Breast Cancer Treatment Outcomes Modeling from METABRIC Genomic Patient Data

Aleksandar Obradovic[1], Carolyn Silverman[2]
[1]Columbia University, New York, NY [2]Columbia University, New York, NY

Abstract

The ability to more accurately predict outcomes of various cancer treatment regimes would be a transformative tool in risk-assessment for clinical therapeutic decision-making in response to cancer diagnosis. Our project aims to predictively model favorable outcomes in response to treatment in a dataset of Invasive Breast Carcinoma patients, where treatments post-surgery are any of a combination of chemotherapy, hormone therapy, and radiotherapy. We build separate random forest models for each therapeutic category in order to predict vital status and overall survival months (both represented as binary variables) in response to treatment. The dataset we are utilizing includes standard clinical diagnostic markers of breast cancer, as well as patient age and inferred menopausal state, which are useful attributes in a predictive model. However, the dataset also includes tumor genetic data, specifically SNP, copy number, and normalized gene expression for each patient. When we augment the clinical variables with genetic copy number and RNA-expression data for each patient, select top predictive variables, and evaluate model performance by multiple metrics, the comparative performance improves across the board. The greatest performance gain is achieved by incorporating gene expression data, although specific genes indicative of treatment response are not stable across treatment categories and do not exhibit over-representation of any particular biological process GO category. We therefore observe a useful contribution of genomic data in risk assessment modeling of breast cancer treatment regimes but not in terms of specific diagnostic gene targets.

Introduction

Since the publication of the first complete cancer genomes in 2009[1,2], it has become increasingly clear that cancer is not one single disease with a common cure. Not only do genetic differences of cancer cells exist between various types of cancer, but even a single cancer type is likely to exhibit a substantial amount of genetic variation[3]. Consequently, patients may require different therapies to treat the same disease – the idea behind "personalized medicine." Currently, Washington University's Genomics and Pathology Services provides testing for 42 gene mutations linked to cancer, which can help doctors tailor treatments[4].

Breast cancer is a prime example of a complex disease in which patients with seemingly similar symptoms and clinical features respond quite differently to the same treatment therapies. The landmark 2012 METABRIC study[5], from which we obtained our data, found breast cancer to exist as at least 10 distinct diseases, each with a different "molecular fingerprint." The researchers involved in the study were able to establish sub-categories by identifying key genetic variations responsible for turning on many cell processes associated with breast cancer (i.e. driver mutations) and mapping these variations to certain clinical features, such as high levels of the estrogen receptor. Our approach will focus on predicting favorable outcomes of breast cancer patients in response to treatment by integrating clinical and molecular data. The large size and completeness of the dataset will allow us to partition the patients based on treatment regimes to determine how genomic and expression data can supplement clinical data in predicting patient response to treatments.

Data Sources

This study utilizes the 2012 METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset, consisting of 2509 cases of Invasive Breast Carcinoma. The dataset is made available through cbioPortal, which allows for easy access to all of the clinical phenotype data, as well as patients' genetic profiles. For this project, we limit our study of personalized genomic data to copy number variation (CNV)—notably, the first ever breast cancer-specific map of CNVs—and normalized mRNA gene expression. Clinical variables include age, cellularity, estrogen receptor (ER) status, progesterone receptor (PR) status, cancer grade, Nottingham Prognostic Index (NPI), inferred menopausal state, laterality, tumor size, and tumor stage, described in Table S1. Below, we explain how we determine the most important variables and variable types to include in our predictive models and evaluate response-prediction results. The dataset also contains binary variables for the type of treatment received—chemotherapy, radiotherapy, and hormone treatment, which we use to partition into treatment categories. We consider two response variables for classification of positive treatment outcome: vital status ("living" or "died of disease"), and overall

survival months. We define positive treatment response alternately as the vital status at the end of the study, and as overall survival months greater than average, in order to account for patients who may have died after the course of the study. Patients who died of other causes and patients with unknown vital status are discounted from predictive analysis and are removed from the data set, as are patients with incomplete data coverage, leaving 1375 usable patient cases.

