AN INTEGRATIVE DATA-DRIVEN APPROACH TO IDENTIFY MOLECULAR PATTERNS IN BREAST CANCER PATIENTS

by
Isis Yanina Narvaez Bandera

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
in
INDUSTRIAL ENGINEERING

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2017

Approved by:

_____	_____
Mauricio Cabrera, PhD. Member, Graduate Committee	Date
_____	_____
Maribella Domenech, PhD. Member, Graduate Committee	Date
_____	_____
Wandaliz Torres-García, PhD. President, Graduate Committee	Date
_____	_____
Daniel Rodríguez Román Representative of Graduate Studies	Date
_____	_____
Viviana Cesaní, PhD. Chairperson of the Department	Date

ABSTRACT

Breast cancer is a heterogeneous disease of the genome in need of better diagnostics and treatments through the characterization of genomic patterns and interactions. Currently, its molecular understanding is still insufficient even with the advances in genomic technologies. Therefore, this thesis presents a multi-stage data mining approach to discriminate breast cancer subtypes through the integration of highly dimensional data from different genomic platforms using feature selection and classification techniques. This methodology allowed us to extract patterns that play a critical role in the classification of breast cancer subtypes (*i.e.* the underexpression of FOXA1 for basal). Furthermore, this thesis provides a new metric capable to assess and rank interactions between relevant features using a prevalence criteria and Random Forest classifier. This metric identified a ranked list of variable interactions to discriminate subtypes. Among those, we found a set of correlated genes frequently interacting with FOXA1 or MLPH like CEP55 and UBET2.

RESUMEN

El cáncer de mama es una enfermedad heterogénea del genoma que necesita mejores diagnósticos y tratamientos a través de la caracterización de patrones genómicos e interacciones. Actualmente, su comprensión molecular es aún insuficiente incluso con los avances en las tecnologías genómicas. Por lo tanto, esta tesis presenta un enfoque de minería de datos en varias etapas para discriminar los subtipos de cáncer de mama a través de la integración de datos altamente dimensionales de diferentes plataformas genómicas utilizando técnicas de selección y clasificación de características. Esta metodología nos permitió extraer patrones que desempeñan un papel crítico en la clasificación de los subtipos de cáncer de mama (es decir, la subexpresión de FOXA1 para basal). Además, esta tesis proporciona una nueva métrica capaz de evaluar y clasificar las interacciones entre las características pertinentes utilizando un criterio de prevalencia y el clasificador Random Forest. Esta métrica identificó una lista de interacciones de variables importantes para discriminar subtipos. Entre las principales interacciones, encontramos un conjunto de genes correlacionados interactuando frecuentemente con FOXA1 o MLPH tales como CEP55 y UBET2.

To God, my mom, my husband and my daughter:

You are my inspiration. I love you so much.

ACKNOWLEDGEMENTS

First, I want to give my special thanks to my advisor, Dr. Wandaliz Torres-Garcia. Thanks for your guidance, support and patience. Thanks for help me grow in so many aspects and help me to become a better researcher. You are, by far, one the best person I have ever met. Also, I want to thanks the members of my graduate committee: Dr. Mauricio Cabrera and Dr. Maribella Domenech for their advice and assistance during the execution of this thesis.

Thanks to the department of Industrial Engineering, their graduated students, professors and administrative personnel, especially to Dr. Betzabe Rodiguez, for believe in me and for always ensuring the well-being of my family. Thanks, to my friends: Heizel Rosado, Cesar Salazar, Miguel Ruiz and Jean Faucher: each of you, have supported me so much, in so many possible ways.

Finally, Thanks to my family, my loving mother, my dear husband and my sweety daugther. Said Cifuentes, you are the best friend and life partner that God could give me. This goal was possible thanks to all your support. This achievement is for you and for our beautiful Sofía.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY OF TERMS

AUC	Area Under the Curve
CFS	Correlation-Based Feature Selection
CV	Cross Validation
FAST	Fast Clustering-Based Feature Selection Algorithm
GEO	Gene Expression Omnibus
GSEA	Gene Set Enrichment Analysis
IG	Information Gain
impSeq	Sequential imputation of missing values
IRMI	Iterative Robust Model-based Imputation
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	K- Nearest Neighbors
MDA	Mean Decrease Accuracy
MDI	Mean Decrease Impurity
MDR	Multifactor Dimensionality Reduction
MDS	Multi-Dimensional Scaling
mRNA	Messenger Ribonucleic acid
NNs	Neural Networks
OOB	Out-Of-Bag
PAM50	Prediction Analysis of Microarray 50
PDP	Partial Dependence Plot
RNAseq	Ribonucleic Acid Sequence
RFs	Random Forest
RPPA	Reverse Phase Protein Array
SVM	Support Vector Machine
SVM-RFE	Support Vector Machine based on Recursive Feature Elimination
TCGA	The Cancer Genome Atlas
VIM	Variable Important Measures

1 INTRODUCTION

1.1 Justification

After skin cancers, breast cancer is the most common cancer among American women and it's the most common cancer leading to death among Puerto Rican women during 2006 to 2010 [1]. Genetic diversity for this disease is described in four (4) breast cancer subtypes: luminal A, luminal B, HER2 enriched and basal-like which have been grouped by mRNA profiling [2]. Also, each subtype is different both in their immunohistochemical characteristics and their gene mutation profile. For instance, basal subtype is mostly characterized with BRCA1 mutations and the absence of progesterone receptors (PR-), while luminal A is defined as PR positive [3], [4]. Molecular differences across breast cancer subtypes makes that each patient responds differently to clinical treatments. Currently, the selection of clinical treatments is defined by the expression of HER2 and hormone receptors [5]. Recent work proposes that gene mutations have a significant effect on the treatments outcome, however these mutations have not been fully characterized [6], [7]. Hence specific genomic characterization within each breast cancer subtype is important to possibly reveal clinical impact that can allow personalized treatment.

Due to the technology advances in the area of cancer genomics we have access to an explosion of large-scale biological datasets to analyze it. Data mining and knowledge discovery methodologies allows for the extraction of implicit information from large

amounts of data using mathematical and statistical methods. Also, added computer advances have enabled knowledge discovery for high dimensional databases.

