDLC (71), and ILC (23). Of the total genes mapped, 54 of them are deemed influential mutations for more than one cancer subtype. It is interesting to note the distribution of certain gene mutations are heavily clustered in chromosome 8, 11, and 17. This points to possible future areas of work, for example to determine the influence of physical proximity of mutuations on tumor progression and potential novel therapeutic targets.

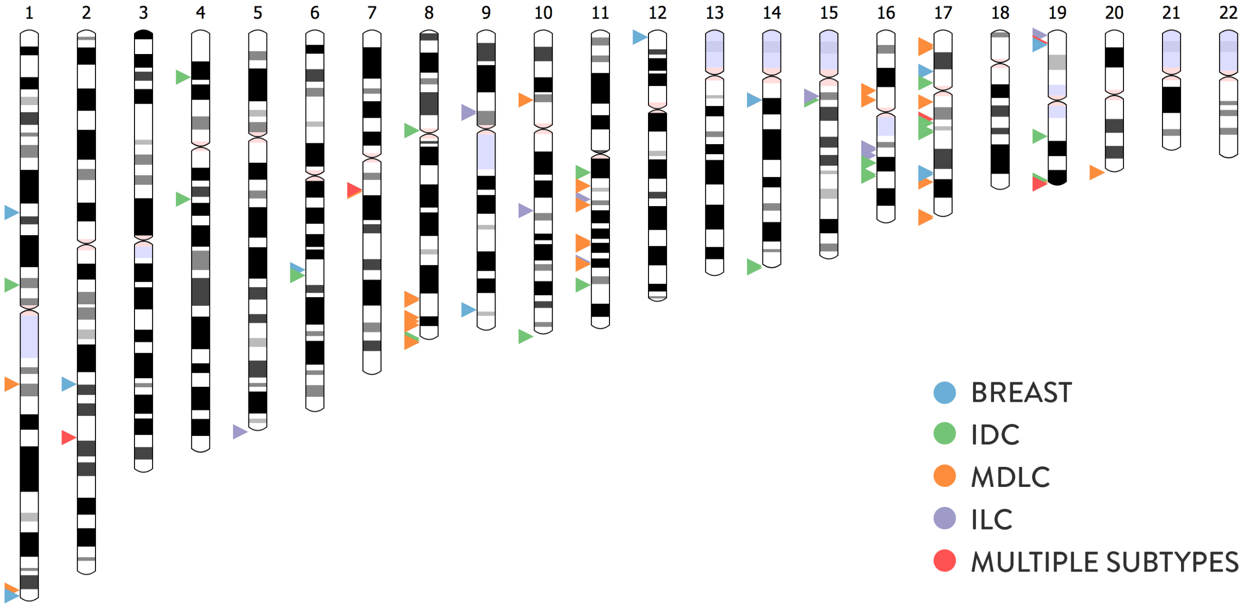


Fig. 8. Locations of the identified genes within the chromosomes.

Some of the mutuated genes highlighted above have an important role in cancer progression. Whole-genome sequencing has made it possible to overlay genetic data unto molecular networks that drive disease evolution. Some of the relevant genes found and their influence on well-established molecular breast cancer networks