Preliminary analysis of the dataset showed six treatment-regime subpopulations of reasonable size for classification (between 150 and 442 patients). For each of these, we move forward with training prediction models for both response metrics. Table S2 shows the number of patients in each subpopulation and the treatment outcomes by vital status and overall survival months greater than average. Figure 1 provides a graphic representation of the survival rates and size of each treatment group.

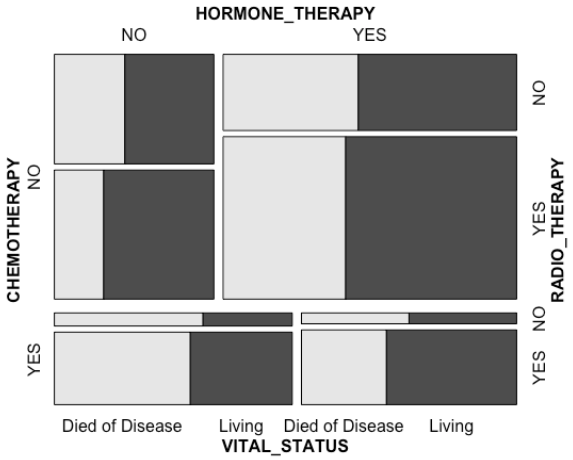

Figure 1: Plot of survival rates by vital status in each treatment subset of the data

Proposed Algorithm

We propose for each of the six sufficiently large treatment-segmented subsets of the data to build two sets of random forest classification models: one predicting favorable outcome defined by vital status at the end of the study, and one by overall survival months greater than average within that treatment regime. For each, we build a model using clinical variables alone, another using clinical variables supplemented by genomic copy number, one with clinical variables and gene expression data, and one using all three sets of variables together.

We perform feature ranking in five training/testing folds for these models by mutual information with the response variable, and average the mutual information for each feature across the five folds. We then heuristically determine the top N features in each configuration by increasing N until random forest out-of-bag-error increases due to over-fitting (see Figure 3 for an example). Finally, we evaluate which combined set of variables results in the most useful models across treatment regimes in terms of accuracy, positive predictive value, and negative predictive value under maximum likelihood classification, as well as area under the roc curve, all with five-fold cross-validation. We examine which features within these model sets were the most relevant for treatment efficacy prediction. The overall flow of the algorithm is described in Figure 2.

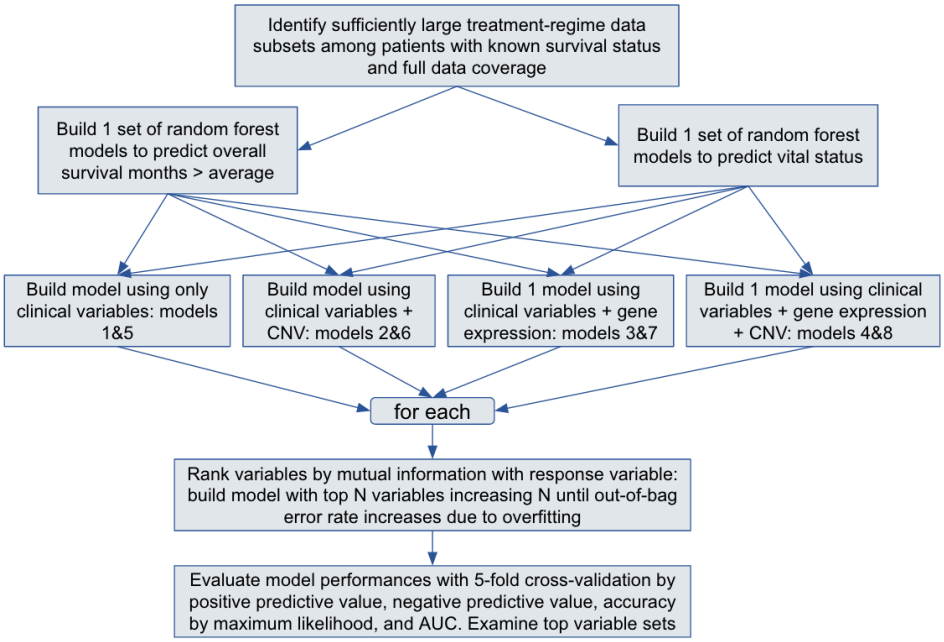

Figure 2. Algorithm flow chart, describing a total of 8 random forest models trained for each treatment regime subset. Models 1 and 5 rely only on clinical variables to predict vital status and overall survival months greater than average, respectively. Models 2 and 6 use clinical variables and CNV. Models 3 and 7 use clinical variables with gene expression data. Models 4 and 8 use all variables.