Several cancer projects such as The Cancer Genome Atlas (TCGA) can provide a wealth of information to better understand cancer biology mechanisms when critical computational challenges are addressed. There is plenty of literature work on microarrays to detect biomarkers oncogenes and tumor suppressors but most are performed using data from similar platform technologies. Therefore, we propose to integrate heterogeneous data types to determine association with breast cancer subtypes and extract interactive patterns. It is increasingly evident that gene interactions play a fundamental role in the proneness to diseases [8]. However, finding gene-gene interactions is a hard problem because of the dimensionality problem, noise in the data, complexity of the systems and experimental protocols. Currently, there are no many methods that can do this efficiently [9].

Consequentely, we aim to find interactive patterns through the integration of diverse datasets in a computationally feasible manner as well as to present a new metric to assess the importance of interacting variables implemented using Random Forest. We have chosen to develop the new metric through the use of an ensemble Random Forest model based on the numerous advantages of this methodology such as mixed type variable integration, nonlinear interactions detection, variable importance estimates and its overall good performance in practice. Results from the multi-step integration process may discover important genes, proteins, methylation regions, clinical factors and their interactions that distinguish breast cancer subtypes well. Hence, the main contribution of

this work consisting of data integration and exploration of a new metric to assess interaction importance lies on the possible biological discoveries that can deepen the current understanding of breast cancer subtypes for clinical treatment and patient outcome improvement.

1.2 Objectives

The main objective of this work is to investigate whether important genes, proteins, methylations, clinical factors or their interactions can have the potential to discriminate among breast cancer patients and their subtypes, through the use of data mining techniques. Also, an implementation of a new algorithm capable of assess and rank the interaction between relevant features, using prevalence criteria and Random Forest classifier. The findings of this work could yield new knowledge necessary to further personalize clinical diagnosis that can later impact clinical treatment and prognosis.

1.2.1 Specific Objectives

- To identify molecular patterns that can discriminate among breast cancer patients and their subtypes by applying feature selection and classifications methods using gene expression (microarray and RNA-sequence platforms) and clinical data.
- To integrate protein, methylation and gene expression towards breast cancer subtype classification.
- To develop a new metric and implement an algorithm capable of assessing the interaction between relevant features resulting from the integration phase using Random Forest method.

1.3 Scope and General Organization of the Thesis

The scope of this thesis encompasses the use of different feature selection methods in order to extract biologically important patterns with relevance at the phenotypic level specifically breast cancer subtype identification. Firstly, the literature review pertaining to all three objectives was discussed in Chapter 2. Then, two stages of integration of highly heterogeneous datasets were evaluated: the first using gene expression and clinical factors (Chapter 3), and the second using gene expression, protein and methylation data (Chapter 4). In addition, as part of a third stage, we develop a new metric capable of measuring interactions' importance between pairs of features implemented using Random Forest (Chapter 5). Finally, general remarks on the overall results in this thesis were summarized (Chapter 6).

2 LITERATURE REVIEW

Data mining approaches have clever implementations to analyze large dataset from different perspectives making it a popular analytic strategy within the research community. Many knowledge discovery models consists of data selection, data transformation, data preprocessing (cleaning and reduction), data integration, data mining and data interpretation which we intend to explore in this thesis work.

Therefore in biomedical data analysis, data selection from appropriate databases, such as 1000 Genome, ENCODE and TCGA, is the first stage. These project repositories allowed for implementations of data mining tools that can extract information and ultimately useful knowledge, to better understand cancer mechanisms. Among the advantages of these repositories are their robustness and high quality of experimental protocols that aim to reduce external noise factors. Usually molecular data used for research undergoes a data transformation before publication (i.e. scaling, normalization), consisting in syntactic modifications on the data without change of its meaning. Generally biomedical data have inherent characteristics such as: large dimensionality, small sample sizes [10], class imbalanced [11], [12] and high complexity [13] that pose as challenges when training classification models.

Data preprocessing is the process to clear data, impute missing and reduction data, using for example, feature selection techniques. Given the high dimensionality of the data available from the cancer genome repositories, implementation of feature selection is

necessary to allow for computational feasibility. Hence in Section 2.1 we review feature selection methodologies that could accomplish the task at hand.

Due to the large number of omics technologies available from heterogeneous data type it is imperative to overcome data integration challenges to allow for discovery of new biological knowledge from improved models that include high-degree interactions. In Section 2.3, a review of data integration efforts for classification using omics data is explored and we investigate Random Forest algorithm to account for interaction among predictors.

Once data has been acquired and preprocessed, then the essential step of knowledge discovery is performed using intelligent classification, clustering, or similarity search algorithms to mine the data and extract knowledge from it. Data mining methods can find patterns over the data and as in biomedical datasets these patterns are believed to have interactions effects with importance in the proneness of diseases especially cancer. In Section 2.4 we focus on the most common used metrics to assess degree of interaction between variables.

Ultimately, the discovered knowledge is presented to the end user with insights gathered through functional analysis and helpful visualization output for easier interpretation. This allows for data interpretation in which the user understands the discovered knowledge obtained by the data mining techniques. These biomedical data are interpreted with databases available such as Gene Set Enrichment Analysis (GSEA) [14] and cBio portal [15] which provide another level of functional and known pathway inner workings to evaluate the importance of the revealed molecular patterns.

2.1 Feature Selection

Feature selection is a useful technique to reduce the computational burden of exploring the effect of predictors (or features) and to ease the interpretability by researchers. Feature selection methods aim to obtain the most relevant predictors, where relevancy is defined based on their effect on improving the prediction model [16]. There are three common classes for feature selection methods: filter, wrapper and embedded methods. Also, other more sophisticated feature selection techniques beyond these three types were studied.

2.1.1 Filters

Traditionally, filters methods are the most employed gene selection approach for its speed and computational simplicity. These assess the goodness of variables by considering only the intrinsic data properties of single genes with its corresponding class label. Then, a score for each variable (i.e. genes) is assigned based on statistics metrics measuring the general behavior between samples in the training set. There are different measures of association such as Euclidean distance, information gain or probabilistic. Some of the most used filters methods in omics data are: Correlation-based Feature Selection (CFS), Information Gain (IG), and ReliefF [17]–[21]. Overviews of these specific techniques are shown below as these are investigated further in the proposed work.

