Breast Cancer Prognostics Using Multi-Omics Data

Sisi Ma, PhD, Jiwen Ren, MS, David Fenyö, PhD
Center for Health Informatics and Bioinformatics,
New York University Langone Medical Center, New York, NY

Abstract

Breast cancer affects one in eight women in America and is a leading cause of death from cancer worldwide. In the current study, four types of Omics data including copy number variation, gene expression, proteome and phosphoproteome were collected from seventy-seven breast cancer patients. Individual types of Omics data were used to separately construct predictive models to predict ten-year survival, an important clinical hallmark. The predictive models constructed with proteome data achieved decent predictivity (mean AUC = 0.725) and outperforms the models constructed with other types of Omics data. This indicates that high quality, large scale protein data is more effective for survival prediction compared to other types of omics data. Further, we experimented with ten different data fusion techniques (generic and Multi-kernel learning based) to test whether combining multi-Omics data can result in improved predictive performance. None of the data fusion techniques tested in the current study outperforms the predictive models built with the proteome data.

Introduction

Breast cancer affects one in eight women in America, and is a leading cause of death from cancer worldwide (1). Extensive genomic characterization of breast cancer has been conducted in the last ten years, leading to clinically relevant molecular subtyping (2), increased accuracy in prognostication (3-5), and success in targeted therapy (6, 7). One missing link in our knowledge is how genomic changes translate into changes in proteome and phosphoproteome which in turn execute the phenotypic characteristics of the disease. Since the proteome and phosphoproteome are more proximal to the manifestation of the disease, it is likely that they contain more predictive information regarding the clinical outcomes of the disease and could lead to the development of more accurate prognostic technologies. While proteomic characterization was performed in the Cancer Genome Atlas (TCGA) breast cancer study using reversed phase protein arrays (8), only 171 cancer-related proteins and phosphoproteins were quantified. To provide greater depth of proteomics characterization, the NCI Clinical Proteomic Tumor Analysis Consortium (CPTAC) (9) analyzed the global proteomes of genomically annotated TCGA breast tumor specimens.

In the current study, we utilized machine learning method to predict ten-year survival, an important clinical hallmark, of breast cancer patients. First, we constructed predictive models with four different types Omics data, including copy number variation, gene expression, proteome, and phosphoproteome, to determine which type of Omics data is the most predictive regarding ten-year survival. Then, different data fusion techniques were employed to combine the above mentioned four types of Omics data, as an effort to improve the predictive performance obtained from the models based on a single type of Omics data. Data fusion for data coming from difference sources has been an active area of research in machine learning. Data fusion techniques have been applied to solve problems in various domains, including robotics, market research and medicine (10-13). However, to the best of our knowledge, the current study is one of the first to benchmark the efficacy of different data fusion technique on genome-scale multi-Omics data, since global proteome data has only recently became available.

Identifying the type of Omics data or the combination of multi-Omics data that lead to the best predictive performance is critical in the following ways. First, it could serve as a guide for experimentalists and clinicians, such that more experimental resource can be dedicated to the identified Omics data domain and generate new data for mechanistically relevant discovery and hypothesis testing. Second, the mathematical predictive models constructed with the most informative Omics data type(s) could be readily translated into prognostic technologies and used in a clinical setting.

Methods

Data

Copy number variation, gene expression, proteome and phosphoproteome data for 77 breast cancer tumors characterized by the TCGA were utilized to predict the ten-year survival of the patients. 11 out of the 77 patients did not survive longer than ten-years. For copy number variation and gene expression data, the original characterization from TCGA were used, and the proteome and phosphoproteome data was generated by the Broad Institute as a part of the CPTAC Consortium (14).Tumor samples were analyzed by high-resolution, accurate mass tandem mass spectrometry that included extensive peptide fractionation by high pH reversed phase chromatography and phosphopeptide enrichment by immobilized metal affinity chromatography (IMAC). An isobaric peptide labeling approach (iTRAQ) was employed to quantify protein and phosphosite levels across samples, with 37 iTRAQ 4-plexes analyzed in total. Each 4-plex contained 3 samples from different subtypes and a common reference that was created by pooling material from 40 tumors (with equal representation by weight of the 4 subtypes of breast tumors). 24,174 features of copy number variation, 16,525 features of gene expression, 12,553 features of proteome, and 32,939 features of phosphoproteome characterizing the global molecular profile of the breast tumors were used in the present study.