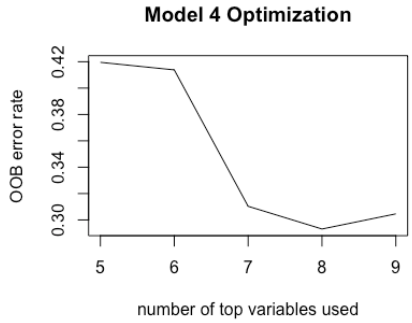

Figure 3. A plot of out-of-bag-error rate for the random forest model 4 trained in the (chemo, no hormone, radio) therapy set. Out-of-bag-error continuously decreases when incorporating the first 8 features by mutual information into the model, and then increases when the 9th feature is added. Therefore the training stops at 8 top features, used for further performance evaluation.

Results

We build models on each of the six identified treatment subsets, using clinical data and varying levels of genomic information to predict treatment outcome defined by two binary classification metrics, vital status and overall survival months greater than average. We see an improvement across the board in mean model performance when incorporating copy number and gene expression data in addition to the clinical data baseline (Figure 4). This improvement is seen by all metrics, including accuracy and area under the roc curve as a general measure of overall model performance, as well as the more targeted positive predictive value, which indicates the degree to which the model accurately predicts positive response to treatment, and negative predictive value, which indicates the degree to which the model accurately predicts negative response to treatment. Best improvement relative to clinical variable baseline is seen for the models incorporating gene expression data, and although positive predictive value is higher than negative predictive value, there is no compromising drop in negative predictive value. Overall area under the

curve (AUC) is comparatively high in the gene expression models at around 0.75, relative to baseline around 0.65, with statistical significance in some comparisons (Table 1).

When examining the top variables for these models, we note that the top variables selected for model 3 and model 4 are universally identical or near-identical, as are the top variables selected for model 7 and model 8 (Table 2). This indicates that the most useful added information for predicting treatment response was present in the gene expression data and not the CNV data. Furthermore, these top-performing variable sets often do not include many clinical variables or any at all. The most commonly recurring clinical variables, and therefore likely the most informative, are the Nottingham Prognostic Index (NPI) and tumor size, but even these are routinely outperformed by gene expression features.

Although gene expression produces an improved model performance, and individual genes selected as top features may have been implicated in cancer progression in prior studies, the top gene expression variables are divergent across treatment regimes. Even within the same treatment regime the top genes selected are almost entirely divergent between models trained on the two different response variables (Table 2). Furthermore, there is no trend of over-represented PANTHER biological processes[6] in these top gene lists, so there is no specific gene target to probe as a link to cancer treatment response and no set of biological processes that are especially implicated as treatment response drivers (Figure 5).

Figure 4. The mean performance with standard error across five folds for each model trained within each treatment set. The upper left plots show area under the roc curve, the upper right show accuracy by maximum likelihood classification, the bottom left show positive predictive value, and the bottom right show negative predictive value. Positive predictive value is higher than negative predictive value, and there is more improvement for PPV in performance of the personalized genetic models relative to baseline (2,3,4 vs 1 and 6,7,8 vs 5), but the greatest improvement is in AUC for models 7 and 8 vs model 5, predicting overall survival months with gene expression data vs without. P-values for the improvements are listed in Table 1.