Correlation-based Feature Selection (CFS) evaluates a subset of features depending to a correlation measure with the class according to a heuristic search strategy such as greedy

hill climbing and best first [22]. It is expected that selected subsets are only contained by variables that satisfy the following two characteristics: first, to be highly predictive for the class and second to not have correlation between them. Hall evaluated the score of a subset as shown in Equation 2-1.

$$M_s = \frac{K\overline{r_{zi}}}{\sqrt{k + k\,(k-1)}r_{ii}}, \qquad \text{Equation 2-1}$$

Hall defined M_S is the score of a feature subset S containing k features. Let z be the class response then $\overline{r_{zi}}$ is the average feature to class correlation ($i \in S$), and r_{ii} is feature-feature correlation [22]. The difference of this method with other filter methods is that CFS reports the best subset found according to the scores of different feature subsets which are generated heuristically through an optimization method. In this thesis work, we focused on implementing the best first search optimization.

Information Gain (IG) is the expected reduction in entropy on the groupings formed by the attribute values. It gives an ordered ranking for all the features comparing the entropy before and after one split and a threshold is needed for choosing those variables with highest information gain [23]. Information gain is given by Equation 2-2, where *H* specifies the entropy function, *X* is the predictors' matrix and *Y* is the response of interest.

$$InfoGain = H(Y) - H(Y|X) \qquad \text{Equation 2-2}$$

Entropy is the measure of uncertainty associated with a probability distribution. The entropy of $H(Y)$ set is defined by Equation 2-3:

$$H(Y) = -\sum_{y} p(y)log_2(p(y)) \qquad \text{Equation 2-3}$$

$H(Y|X)$ is the conditional entropy of *Y*, where *X* is known and is defined by Equation 2-4:

$$H(Y|X) = -\sum_{X} p(x) \sum_{y} p(y|x)\, log_2(p(y|x)) \qquad \text{Equation 2-4}$$

Where $p(y)$ is the probability to select class y and $log_2(p(y))$ is the information associated to this class. Similarly for $p(x)$ is the probability to select class x and $log_2(p(y|x))$ is the information given this class.

ReliefF is another recognized filter method, a feature weighting algorithm extended from the original Relief method that enables multi-class problems and deals better with noisy and incomplete data [24]. This method intends to group samples from the same class while distinguishing between different ones as shown in **Figure 2-1**. Kononenko *et al*. stated the steps of this algorithm as follows: (1) a sample (*R*) is randomly chosen and the (*K*) nearest neighbors from its same class (*H*) and from different classes (*M*) are selected; (2) the differences between the instance *R* and the instances *H* and *M* are calculated with respect to the values of each feature *A* (i.e. $diff(A, R, H\ or\ M)$) to estimate their scores (*W(A)*) using Equation 2-5; lastly, (3) to improve the score estimate steps (1) and (2) are repeated *m* times [25]. For the problem of multiple classes these differences metrics are calculated by conditioning on the probabilities of each class.

$$W(A) = W(A) - \frac{diff(A, R, H)}{m} + \frac{diff(A, R, M)}{m} \qquad \text{Equation 2-5}$$

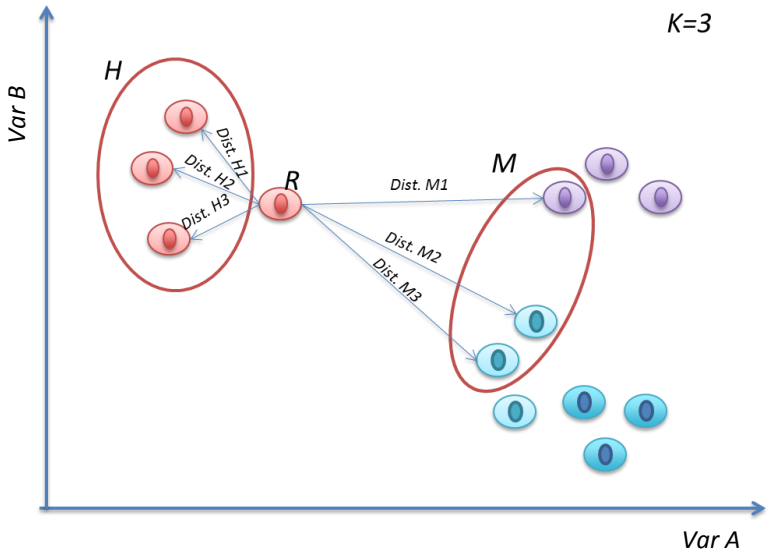


Figure 2-1. Example of ReliefF. Illustration for a reliefF feature selection problem of multiple classes using $k = 3$ as an example.

Among the benefits of filter methods are that they are fast and computationally simple, also they can scale data with hundreds of thousand variables very well. However the filters methods use a univariate procedure, which is one of its major drawbacks, because interactions or dependencies are not captured. Though, CFS does capture dependencies to small degree with their optimization logic to select the subtypes. While filters methods first looks for a good subset of features and then do the model selection, wrapper methods integrates both steps.

2.1.2 Wrappers

In the ***Wrapper*** approaches [26], every possible subset of candidate variables is evaluated through an classification algorithm, from the subset that contains the selected variables and the variable that determines the sample class. This evaluation can be performed based on specific classification model. To define the space of all subsets, the classification

model and the search algorithm work simultaneously. Compared to filter methods, this type of approach is much more demanding computationally and has higher risk of overfitting (performs too well in training data, but fails in test data). Consequently, the wrapper approach is not commonly used with microarray datasets [27].

2.1.3 Embedded

Other classical feature selection method are ***Embedded*** techniques. This type of technique differs from other methods in the way feature selection and learning are implemented iteratively as these two parts cannot be separated. The search for the optimal subset of variables is done in the process of constructing the classifier system. This technique is bound to a specific learning algorithm similar to wrappers. However, it has the advantage of including the interaction with the classification model while requiring fewer computational resources; therefore several proposals have emerged in recent years for microarray data classification.