Predictive Models

All analytical experiments conducted in this study are summarized in table 1.

To assess which set of Omics data is the most informative in predicting ten-year survival, predictive models were built with each type of Omics data, i.e. copy number variation, gene expression, proteome, and phosphoproteome respectively.For individual type of Omics data, we trained the following five classification models: (A) support vector machine (SVM) with linear kernel, (B) SVM with polynomial kernel, (C) SVM with rbf kernel (15), (D) Bayesian logistic regression (16), and (E) random forest (17). For the SVM models and Bayesian logistic regression, hyperparameters were selected via nested cross validation. Three feature selection strategies were explored: (a) no feature selection: all features were used; (b) univariate association: all features that are associated with the target univariately were selected; (c) SVM RFE: features were selected via recursive elimination process based on the predictive performance of a linear SVM model (18).

To assess whether combining different types of Omics data can improve the predictive performance over predictive models built with a single type of Omics data, we explored the following data fusion techniques (summarized in Figure 1): (i) All Omics concatenated: features from individual Omics data concatenated. (ii) All Omics concatenated with feature selection: features from individual Omics data concatenated with feature selection employed on the concatenated features. (iii) Selected features concatenated: features were selected first for individual types of Omics data by some feature selection method. The selected features were then concatenated.(19) (iv) Multi-kernel learning (MKL): 8 kernel matrices (using linear kernel; polynomial kernel with degree 1, 2, and 3; and rbf kernels with sigma (10^{-6}, 10^{-4}, 10^{-2}, and 10^{-1}) were computed for individual types of Omics data, and MKL was applied for classification. (v) Selected features MKL: feature selection was conducted on individual type of Omics data. Then, 8 kernel matrices were computed for the selected features for individual type of Omics data. MKL was applied for classification (20, 21). It is worth noting that, different from the previous two data fusion techniques, which only takes care of combining data from different sources, the MKL also learns the decision boundary for the classification. The above five data fusion techniques were used to combine two sets of Omics data: the first set comprised of data types including copy number variation, gene expression, proteome, and phosphoproteome. We denote data fusion strategies resulting from this set with "_4"; the second set comprised of data types including proteome and phosphoproteome. We denote data fusion strategies resulting from this set with "_2"As a result, ten data fusion strategies were obtained: (i)_4 All Omics concateneated_4, (ii)_4 All Omics concatenated with feature selection_4, (iii)_4 Selected features concatenated_4, (iv)_4 MKL_4, (v)_4 Selected features MKL_4, (i)_2 All Omics concateneated_2, (ii)_2 All Omics concatenated with feature selection_2, (iii)_2 Selected features concatenated_2, (iv)_2 MKL_2, and (v)_2 Selected features MKL_2. The rationale of combining proteome and phosphoproteome is that these two type of Omics data results in better predictive performance when used to build predictive models individually. Figure 1 gives a graphical illustration of the above mentioned ten data fusion strategies.

For data fusion strategy (i)-(iii), we trained the following five classification models: support vector machine (SVM) with linear kernel, SVM with polynomial kernel, SVM with rbf kernel, Bayesian logistic regression, and random forest. For data fusion strategies (iv) and (v), MKL were used as the classifier. For the SVM models, Bayesian

logistic regression and MKL, hyperparameters were selected via nested cross validation. Two feature selection strategies were explored for data fusion strategies (ii), (iii), (v), univariate association and SVM RFE.

The SVM models were constructed with LibSVM software (22). The Bayesian logistic regression models were constructed with the BBR software (23). And the random forest models were constructed with the random forest package in R (24). The SKMsmo software was used for MKL (20, 21).

Performance Estimation and Statistical Comparison

For performance estimation, 4 fold cross validation were implemented with 10 stratified (on ten year survival) random splits, resulting in 40 performance estimates for each model. Area under the ROC curve was used as the performance estimation metric. Paired sample t test were used to compare the performance between pairs of models. p-values were FDR adjusted to correct for multiple comparisons (25).