Accuracy T-tests

No Chemo, No Hormone, No Radio	
Models	P-Value
M1 vs M2	0.478513
M1 vs M3	0.2741938
M1 vs M4	0.9899183
M5 vs M6	0.6793166
M5 vs M7	0.04448803
M5 vs M8	0.02406805

No Chemo, Hormone, No Radio	
Models	P-Value
M1 vs M2	0.4894488
M1 vs M3	0.09879367
M1 vs M4	0.02446095
M5 vs M6	0.03956691
M5 vs M7	0.2174603
M5 vs M8	0.07499059

No Chemo, No Hormone, Radio	
Models	P-Value
M1 vs M2	0.1294978
M1 vs M3	0.3742491
M1 vs M4	0.2487952
M5 vs M6	0.1870807
M5 vs M7	0.0375687
M5 vs M8	0.1058182

No Chemo, Hormone, Radio	
Models	P-Value
M1 vs M2	0.4535787
M1 vs M3	0.6439052
M1 vs M4	0.3057495
M5 vs M6	0.2220088
M5 vs M7	0.07021734
M5 vs M8	0.1710669

Chemo, No Hormone, Radio	
Models	P-Value
M1 vs M2	0.5957921
M1 vs M3	0.4820906
M1 vs M4	0.2981639
M5 vs M6	0.2514875
M5 vs M7	0.03724758
M5 vs M8	0.1106795

Chemo, Hormone, Radio	
Models	P-Value
M1 vs M2	0.275535
M1 vs M3	0.9075649
M1 vs M4	0.8457223
M5 vs M6	1
M5 vs M7	0.1710491
M5 vs M8	0.2432673

Positive Predictive T-tests

No Chemo, No Hormone, No Radio	
Models	P-Value
M1 vs M2	0.1673389
M1 vs M3	0.2557431
M1 vs M4	0.4927239
M5 vs M6	0.5661934
M5 vs M7	0.08673818
M5 vs M8	0.05418413

No Chemo, Hormone, No Radio	
Models	P-Value
M1 vs M2	0.5809613
M1 vs M3	0.135601
M1 vs M4	0.09068075
M5 vs M6	0.009985125
M5 vs M7	0.2148639
M5 vs M8	0.07855994

No Chemo, No Hormone, Radio	
Models	P-Value
M1 vs M2	0.2809974
M1 vs M3	0.2077093
M1 vs M4	0.452376
M5 vs M6	0.5302726
M5 vs M7	0.03992262
M5 vs M8	0.2118689

No Chemo, Hormone, Radio	
Models	P-Value
M1 vs M2	0.6981967
M1 vs M3	0.5348362
M1 vs M4	0.1160908
M5 vs M6	0.2678359
M5 vs M7	0.06452643
M5 vs M8	0.1814283

Chemo, No Hormone, Radio	
Models	P-Value
M1 vs M2	0.9205452
M1 vs M3	0.2110895
M1 vs M4	0.1639234
M5 vs M6	0.1854158
M5 vs M7	0.03489473
M5 vs M8	0.1500748

Chemo, Hormone, Radio	
Models	P-Value
M1 vs M2	0.4060488
M1 vs M3	0.8703398
M1 vs M4	0.6518554
M5 vs M6	0.4752915
M5 vs M7	0.148668
M5 vs M8	0.2677359

Negative Predictive T-tests

No Chemo, No Hormone, No Radio	
Models	P-Value
M1 vs M2	0.7673451
M1 vs M3	0.3988461
M1 vs M4	0.6438287
M5 vs M6	0.1088352
M5 vs M7	0.03748736
M5 vs M8	0.01755876

No Chemo, Hormone, No Radio	
Models	P-Value
M1 vs M2	0.362713
M1 vs M3	0.1178536
M1 vs M4	0.01445173
M5 vs M6	0.4309919
M5 vs M7	0.3289958
M5 vs M8	0.06875703

No Chemo, No Hormone, Radio	
Models	P-Value
M1 vs M2	0.1576195
M1 vs M3	0.5561789
M1 vs M4	0.3170737
M5 vs M6	0.07297795
M5 vs M7	0.04244762
M5 vs M8	0.05482166

No Chemo, Hormone, Radio	
Models	P-Value
M1 vs M2	0.4493746
M1 vs M3	0.8623454
M1 vs M4	0.612334
M5 vs M6	0.106554
M5 vs M7	0.06639298
M5 vs M8	0.2002502

Chemo, No Hormone, Radio	
Models	P-Value
M1 vs M2	0.2336788
M1 vs M3	0.7738357
M1 vs M4	0.530582
M5 vs M6	0.8356422
M5 vs M7	0.06551687
M5 vs M8	0.2270437