One common embedded method is ***Support Vector Machine based on Recursive Feature Elimination (SVM-RFE)***, proposed in [28]. The algorithm begins with all variables and removes one at a time iteratively. The removed variable is the least important according to its weight on the SVM classifier. Then feature subsets are selected through this backward elimination methodology. SVM-RFE can easily deal with large variables and a small number of samples that generally are intrinsic characteristics of biological data, which is why it has become very attractive in the use of this data type [29]–[32].

2.1.4 Other Sophisticated Methods

On the other hand, since molecular pathway interactions are thought to have a prominent role in the susceptibility to cancer, clustering based methods for microarray data have been recently proposed to capture these interactions [27]. Beyond the traditional filter, wrapper and embedded feature selection methods there has been the need to reduce predictors using more complex methodologies such as ***fast clustering-based feature selection algorithm (FAST)*** [33]. FAST works in two steps: (i) irrelevant features are removed and simple clusters of correlated variables are created, and (ii) the most important features of each cluster are selected to form create a subset of relevant features. Is expected that these subsets are formed by independent variables. Qinbao *et al*, demonstrated that FAST is efficient in several publicly available data. This method outperformed four commonly known classifiers (i.e. Naïve Bayes, C4.5, instance-based lazy learning algorithm, RIPPER) when using microarray data based on classification accuracy.

2.2 Feature Selection Assessment through Classification Methods

In this work five feature selection methods were evaluated. The selected methods (CFS, Information Gain, ReliefF, SVM-RFE and FAST Clustering based) were selected based on their ability to work with large data dimensionality. To assess the performance of feature selection techniques, we will use AUC and error rate measures through the application of three well-known classification methods: K- Nearest Neighbors (KNN)

[34], Support Vector Machines (SVM) [35] and Random Forest (RFs) [36]. A brief description of these classifiers is presented next:

2.2.1 K- Nearest Neighbors (KNN)

KNN is an algorithm that classifies a new instance as the class of the 'k' nearest neighbors by taking into account its similarity matrix calculated from a distance measure (See **Figure 2-2**). Among the parameters required to implement this classifier are: fixed "k" nearest neighbors and the type of distance to calculate (i.e. Euclidean, Manhattan, Mahalanobis).

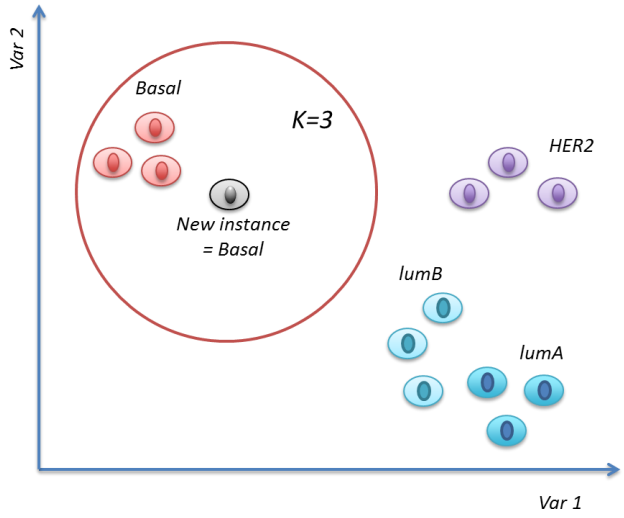


Figure 2-2. Example of KNN. Illustration for a KNN classifiers problem of multiple classes using $k = 3$ as an example. The new instance (grey color) will be classified as basal subtype, since the most closely samples are of this subtype.

2.2.2 Support Vector Machines (SVM)

Another well-known classifier model in the field is SVM. Essentially, this type of algorithm uses a kernel function to map the original data into higher dimensional spaces. This allows for a binary linear separation between classes when using an appropriate

kernel. Once the data has been mapped through the kernel trick, a two-class classification is performed by detecting two support vectors which transformed features from each class. These are selected by maximizing the distance between the support vectors (i.e. margin) (See **Figure 2-3**). To perform the classification, it is required that the user defines two parameters: the C regularization constant and the specific kernel function.

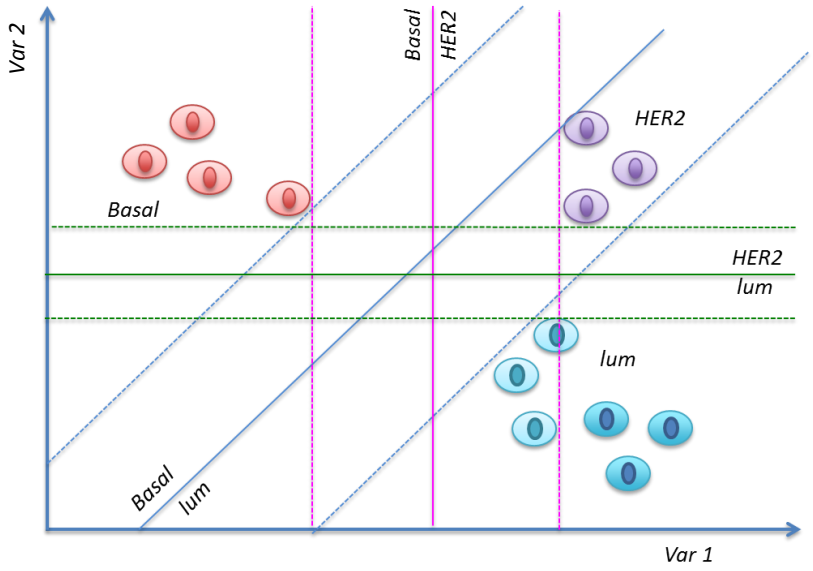


Figure 2-3. Example of SVM. Illustration for a SVM classifiers problem of three classes with one-against-one approach. The blue, green and purple solid lines are the boundaries between Basal-luminal, Her2-luminal and Basal-Her2 subtypes respectively.

2.2.3 Random Forest (RFs)

Lastly, we reviewed the classification performance of RFs, an ensemble method. This algorithm uses a group of individual classification trees, where each tree has been trained using a bootstrap sample of the data. During the construction of each tree, a random set of input features are selected and considered as candidate variables at each split of the tree.