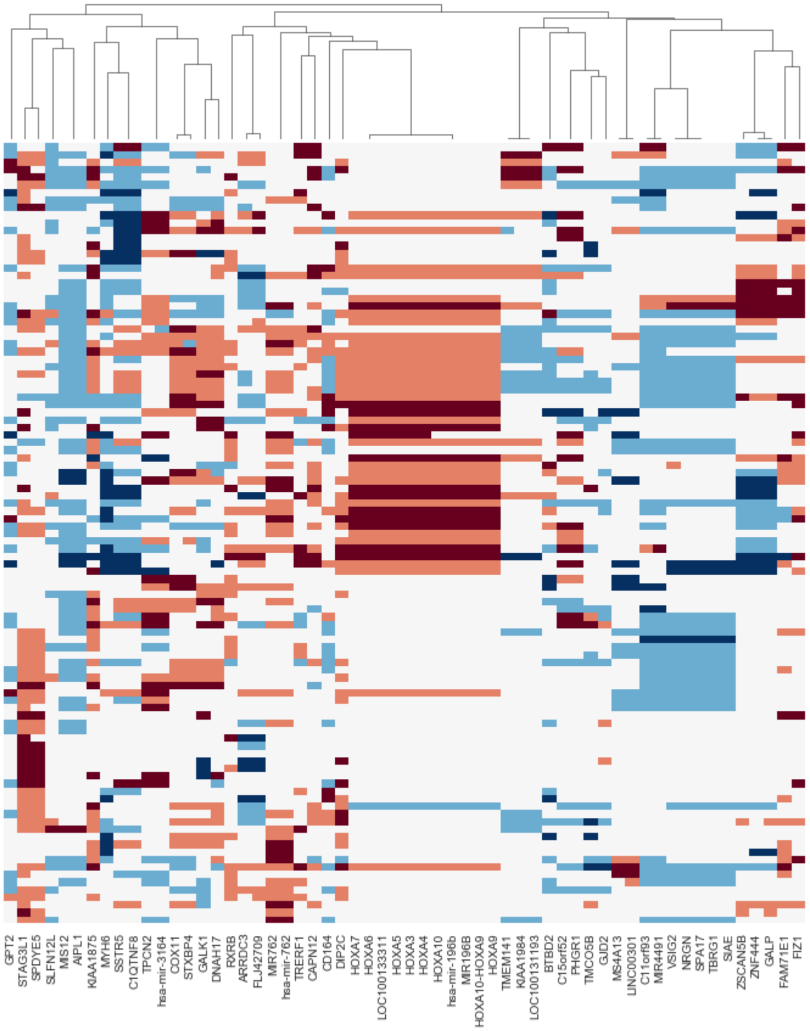


Fig. 9. Correlation analysis and hierarchical clustering reveals gene families associated with different cancer subtypes.

such as the KEGG pathway (link) include the following:

- CCDN1 - amplification of this gene in the q13 region of chromosome 11 is found to lead to cyclin D1 protein overexpression in breast cancer. Previous studies have shown Cyclin D1 acts as an oncogene when overexpressed, given its role in enhancing DNA synthesis in the cell cycle and leading to tumorigenesis.(5)
- GADD45B - this gene, found in the p13.3 region of chromose 19, is a member of the growth arrest DNA damage-inducible gene family associated with cell growth control, apoptosis, and DNA damage repair response. Under stressful environments, expression increases and activates the p38/JNK pathway. However, downregulation through inhibition leads to resistance to environmental stresses, preventing apoptic signals and thus making it a lucrative area of research for potential cancer drug targets.(6)

Next, we compute the correlation between genes' CNA profiles across the different tumor samples, and cluster the genes based on these correlations using hierarchical clustering. We observe several distinct families based on these correlations, some of which make intuitive sense, such as the HOX family cluster, while others are less obvious. Further work should be done to understand the relationships within these families and the influence on downstream protein-protein interaction.

However, we can observe certain genes that are known to be related to cancer. For example, TBRG1 is shown to be negatively correlated in our analysis, and has previously been shown to collaborate with CDKN2A to restrict cancer proliferation. Similarly, the HOX genes are shown to be positively correlated, as well as MIR762 which promotes breast cancer cell proliferation and invasion.

6. Predicting Therapy and Location of Metastasis

We analyzed the performance of our classifier on predicting location, but found that the model could not reliably predict location of metastasis, achieving 20% accuracy. Similarly, we were able to predict therapies which would be highly effect with accuracy of 35% at best.

7. Conclusion

- Effectively predicting phenotypic cancer subtype is difficult. CNA and patient/sample data show promise in achieving this, but can be complemented with other data sources (e.g. gene expression) in order to make better predictions.
- The best cancer subtype prediction we achieved in this project was 78% balanced accuracy using using LASSO regression on merged clinical and CNA data.
- We analyzed selected features and found influential gene mutations. This valuable list of targets provides insight into potential downstream protein effects and the underlying molecular basis that influence both disease evolution and effectiveness of therapies.
- Methodologies outlined provide the foundation for further research as the study is expected to expand to more than 1500 patients in the near future.

8. Further Work

Cancer genes need to be understood as being part of a complex network (7). We identified the most important genes in determining the breast cancer subtypes, and the future work could be to develop a gene regulatory network to analyze the molecular mechanistic interplay and physical interactions between those individual genes.

9. Code

The code used for this project is available on github.

Bibliography

1. Hofree M, Shen JP, Carter H, Gross A, Ideker T (2013) Network-based stratification of tumor mutations. *Nature methods* 10(11):1108–1115.
2. Gusnanto1 A, et al. (2015) Stratifying tumour subtypes based on copy number alteration profiles using next-generation sequence data. *Bioinformatics* 31(16):2713–2720.
3. Ciriello G, et al. (2013) Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics* 45:1127–1133.
4. (2017) The Metastatic Breast Cancer Project (Provisional, October 2017) (http://www.cbioportal.org/study?id=brca_mbcproject_wagle_2017#summary).
5. Elsheikh S, et al. (2008) Ccnd1 amplification and cyclin d1 expression in breast cancer and their relation with proteomic subgroups and patient outcome. *Breast Cancer Research and Treatment* 109(2):325–335.
6. Tamura R, et al. (2012) Gadd45 proteins: central players in tumorigenesis. *Current molecular medicine* 12(5):634–651.
7. Kitano H (2004) Cancer as a robust system: implications for anticancer therapy. *Nature Reviews Cancer* 4(3):227–235.

Acknowledgements

We extend our appreciation to Professor Uhler and Professor Jegelka for all they have taught us this semester and for their guidance on this project. We are also grateful to the course teaching assistants, particularly Tahin, for his helpful feedback and advice. Thanks to the patients who donated their samples to this public dataset in the name of scientific advancement. Finally, thanks to the IDSS coffee and hot chocolate machine, without which this project would not be possible.