Chemo, Hormone, Radio	
Models	P-Value
M1 vs M2	0.2230425
M1 vs M3	0.7652534
M1 vs M4	0.6851914
M5 vs M6	0.2941189
M5 vs M7	0.7897161
M5 vs M8	0.5163227

AUC T-tests

No Chemo, No Hormone, No Radio	
Models	P-Value
M1 vs M2	0.5338549
M1 vs M3	0.1186961
M1 vs M4	0.445543
M5 vs M6	0.9650166
M5 vs M7	0.04007806
M5 vs M8	0.03070078

No Chemo, Hormone, No Radio	
Models	P-Value
M1 vs M2	0.5566943
M1 vs M3	0.05142955
M1 vs M4	0.02426606
M5 vs M6	0.1085977
M5 vs M7	0.08276847
M5 vs M8	0.07208823

No Chemo, No Hormone, Radio	
Models	P-Value
M1 vs M2	0.1062091
M1 vs M3	0.02197213
M1 vs M4	0.3153895
M5 vs M6	0.03679609
M5 vs M7	0.01369019
M5 vs M8	0.02329823

No Chemo, Hormone, Radio	
Models	P-Value
M1 vs M2	0.575567
M1 vs M3	0.4721013
M1 vs M4	0.468723
M5 vs M6	0.01662181
M5 vs M7	0.004311021
M5 vs M8	0.01238407

Chemo, No Hormone, Radio	
Models	P-Value
M1 vs M2	0.9793804
M1 vs M3	0.2719921
M1 vs M4	0.1573064
M5 vs M6	0.6957178
M5 vs M7	0.01934901
M5 vs M8	0.04586046

Chemo, Hormone, Radio	
Models	P-Value
M1 vs M2	0.9310573
M1 vs M3	0.6890556
M1 vs M4	0.5379708
M5 vs M6	0.5835549
M5 vs M7	0.03467002
M5 vs M8	0.06134346

Table 1. This table shows p-values for the model 1,2,3 vs 4 and 6,7,8 vs 5 performance comparisons in Figure 4. P-values are derived by 2-sample t-test, since performances across the five folds are normally distributed around the mean. Significant p-values ($p<0.05$) are highlighted in yellow. Significant improvement of experimental model vs clinical baseline is seen most often for the AUC of models 7 and 8 compared to model 5, predicting overall survival months with gene expression data vs without.

No Chemo, No Hormone, No Radio							
Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
TUMOR_SIZE	CORT	CD1C	CORT	NPI	NPI	RAMP3	RAMP3
NPI	DFFA	KIF14	DFFA	TUMOR_SIZE	CYP4F3	ABCB1	ABCB1
AGE_AT_DIAGNOSIS	PGD	MKI67	PGD	AGE_AT_DIAGNOSIS	CYP4F8	CD1C	CD1C
GRADE	UBE4B	HOXC10	UBE4B	GRADE	DNAJB1	LAT2	LAT2
HISTOLOGICAL_SUBTYPE	NOV	AURKA	NOV	ER_STATUS	NDUFB7	KLK1	KLK1
PR_STATUS	PIK3CD	PPP1CB	PIK3CD	HISTOLOGICAL_SUBTYPE	PKN1	NPI	NPI
ER_STATUS	ENPP2	EZR				INHBA	INHBA
BREAST_SURGERY	MTOR					GNRH1	GNRH1
	EXOSC10						

No Chemo, Hormone, No Radio							
Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
NPI	NPI	NPI	NPI	AGE_AT_DIAGNOSIS	AGE_AT_DIAGNOSIS	AGE_AT_DIAGNOSIS	AGE_AT_DIAGNOSIS
TUMOR_SIZE	TUMOR_SIZE	TUMOR_SIZE	TUMOR_SIZE	INFERRED_MENOPAUSAL_STATE	PRKAR1B	BIRC2	BIRC2
AGE_AT_DIAGNOSIS	CCNB1	CDK3	CDK3	NPI	ZNF217	UBE2L3	UBE2L3
GRADE	FADD	EZR	EZR	HISTOLOGICAL_SUBTYPE	PDGFA	OLFM4	OLFM4
LATERALITY	CDK7	RECQL5	RECQL5	CELLULARITY	NUDT1	CSH2	CSH2
BREAST_SURGERY	FLI1	HMBS	HMBS	PR_STATUS	GPER1	MYBPC1	PRKAR1B
						ZNF185	ZNF217
						ROBO2	
						STK25	