The class of a sample is assigned according to the aggregate votes obtained by each tree. The tree construction is according to the following steps:

i. Randomly sample *n* instances with replacement to form the training set containing about two thirds of the original data. Samples not selected for training are then considered for testing which is commonly known as "out-of-bag" (OOB) data (i.e *BootstrapSampling).*

ii. Build a decision tree using the training dataset by choosing the most important variable as the root node. Then, every following split is determined using the GINI index importance criteria for variables. Not all the variables are evaluated at every split, only a random sample is selected. Once the tree is constructed on the training set, the OOB data looked over on the tree and a prediction for each OOB case is obtained.

iii. The construction of each tree (step 2) is repeated a number of times (*ntree*) to construct the forest. The OOB prediction of each tree is averaged across the entire forest to obtain the final prediction.

2.3 Data Integration

Cancer projects are generating large amounts of genomic data from different experimental platforms due to rapid technological advances in biotechnology. Many of these data types are now available in public or government-funded repositories to the research community interested in performing data integration, exploration, and analytics.

Yet, many computational hurdles need to be overcome to uncover new knowledge from these data repositories. Not long ago, most studies focused their attention only at microarray analysis, which represent a one type of genomic data, yielding important discoveries in the field of biomarker detection in cancer. However, current research has concluded that these do not provide a complete picture of tumor behavior [37]. Therefore, to fully understand the roles of molecular inner workings it is necessary to integrate gene expression data with other omics technologies.

Each technology provides unique data which may not capture information from another platform. It is hypothesized that integration of multiple platforms of information should improve the biological understanding of cancer and the effectiveness of therapies. Then it is imperative to address the challenges of integrating currently available multi-omics data in order to obtain a better perspective of system functionality [38], [39]. For instance, in the data selection step, it is critical to study heterogeneous data types (Transcriptome, Glycome, Proteome, Metabolome) since their levels and characteristics are often different across platforms and experimental designs. At this stage, cleaning, missing value imputation, normalization and standardization should be properly performed. Another challenge in integration is the computational feasibility of these types of studies. Althougt there are many machine learning and data mining methods to extract important features, learn and predict complex structures, with the highly dimensional OMICS data, many of these methods cannot be directly applied because they are computationally demanding. This requires the reduction of features through feature selection and extraction techniques resulting in another criterion to take into account. Integrative models can be even more

difficult if it is desired to detect high degrees of interaction between features (four-way, three-way or pairwise), increasing even more the computational complexity requiring high performance computing to address those. Lastly, it is vital to translate computational methods into meaningful biological interpretation for physicians and researchers in the oncology field through summary statistics, pathway enrichment analysis and visualization, which are often demanding tasks.

Consequently, recent studies address some of these challenges, often through the application of machine learning tools such as classification and feature selection algorithms. For example, Kim *et al*. [40] developed a tool called ATHENA that implements a feature selection model based on a grammatical evolution neural network to analyze survival in cancer rates using copy number, gene expression, DNA methylation, and protein expression data. Their results showed that integrating these variables provided better survivability than with individual platforms data. Nonetheless, their methodology performance for prediction could be improved further (73%) though the adjustment (they only adjusted by age) or inclusion of other clinical factors. Other powerful technique currently used in this area are Support Vector Machines [41] and Random Forest [42]. List *et al* [43] applied RFs to integrate two technologies (methylation and gene expression) to find unique features for breast cancer classification (i.e. PAM50). Among the numerous advantages of using RFs is their feasibility to integrate mixed type variables, self-estimated variable importance scores and overall good performance in practice. Also, RFs has shown to be more robust when was evaluated against Neural Networks [44].

2.4 Evaluation Metrics for Interaction Between Features

Usually, common diseases are caused by the combinations of multiple genomic, pathological and lifestyle factors studied by the research community at large. It is increasingly evident that gene interactions plays a fundamental role in the proneness to common diseases [45]. Therefore identifying attribute interactions is progressively being accepted as a challenge in genetic epidemiology and human genetics needed to be tackled in order to fully understand cancer mechanisms [46], [47]. Data mining methods are becoming popular approaches to detect interactions in genetic studies. Many approaches have proven good performance in detecting gene-gene interactions, such as: penalized regression [48], multifactor dimensionality reduction (MDR) [49], neural networks (NNs) [50] and RFs [36].

MDR method was proposed as complimentary of logistic regression to detect gene interactions [49]. Since its initial description numerous variations have been proposed, such as extensions to imbalanced data [51], missing data [52], sparse or empty cells [53]. The drawback of MDR is its combinatorial nature which tends to become computationally expensive when using large amounts of data.

NNs is other method that has been used in genetics studies due to their ability to detect interactions in addition to main effects [54]–[56]. However, among the limitations of NNs is interpretability given its black box nature. Additionally, the connections between neurons are complex and considerable trial-and-error efforts are needed to obtain good network parameters.

One method that is very attractive to study gene-gene or gene-environment interactions is RFs, given that, as we have previously discussed, they have several intrinsic characteristics that fit very well with the requirements of molecular dataset [57]: (i) RFs give an explicit representation of feature interactions where several variable importance measures can be derived to identify driver genes in isolation but also in combination, (ii) they allow for the development of predictive model, without the need for strict assumptions on the underlying relationship between variables [58], (iii) they are computationally efficient and easily applicable to high dimensional problems as they are non-exhaustive and can be constructed in parallel. Moreover, to deal with certain types of genetic heterogeneity, tree methods are appropriate as we mentioned earlier. Studies show that RF performs better than univariate tests in the ability to detect interactions due to their tree construction process [58], [59]. Compared with methods based on traditional statistics such as, Fisher's Exact test, RF has presented better results [47].

For each input variables, tree-based ensemble methods can provide scores to obtain variable important measures (VIM) to rank variables (i.e. genes). Two popular measures to rank genes with Random Forest are the Mean Decrease Impurity (MDI) importance [36] and the Mean Decrease Accuracy (MDA) importance [60].

2.4.1 Mean Decrease Impurity

The impurity is a measure on which the optimal condition is normally chosen based on information gain/entropy (see Equation 2) as an objective function. It is expected that each variable used to build a tree, should reduce the impurity of this specific tree. This total reduction is called mean decrease impurity (MDI) [61]. In this sense, the impurity

decrease of a trained tree is calculated taking in account the weighted of the decrease impurity by each variable.