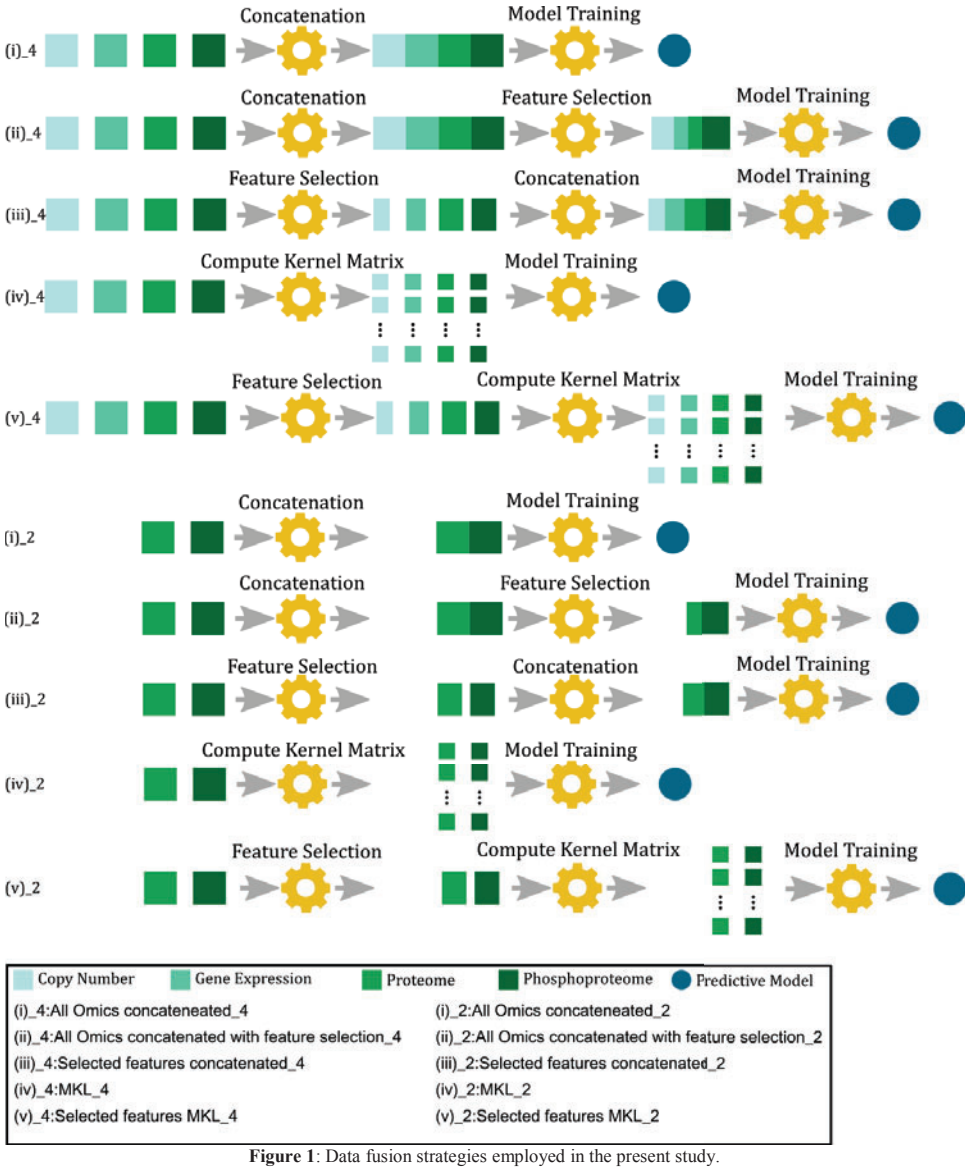

Figure 1: Data fusion strategies employed in the present study.