No Chemo, No Hormone, Radio							
Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
BREAST_SURGERY	CPT2	PDHA1	CPT2	LATERALITY	MEF2A	SORBS1	SORBS1
AGE_AT_DIAGNOSIS	MAGOH	GADD45A	MAGOH	NPI	HSP90B2P	PDIA4	MEF2A
TUMOR_SIZE	LRP8	PTGER3	LRP8	GRADE	NR2F2	STIP1	HSP90B2P
PR_STATUS	C8B	ASAH1	C8B	PR_STATUS	GPR37	LRP2	NR2F2
NPI	C8A	TLR3	C8A	TUMOR_SIZE	IGF1R	CCDC6	PDIA4
INFERRED_MENOPAUSAL_STATE	CYP2J2	SRSF4	CYP2J2	AGE_AT_DIAGNOSIS	ABHD2	AMHR2	STIP1
GRADE							GPR37
HISTOLOGICAL_SUBTYPE							

No Chemo, Hormone, Radio							
Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
NPI	NPI	NPI	NPI	AGE_AT_DIAGNOSIS	AGE_AT_DIAGNOSIS	AGE_AT_DIAGNOSIS	AGE_AT_DIAGNOSIS
TUMOR_SIZE	TUMOR_SIZE	FOXM1	FOXM1	TUMOR_SIZE	SYT4	AGFG1	AGFG1
AGE_AT_DIAGNOSIS	AGE_AT_DIAGNOSIS	E2F2	E2F2	INFERRED_MENOPAUSAL_STATE	RIT2	OR7E14P	OR7E14P
BREAST_SURGERY	BCL2	AURKB	AURKB	PR_STATUS	SLC14A1	SLC25A44	SLC25A44
GRADE	BREAST_SURGERY	KIFC1	KIFC1	NPI	ZNF24	ZNF214	ZNF214
INFERRED_MENOPAUSAL_STATE	DSG1	TUMOR_SIZE	TUMOR_SIZE	BREAST_SURGERY	PSTPIP2	CD68	CD68
PR_STATUS	LDLRAD4						SYT4
	CDH7						

Chemo, No Hormone, Radio							
Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
BREAST_SURGERY	NEURL1	AQP4	AQP4	TUMOR_SIZE	EYA2	GPR17	GPR17
TUMOR_SIZE	SH3PXD2A	PRKCZ	PRKCZ	BREAST_SURGERY	LILRB1	KCNH1	EYA2
NPI	CREBBP	SEPW1	SEPW1	INFERRED_MENOPAUSAL_STATE	LILRB4	MAPKAPK2	KCNH1
GRADE	IGFBP1	CIRBP	CIRBP	GRADE	SPAST	INPP5A	LILRB1
PR_STATUS	IGFBP3	PCGF2	NEURL1	AGE_AT_DIAGNOSIS	LAIR2	EPHB3	LILRB4
ER_STATUS	MEFV	CHST4	PCGF2	ER_STATUS	LILRA1	MSX2	SPAST
		GPX4	CHST4				LAIR2
		PALM	GPX4				
		KCNH1					

Chemo, Hormone, Radio							
Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
NPI	NPI	MARS	MARS	GRADE	CYP2E1	RPS14	RPS14
GRADE	HSD17B3	PRC1	PRC1	NPI	ADAM8	PSMD2	PSMD2
AGE_AT_DIAGNOSIS	ETV6	GUCY1B2	GUCY1B2	LATERALITY	ECHS1	DIAPH1	DIAPH1
ER_STATUS	GNAQ	FEN1	FEN1	HISTOLOGICAL_SUBTYPE	TUBGCP2	CSTF2	CSTF2
LATERALITY	CDC14B	PYCR1	PYCR1	AGE_AT_DIAGNOSIS	DPYSL4	SDS	SDS
PR_STATUS	SLC35D2	E2F2	E2F2	ER_STATUS	UTF1	SLC25A11	SLC25A11
HISTOLOGICAL_SUBTYPE	FANCC					TYMS	TYMS

Table 2. Top variables used for best-performing models in each treatment subset, as selected by the method shown in Figure 3. Between 6 and 9 top variables are used in each model, making them comparable by number of variables. Model performances in five-fold cross-validation can be seen in Figure 4.