2.4.2 Mean Decrease Accuracy

Mean decrease accuracy (MDA) measure the impacts of each feature on accuracy of the general model by permuting the values of each feature and measure how much the permutation decreases the accuracy of the model. This is, importance of one specific variable is calculate by the mean decrease in accuracy from the original forest versus a model with randomly permuted variable values in the out-of-bag samples. For important variables the permutation must have a significant effect on prediction accuracy.

Despite the well-known ability of RFs to detect interactions these does not automatically provide the user with deeper degree of interactions as an output [47], [59], [62]. Therefore we will develop a new metric that will measure the prevalence of feature sets, which to the best of our knowledge, has not been explored in this manner before.

3 DATA-DRIVEN APPROACH TO EXTRACT MOLECULAR PATTERNS IN BREAST CANCER USING TRANSCRIPTOMIC AND CLINICAL DATA

3.1 Introduction

Despite significant biotechnology growth in recent years, cancer mechanisms are still not fully understood. There is plenty of literature work on microarrays to detect biomarkers oncogenes, however studying only one type of data is not sufficient to fully understand tumor behavior [37]. Therefore, we propose to integrate large and heterogeneous data types to associate with breast cancer subtypes. We aimed to address the high dimensionality challenge by applying feature selection methods (Correlation-based [22], Information gain [23] and ReliefF [24]) to reduce the feature space by choosing feature groups with the best classification performance [10]. Traditionally, these filter methods are the most employed gene selection approaches for their speed and computational simplicity [18], [19].

Once we have reduced the dimensionality of our features for the different genomic assays (i.e. gene expression, immunohistochemistry, clinical factors and pathological), we planned to integrate those using suitable classifiers. Machine learning methods are often used to model a response variable in terms of a diverse number of predictors facilitating the integration of heterogeneous data types. Some of the most powerful methods currently used are Support Vector Machines (SVM) [41] and Random Forest (RFs) [42], [43]. Among the numerous advantages of using RFs are their feasibility to integrate mixed type variables and self-estimated variable important scores. RFs has shown to be

more robust in comparison to other methods, such as classical Bayesian regression or Neural Networks [44], either on simulated [63] or on real data [46]. Consequently in this work we aim to extract biological important patterns with relevance at the phenotypic level, specifically breast cancer subtype identification. The extraction can be adressed through the exploration of new technology available and the integration of highly heterogeneous clinical datasets using different methods with overall good performance in practice.

3.2 Objective

Initially we will focus on identifying molecular patterns that help to discriminate among cancer patients and their subtypes, applying feature selection and classification methods towards integration of gene expression (microarray and RNA-sequence platforms) and clinical data. Since for the measurement of gene expression, is available in both microarray and RNA-sequencing format, the first objective is to discover which platform is the most informative with highest quality data, namely the platform with highest AUC and less error rate in the predicted model.

3.3 Methodology

The general procedure can be summarized in the diagram shown in **Figure 3-1**, which consists in four important steps: select data, preprocess, integration and biological interpretation.

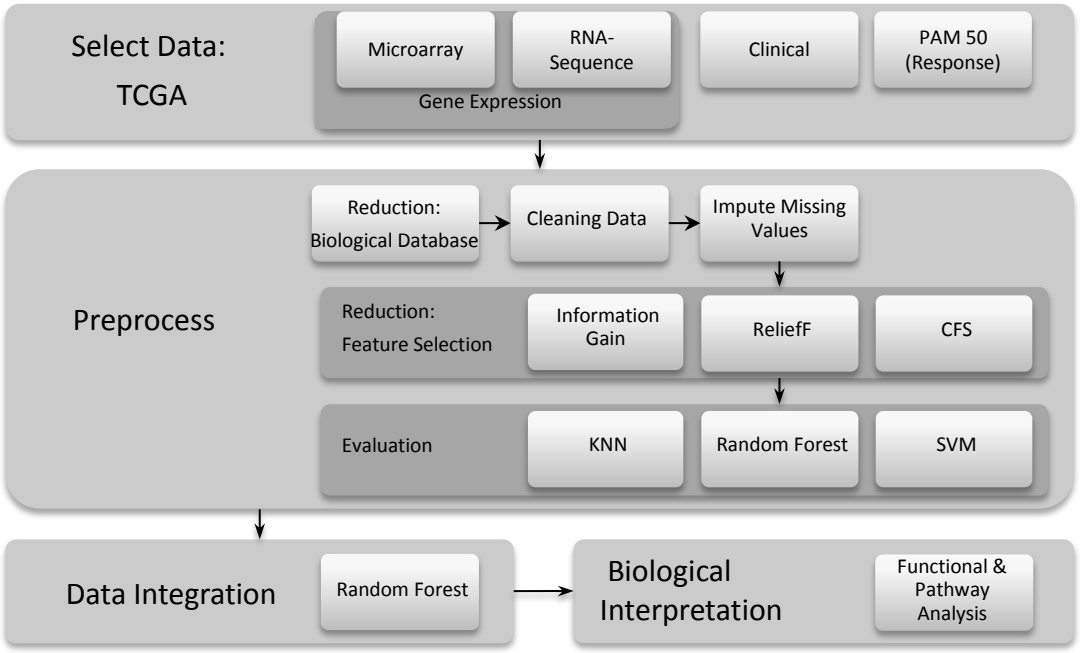


Figure 3-1. Phase 1 methodological framework. Consists of four important steps: select data, preprocess, integration and biological interpretations

3.3.1 Data Description

To build an integrative model aimed to predict breast cancer subtypes, we gathered all gene expression, clinical and subtype information from The Cancer Genome Atlas (TCGA) public repository (http://cancergenome.nih.gov) in processed and normalized form. TCGA has available microarray and RNA sequencing technology that describes profile transcriptome of the samples. Along with these data, TCGA provides a subtype classification of all gene expression samples via PAM50 that contain samples of 547 breast cancer patients, which becomes our response variable. **Error! Reference source not found.** contains a detailed description of the data of first phase.