Data Used for Modeling	Feature Set/Data Fusion Stratergy Name	Classifier	Feature Selection Method		
			No Feature Selection	Univariate Association	SVM RFE
Single Omics Data	Copy Number Variation	classifier(A-E)	✔	✔	✔
	Gene Expression	classifier(A-E)	✔	✔	✔
	Proteome	classifier(A-E)	✔	✔	✔
	Phosphoproteome	classifier(A-E)	✔	✔	✔
Combination of 4 types of Omics Data	(i)_4:All Omics concatenated_4	classifier(A-E)	✔		
	(ii)_4:All Omics concatenated with feature selection_4	classifier(A-E)		✔	✔
	(iii)_4:Selected features concatenated_4	classifier(A-E)		✔	✔
	(iv)_4:MKL_4	MKL	✔		
	(v)_4:Selected features MKL_4	MKL		✔	✔
Combination of 2 types of Omics Data	(i)_2:All Omics concatenated_2	classifier(A-E)	✔		
	(ii)_2:All Omics concatenated with feature selection_2	classifier(A-E)		✔	✔
	(iii)_2:Selected features concatenated_2	classifier(A-E)		✔	✔
	(iv)_2:MKL_2	MKL	✔		
	(v)_2:Selected features MKL_2	MKL		✔	✔

Table 1: Classification models and feature selection methods used for 4 individual Omics data and data fusion strategies used to combine 4 individual Omics data/ 2 individual Omics data. For some data fusion strategies, not all feature selection methods were explored, to avoid repetition. For example (i)_4 with univariate association is the same as (ii)_4 with univariate association.

Results

Proteomics Data is the most Informative in Predicting Ten-year Survival

Among the predictive models constructed with a single type of Omics data, models constructed with proteomics data achieved better predictive performance (as measured by AUC) for all classification models except random forest. The random forest classifier did not perform well in general in the current study. The best performance of the random forest results in a AUC of 0.634±0.157 using copy number data modality and univariate feature selection method. Feature selection using univariate association improved the predictive performance compared to when feature selection was not employed, whereas feature selection by SVM-RFE did not improve the predictive performance in most of the cases (Table 2). Figure 2 shows the average AUC of the models trained with SVM rbf using univariate association as the feature selection method. The predictive performance for the model trained with the proteomics data (0.725±0.222, Mean±Std) is significantly better (p-values<0.05) compared with the models trained with phosphoproteome (0.671±0.234), gene expression (0.547±0.207), and copy number variation (0.463±0.218). The predictive performance of models constructed with features from a single type of Omics data with SVM rbf as the classifier and three different feature selection methods were shown in Table 2. The predictive performance of the other classification methods follows a similar pattern as the SVM rbf (see online appendix table 1 at sisima.net/breast_cancer_study/index.html).

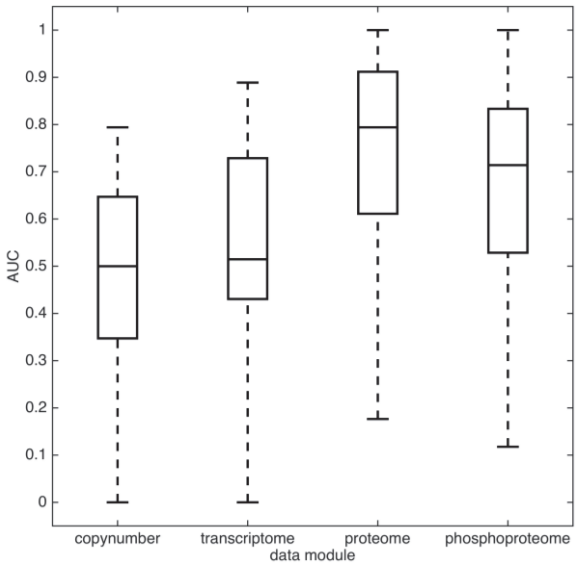

Figure 2: The AUCs for predictive models built with one type of Omics data with SVM rbf and univariate association for feature selection. Models built with proteome data out performs models built with other data modules.