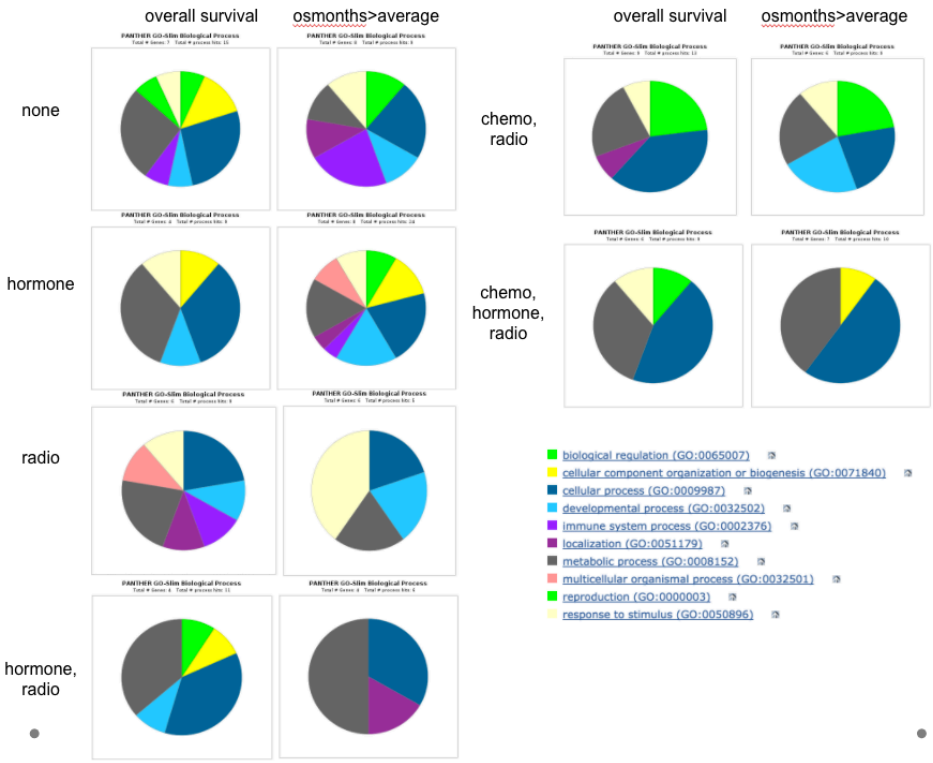

Figure 5. PANTHER biological process categories associated with the top genes selected in training model 3 and model 7 (Table 2). There are a lot of genes associated with metabolic processes (grey), and cellular processes (blue), but these are not over-represented with statistical significance.

Conclusion

The goal of this project was to take advantage of the 2012 METABRIC Breast Cancer genomic phenotyping study in order to predict breast cancer treatment responses by integrating several types of personalized genetic data. Ultimately, tumor genetic data combined with clinical history and demographic information should be able to help practitioners make better-informed decisions about an appropriate and effective treatment regime for a particular patient, as improvement in predictive model performance was seen across the board when incorporating genetic data as a supplement to clinical data. mRNA gene expression data has shown to be more useful in this regard than CNV, capturing all of the most informative features in the top models even when used in conjunction with CNV. The better performance improvement when predicting overall survival months vs vital status is also encouraging, as overall survival months seems like a more informative response variable in this study, where patients may have died following data collection or been diagnosed at different initial time points. Unfortunately, the top gene expression variable set used in the models (Table 2) is difficult to biologically validate and not very interpretable in terms of novel gene targets for clinical tracking of cancer progression. The gene expression data overall still provided a valuable improvement in model performance, which may translate to a potential for better-informed clinical decision-making in the course of genetically personalized breast cancer treatment.