Table 3-1. Description of microarray, RNA-seq, clinical and response data

	Attributes			Response
Name	Transcriptomic		Clinical data	Subtype breast cancer (PAM50)
	Microarray	RNA-seq		
Sample Size	547	1219	1077	547
Number of Attributes / Attribute Response	17814	20531	109	luminal A (232), luminal B (129), HER2 (58), basal (98), normal (30)
Missing Values	1570 (0.016%)	0	56465 (47.96%)	0

Microarray Dataset: This technology is commonly used to obtain genomic expression at the mRNA transcript level. Through cDNA microarray dataset is possible study several genes at the same time and extract important information of the cellular activity. The data to be used in this project was normalized using the common Lowess normalization methods, which is based for two fluorescent color assays to remove noise.

The microarray data set contains 547 observations (patients) and 17814 variables (genes) per patient, all variables are numerical, in similar scales, where the observations represent a logarithmic normalization (log2 lowess normalized (cy5/cy3)) by gene. This normalization results in values ranging from -11.8410 to 14.0395 when gene expression from a patient (cy5) is compared with a reference (cy3): positive values for more expression, negative values for less expression, and zero for no change between the two.

RNA-sequence Version 2 Dataset (RNASeq V2): This technology provides gene expression data through the mapping of nucleotide sequences of mRNA to the reference genome used. The method for Version 2 uses MapSplice software for mapping RNA-seq and RSEM (RNA-Seq by Expectation Maximization) to quantify gene expression. This

dataset contains 1219 observations (patients) on 20531 variables (genes) where all variables are numerical.

Clinical Dataset: Provides important pathological, histological and patient descriptive features provided by the specialized clinicians that attended patients' surgery to acquire samples. This data contains 1038 patients on 107 variables (note: not all patients have the two-array transcriptomic data described earlier). These variables will be inspected based on their missing value composition to include or not in the analysis (see Table 3-2).

To analyze the intrinsic characteristics of our data we can see that this poses a challenge for computational techniques to have a large dimensionality with a proportionately small sample size. The prediction models of microarray data can be affected by the sample size. For instance, in variable response found higher occurrence of luminal A subtype and low recurrence of normal subtype, with frequency equal to 232 and 30 samples respectively.

Table 3-2. Description of clinical data

# Var	Name Variable	Qualitative / Quantitative	Unique		Frecuency	%	Miss Val	Values		
								Min	Mean	Max
5	Gender	Qualitative	2	Female Male	1066 11	99 1	0			
6	menopause_status	Qualitative	3	peri* post** pre***	37 697 229	3 65 21	114			
7	Race	Qualitative	4	Indian Asian Black White	1 61 171 747	0 6 16 69	97			
58	vital_status	Qualitative	2	Alive Dead	974 103	90 10	0			
78	age_at_diagnosis	Quantitative						26	57	90
86	HER2_copy_number	Quantitative						2	53.56	441
107	tumor_tissue_site	Qualitative	1	Breast	1077	100	0			

*Peri (6-12 months since last menstrual period) (37, 3%)

**Post (prior bilateral ovariectomy OR >12 mo since LMP with no prior hysterectomy) (697, 65%)

***Pre (<6 months since LMP AND no prior bilateral ovariectomy AND not on estrogen replacement) (229, 21%)
Mixed-type - n = 107

The methodology for the preprocessing phase solved the high dimensionality of microarray datasets. Firstly, through a previous selection of features using biological data available, we compiled cancer-related genes from Cosmic [64], Vogelstein [65] and the Candidate Cancer Gene Database (CCGD) [66]. Therefore, the resulting data subset used to extract important features only included genes that were found in these three databases. We applied features selection methods (i.e. Information Gain, ReliefF, Correlation-based) and extracted the most relevant variables to distinguish between different subtypes. The performance of feature selection methods was evaluated through the following commonly known classifiers: KNN, SVM and RFs. According to the most accurate prediction model, we choosed between microarray and RNA-sequencing subsets to integrated with clinical data using Random Forest approach. This subset of important variables was evaluated to determine if the integration of transcriptomic and clinical data have significant effect on prediction accuracy. Finally, the biological meaning of the results were interpreted through functional and enrichment analysis. The results from this phase were used as a foundation towards a better implementation of extraction of features and its integration towards better subtype classification in the second phase.

3.3.2 Preprocess

- Reducing transcriptomic dataset through biological databases:

First, due to the high dimensionality of transcriptomic datasets (~20,000 genes) and the computational infrastructure needed to implement the suggested feature selection methods, we decided to initially reduce the amount of genes for the transcriptome by retaining genes with some implication in any type of cancer. To accomplish this task we

used the union of three cancer lists: Cosmic [64] with an aggregated list of 572 genes, Vogelstein [65] with 307 genes and lastly the Candidate Cancer Gene Database [66] with a list of 6790 known genes. This resulted in a subset of approximately 6500 genes for each dataset, microarray and RNA-seq.

- Cleaning data and imputing missing values:

For clinical data we removed eleven samples due to lack of information from immunohistochemistry status of ER and PGR receptor genes, yielding a total of 536 patient samples. From the 109 clinical variables available to be incorporated in our model, we removed 82 due to redundancies and missing value percentages that were higher than 14%. For imputing missing values in microarray and in reduced-clinical data, several methods were evaluated. These methods are: 1) iterative robust model-based imputation (IRMI) [67], 2) sequential imputation of missing values (impSeq) [68], 3) scalable robust estimators with high breakdown point for incomplete data (rrcovNA) [69], and 4) RFs modeling using R software. RNA-seq data did not have any missing values however microarray and reduced-clinical attributes showed 0.005% and 0.52% of missing information that were imputed using the discussed methods (these percentages of missing data were computed after the 11 samples were removed).

We selected IRMI method for microarray and clinical data because it showed minimum error rates after values were imputed. Finally, imputed dataset was used in the evaluation of feature selection methods.