Combining Multi-Omics Data did not Improve Predictive Performance in the Current Experimental Setting
To examine whether combining multi-Omics data could improve the predictive performance of the models constructed with data from a single type of Omics data, five different data fusion strategies were employed. The five different methods were used to combine two sets of Omics data: (i) all four types of Omics data available, (ii) proteome and phosphoproteome. Ten data fusion strategies were obtained from the above process. None of the ten data fusion strategies tested in the current study resulted in significantly better predictive performance compared to the model constructed with the proteome data alone (Figure 3 and Table 2). When using the SVM rbf classifier the best data fusion strategy is (vii) All Omics concatenated with feature selection_2 with univariate association as the feature selection method. This strategy achieved a predictive performance of 0.676±0.234 and is not significantly different from the model built with proteome data (0.725±0.222). It is worth noting that strategy (ii) and strategy (iii), as well as strategy (vii) and strategy (viii), are equivalent when univariate association was applied as the feature selection method, since the order of applying data concatenation and applying univariate association as feature selection does not affect what features would be selected. The predictive performance of the other classification methods follows a similar pattern as the SVM rbf (see online appendix table 1 at sisima.net/breast_cancer_study/index.html).

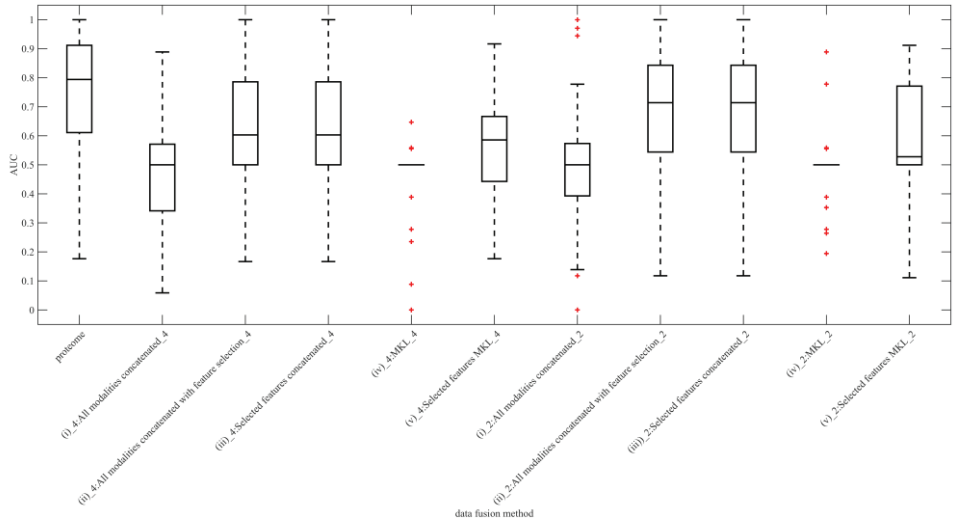

Figure 3: Predictive performance for model built with proteome data vs. that with data fusion methods. Figures shows predictive performance of models built with SVM rbf with univariate association as feature selection methods when applicable.

Data Used for Modeling	Feature Set/Data Fusion Stratergy Name	Classifier	Feature Selection Method		
			No Feature Selection	Univariate Association	SVM RFE
Single Omics Data	Copy Number Variation	SVM rbf	0.551(0.187)	0.463(0.218)	0.546(0.244)
	Gene Expression	SVM rbf	0.431(0.152)	0.547(0.207)	0.431(0.276)
	Proteome	SVM rbf	0.557(0.263)	**0.725(0.222)**	0.536(0.262)
	Phosphoproteome	SVM rbf	0.516(0.257)	0.671(0.234)	0.582(0.219)
Combination of 4 types of Omics Data	(i)_4:All Omics concatenated_4	SVM rbf	0.482(0.216)		
	(ii)_4:All Omics concatenated with feature selection_4	SVM rbf		0.607(0.217)	0.517(0.219)
	(iii)_4:Selected features concatenated_4	SVM rbf		0.607(0.217)	0.476(0.259)
	(iv)_4:MKL_4	MKL	0.469(0.118)		
	(v)_4:Selected features MKL_4	MKL		0.561(0.170)	0.504(0.268)
Combination of 2 types of Omics Data	(i)_2:All Omics concatenated_2	SVM rbf	0.495(0.235)		
	(ii)_2:All Omics concatenated with feature selection_2	SVM rbf		0.676(0.235)	0.566(0.222)
	(iii)_2:Selected features concatenated_2	SVM rbf		0.676(0.235)	0.557(0.253)
	(iv)_2:MKL_2	MKL	0.494(0.109)		
	(v)_2:Selected features MKL_2	MKL		0.588(0.190)	0.530(0.233)

Table 2: Predictive Performance of models constructed with a single type Omics data vs. that with multi-Omics Data. Results are shown for SMV rbf classifier and MKL. Different feature selection methods were applied when applicable. Results are shown in the form of mean AUC with standard deviations inside a pair of Parentheses.