Since we were unable to select an informative gene set as is, future work might further improve model performance with clinical and gene expression data by performing Principal Component Analysis to transform the feature set. This will obscure the driving features from the original untransformed data, but those are already unstable, so the reduced auto-correlation of variables may be a worthwhile tradeoff. We may also modify the core classification algorithm to a regularized logistic regression or other alternative, simply to further confirm the observations from this set of experiments. If given additional time to explore the dataset, we might take advantage of the cancer phenotypes derived in the 2012 METABRIC paper[5] as a potential feature set, but these would need to be re-derived using only secondary genomic features to avoid incorporating treatment-regime information in the phenotype features, as was done by Curtis et al. This is also less desirable for our treatment response task, since it requires incorporation of every level of genomic information from the study—CNV, gene expression, and SNP—which is more expensive diagnostically than any one of those alone.

References

1. Pleasance, Erin D., et al. "A comprehensive catalogue of somatic mutations from a human cancer genome." *Nature* 463.7278 (2010): 191-196.
2. Pleasance, Erin D., et al. "A small-cell lung cancer genome with complex signatures of tobacco exposure." *Nature* 463.7278 (2010): 184-190.
3. Shah, Sohrab P., et al. "The clonal and mutational evolution spectrum of primary triple-negative breast cancers." *Nature* 486.7403 (2012): 395-399.
4. "Sequencing Tests - Genomics and Pathology Services." *Genomics and Pathology Services*. Web. https://gps.wustl.edu/patient-care/sequencing-tests/.
5. Curtis, Christina, et al. "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups." *Nature* 486.7403 (2012): 346-352.
6. Mi, Huaiyu, et al. "Large-scale gene function analysis with the PANTHER classification system." *Nature Protocols* 1754.2189 (2013): 1551—1566

Treatment Variables	
Chemotherapy	Binary: 1094 NO, 398 YES
Radiotherapy	Binary: 549 NO, 934 YES
Hormone Treatment	Binary: 595 NO, 888 YES
Clinical Variables	
Age at diagnosis	Min. 1st Qu. Median Mean 3rd Qu. Max. 21.93 49.22 58.56 58.23 67.61 96.29
Breast surgery type	623 breast conserving, 842 mastectomy
cellularity	726 high, 160 low, 548 moderate
ER status	406 negative, 1077 positive
PR status	746 negative, 737 positive
Cancer grade	118 one, 558 two, 758 three
NPI- Nottingham Prognostic Index	Min. 1st Qu. Median Mean 3rd Qu. Max. 1.000 3.050 4.044 4.103 5.050 6.360
Inferred menopausal state	1080 post, 403 pre
Laterality	722 L, 679 R
Tumor size	Min. 1st Qu. Median Mean 3rd Qu. Max. 0.00 17.00 22.00 26.25 30.00 182.00
Tumor stage	10 zero, 386 one, 625 two, 99 three, 10 four
mRNA expression variables	z-scores available at gene level for 7,281 genes
CNV state variables	Inferred copy number count available at gene level for 7,782 genes
Response Variables	
Overall Survival Months	Min. 1st Qu. Median Mean 3rd Qu. Max. 0.00 55.68 113.70 123.60 185.50 337.00
Vital Status	Binary: 837 Living, 646 Died of Disease

Table S1: Experimental variables extracted from the Data Set for patients with known vital status (excluding "NA" and "Died of other Causes")

Treatment Regime	Number of patients	patients with vital status == “Living”	patients with overall survival months > average
hormone, chemo, radiation	150	86	79
hormone, chemo, no radiation	27	10	11
hormone, no chemo, radiation	442	259	279
hormone, no chemo, no radiation	269	144	185
no hormone, chemo, radiation	173	76	67
no hormone, chemo, no radiation	46	17	17
no hormone, no chemo, radiation	166	123	124
no hormone, no chemo, no radiation	207	122	155

Table S2: This table shows the number of patients in each treatment category defined on the dataset, including only patients where treatment labels are known and vital status is either “Living” or “Died of Disease”. It also shows the number of patients positively responding to treatment according to both metrics described above. The sizes of most treatment regime sets and positive response rates within them are large enough to perform practical classification model training and testing, and so we move forward with the project on all treatment categories except (hormone, chemo, no radiation) and (no hormone, chemo, no radiation).