Discussion

The primary contribution of this study is to show that the global/system level proteome data have superior predictive value for predicting ten-year survival for breast cancer patients. The predictive signal in the proteome data is significantly greater compared to that in the copy number variation data and the gene expression data. This is likely due to the fact that changes on the gene level need to manifest in alternations on the proteome level to drive phenotypical, functional and pathological changes. Examining the most predictive proteins for ten-year survival can

help identify the biological and/or pathological pathways leading to resilience/vulnerability of breast cancer death. Molecular treatment targets might be discovered from these pathways.

This work can be extended and validated by examining a larger population of breast cancer patients. The current sample size (N=77) is limited by the availability of proteome data and phosphoproteome data. The superior predictive value of the proteome for ten-year survival identified in this study warrants the collection of global/system level proteome data from more breast tumors and other types of tumors, in order to systematically assess the value of proteome data in cancer outcome prediction. Moreover, of the 77 breast cancer samples, 18 of them were Basal-like type; 12 of them HER2; 23 of them were Luminal A; 24 of them were Luminal B type. In terms of anatomic stages, 7 patients were determined to be Stage I; 50 patients were Stage II; 19 patients were at Stage III; the rest of the patients were at Stage IV. Surveying more breast cancer samples of different molecular subtype and stages and examining their multi-Omics profile could provide more insight into the different pathology of different types of breast tumor and potentially result in tumor type specific prognostic models and treatment strategies.

Furthermore, recent study has indicated that integrating mammography features and selected GWAS data can improve breast cancer diagnosis (26). Similarly, prognostics of breast cancer might be improved by combining molecular profile of the patients (e.g. proteome) with other clinical data sources, such as family history, imaging data, life style information, previous treatment information, medication and various laboratory test results.

Another area for extending the current work is by exploring additional data fusion techniques. In the current study, we explored generic data fusion methods (concatenation, concatenation with feature selection and feature selection prior to concatenation) and data fusion methods based on MKL. None of these methods improved predictivity compared to model built with proteome data. However, other data fusion methods might be more effective. Ensemble methods that build separate predictive models on different types of Omics data and obtain the final prediction through an ensemble classifier might be a viable candidate (27, 28). In addition, data driven feature construction/feature reduction techniques can be examined. Examples of these techniques include principle component analysis (PCA) based feature reduction, kernel PCA based feature reduction(29), and genetic algorithm(30). Moreover, domain knowledge based feature selection and feature construction methods could be beneficial as well. One potential feature construction strategy is to combine proteome data phosphoproteome data by computing the ratio between the quantities of a particular phosphoprotein and its corresponding protein, since this ratio reflects the activation of a particular protein. Ritchie et. al. has provided a comprehensive review of data fusion methods in multi-Omics data fusion including concatenation based methods, transformation based methods, model based methods and knowledge based methods (31). The field will benefit from a comprehensive benchmark study evaluating those data fusion techniques.

Conclusion

The current study examined global/system scale multi-Omics data, and identified proteome data as the most informative data modality for breast cancer prognosis prediction. Various data fusion strategies were implemented for combining global/system scale multi-Omics data, however, these strategies did not result in better predictive performance. It is our hope that future development of data driven and domain knowledge based data fusion methods could lead to improvement in predictive tasks in biomedicine.

Reference

1. Pfister R, Schwarz KA, Janczyk M, Dale R, Freeman JB. Good things peak in pairs: a note on the bimodality coefficient. Frontiers in psychology. 2013;4:700. PubMed PMID: 24109465. Pubmed Central PMCID: 3791391.
2. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000 Aug 17;406(6797):747-52. PubMed PMID: 10963602.
3. Carey LA, Dees EC, Sawyer L, Gatti L, Moore DT, Collichio F, et al. The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. Clinical Cancer Research. 2007;13(8):2329-34.
4. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology. 2009;27(8):1160-7.
5. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clinical Cancer Research. 2005;11(16):5678-85.
6. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. The Journal of clinical investigation. 2011;121(7):2750.
7. Nahta R, Yu D, Hung M-C, Hortobagyi GN, Esteva FJ. Mechanisms of disease: understanding resistance to HER2-targeted therapy in human breast cancer. Nature clinical practice Oncology. 2006;3(5):269-80.
8. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature genetics. 2013;45(10):1113-20.
9. Rivers RC, Mesri M, Kinsinger C, Boja E, Hiltke T, Rodriguez H. NCI's Clinical Proteomic Tumor Analysis Consortium. Cancer Research. 2012;72(8 Supplement):1282-.
10. Adali T, Levin-Schwartz Y, Calhoun VD. Multimodal Data Fusion Using Source Separation: Application to Medical Imaging. Proceedings of the IEEE. 2015;103(9):1494-506.
11. Breur T. Data analysis across various media: Data fusion, direct marketing, clickstream data and social media. Journal of Direct, Data and Digital Marketing Practice. 2011;13(2):95-105.
12. Hall DL, Llinas J. An introduction to multisensor data fusion. Proceedings of the IEEE. 1997;85(1):6-23.
13. Lovett T, O'Neill E, Irwin J, Pollington D, editors. The calendar as a sensor: analysis and improvement using data fusion with social networks and location. Proceedings of the 12th ACM international conference on Ubiquitous computing; 2010: ACM.
14. Philipp Mertinsa DM, David Fenyob, Michael A. Gillettea, Karl R. Clausera, Pei Wangc, Kelly V. Rugglesb, Jana W. Qiaoa, Sean Wangd, Ping Yand, Chenwei Lind, Sherri R. Daviese, Mike Gatzaf, Charles M. Perouf, Venkata Yellapantulag, Michael McLellang, Bing Zhangh, Filip Mundta,i, Li Dingg, Song Caog, Reid Townsende, Henry Rodriguezj, Amanda Paulovichd, Matthew Ellisk, Steven A. Carra, the NCI CPTAC. Proteogenomic analysis of human breast cancer connects genetic alterations to phosphorylation networks. Nature. Submitted.
15. Boser BE, Guyon IM, Vapnik VN, editors. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory; 1992: ACM.
16. Genkin A, Lewis DD, Madigan D. Large-scale Bayesian logistic regression for text categorization. Technometrics. 2007;49(3):291-304.
17. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
18. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine learning. 2002;46(1-3):389-422.

19. Ray B, Henaff M, Ma S, Efstathiadis E, Peskin ER, Picone M, et al. Information content and analysis methods for Multi-Modal High-Throughput Biomedical Data. Scientific reports. 2014;4.
20. Bach FR, Lanckriet GR, Jordan MI, editors. Multiple kernel learning, conic duality, and the SMO algorithm. Proceedings of the twenty-first international conference on Machine learning; 2004: ACM.
21. Bach FR, Lanckriet GRG, Jordan MI. Fast kernel learning using sequential minimal optimization: Computer Science Division, University of California Berkeley, CA; 2004.
22. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011;2(3):27.
23. BBR: Bayesian Logistic Regression Software. Available from: www.bayesianregression.org/bbr.html
24. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18-22.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995:289-300.
26. Liu J, Page D, Peissig P, McCarty C, Onitilo AA, Trentham-Dietz A, et al. New genetic variants improve personalized breast cancer diagnosis. AMIA Summits on Translational Science Proceedings. 2014;2014:83.
27. Dietterich TG. Ensemble methods in machine learning. Multiple classifier systems: Springer; 2000. p. 1-15.
28. Hansen LK, Salamon P. Neural network ensembles. IEEE Transactions on Pattern Analysis & Machine Intelligence. 1990 (10):993-1001.
29. Cao L, Chua K, Chong W, Lee H, Gu Q. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. Neurocomputing. 2003;55(1):321-36.
30. Raymer ML, Punch WF, Goodman ED, Kuhn L, Jain AK. Dimensionality reduction using genetic algorithms. Evolutionary Computation, IEEE Transactions on. 2000;4(2):164-71.
31. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nature Reviews Genetics. 2015;16(2):85-97.