- Implementing feature selection methods to transcriptomic data:

Clinical data was already reduced plus the number of resulting variables was manageable for modeling, hence we did not implement further reduction to this attribute type. To extract relevant variables in microarray data, we applied into the gene expression data three feature selection methods within the filters type using WEKA software [70]. The implemented filter methods were Correlation-based (CFS), Information Gain and ReliefF. These feature selection methods were assessed through the commonly known classifiers: k-nearest neighbor (KNN), SVM and RFs. Variations of cross validation (CV) were performed in all classifiers including leave-one-out and 10-fold to improve the model error estimate. For each trained KNN model, we use leave-one-out CV and evaluated different number of neighbors (K= 7, 11, 31 and 17) which was calculated based on the square root of the number of testing samples a common estimator guideline. The Knn.cv function from class R package [71] was applied using the discussed parameters. For the SVM implementation we selected the linear function as the kernel as well as 10-fold CV using the SVM from the e1071 R package [72]. Lastly, we used randomForest [73] with default parameters for the minimum number of randomly sampled candidates at each split (mtry= sqrt(#variables)) and the size of terminal nodes (nodesize=1). Furthermore, Random Forest models with a 5000 trees ensemble were implemented using equal-class sampling to reduce the effects of class imbalance . The number of trees was determined through initial tests until error estimation was stable. The metrics used to evaluate the performance of the feature selection methods were the Area Under the Curve (AUC) of Receiver Operating Characteristic and the error rate. Based on the combinatorial experiments (feature selection - classifier) of these metrics we chose the best selection

method and included the relevant attributes in the integrative model for breast cancer subtype prediction.

3.3.3 Data Integration Model

Once we have reduced the dimensionality of our features and selected the most important variables, we proceed to evaluate the performance of integrating transcriptomic and clinical data (microarray-clinical & RNA-seq-clinical). We aim to evaluate with these experiments whether this interaction has a significant effect to enhance subtype prediction. Also we aim to determine which of the two transcriptomic platforms yields higher accuracy. It is important to note that these variables are of different nature (i.e. numerical, categorical) and therefore there is a need to use an approach that can join this information in an applicable manner.Hence, we used a RFs model.

3.3.4 Biological Interpretation

Finally, genetic alterations analysis were executed using the cBioPortal (http://cbioportal.org) [74], [75] with the objective to measure the alteration prevalence per breast cancer subtype of the selected subset of attributes which are deemed to be important. We hypothesized to find a high percentage of extracted attributes in the list of commonly known markers in breast cancer as well as to find variables whose contribution has not been fully studied. Both findings validate the presented methodology and extend it to find patterns not previously understood. These attributes can provide insights into better understanding of how this disease is characterized.

3.4 Results

3.4.1 Feature Selection

For the microarray dataset, among the three evaluated feature selection methods, CFS yielded better results across all three classification approaches with error rates of 9.45%, 11.04% and 13.62% for RFs, SVM and KNN, respectively (see Table 3-3). CFS also outperformed Information Gain and ReliefF when evaluating AUC values (90.6%, 91.25%) for RFs and SVM models,ReliefF outperformed when applying KNN (91.1%). However, these AUC values were very similar ranging around 0.91 ± 0.004. Similar results were obtained when extracting attributes from RNA-seq data with best values (AUC: 90.57%, error rate: 11.42%) comparable to results from microarray selection. Therefore, this study finds CFS as the method with best prediction accuracy as well as AUC values for distinguishing PAM50 subtyping.

CFS is a method that selects subgroup of features depending on a correlation measure with the response class per a heuristic search strategy, such as Greedy Hill Climbing and Best First [22]. Every selection of subsets was performed with a 10-fold CV form. In this study, we have considered three different groups selected based on a threshold condition that considers retaining attributes that were included in all 10 folds, or at least 5 folds and at least 1 fold. Based on the performance metrics and the parsimony principle, we selected the subset of 76 and 79 variables reported by CFS as the relevant variables for subtype identification for microarray and RNA-seq respectively.

Lastly, our study reveals that the accuracy for the microarray platform was just slightly better than RNA-seq (AUC: +0.0003 | error Rate: +0.0197), allowing us to conclude that

the information from these two platforms are highly correlated and providing the same degree of information regarding the response variable. These are encouraging findings as these two platforms are theoretically thought to assess the same information at the genomic level but with different technology protocols.

Table 3-3. Results of evaluations feature selection methods for microarray data

			RFs		SVM		KNN		
F. Selection	Threshold	# Features	AUC	Error Rate	AUC	Error Rate	*K**	AUC	Error Rate
IG	1.8	58	0.6476	0.4100	0.5927	0.4103	17	0.6868	0.4403
	0.4	194	0.8943	0.1252	0.8962	0.1366	7	0.8837	0.1586
	0.3	368	0.8825	0.1272	0.8935	0.1319	7	0.9027	0.1418
	0.1	2300	0.8839	0.1493	0.8901	0.1424	17	0.8925	0.1866
ReliefF	0.08	70	0.8957	0.1185	0.9002	0.1141	7	0.8893	0.1754
	0.06	184	0.8900	0.1190	0.9052	0.1138	11	0.9082	0.1604
	0.04	519	0.8786	0.1293	0.9032	0.1102	**7**	**0.9110**	0.1437
	0.02	2060	0.8767	0.1450	0.8788	0.1439	17	0.8816	0.1954
CFS	10 folds	79	**0.9060**	0.1001	0.8963	0.1253	**11**	0.8999	**0.1362**
	>= 5 folds	282	0.9019	**0.0945**	**0.9125**	**0.1104**	11	0.9056	0.1381
	>= 1 folds	1007	0.8942	0.1215	0.9071	0.1153	11	0.9061	0.1679

**K*: best number of nearest neighbor

3.4.2 Integration by Random Forest

We accomplished to reduce the about 6500 genes to 79 relevant ones for microarray and 76 for RNA-seq, while maintaining high performance metrics to distinguish between subtypes using the feature selection methods discussed. These extracted genes per platform together with 27 other clinical factors were modeled using RFs. The result reflected no significant improvements in the metrics when compared with the models using transcriptomic and clinical data separately. This can be observed in the Multi-Dimensional Scaling (MDS) plots of proximity matrix from the RFs model in

Figure **3-2**, where plots with or without clinical variable yield similar AUC and error rates. Also, transcriptomic data in both microarray and RNA-seq format showed higher AUC values and lower error rates in contrast to the performance of the clinical-only model, revealing a higher importance to gene expression variables.

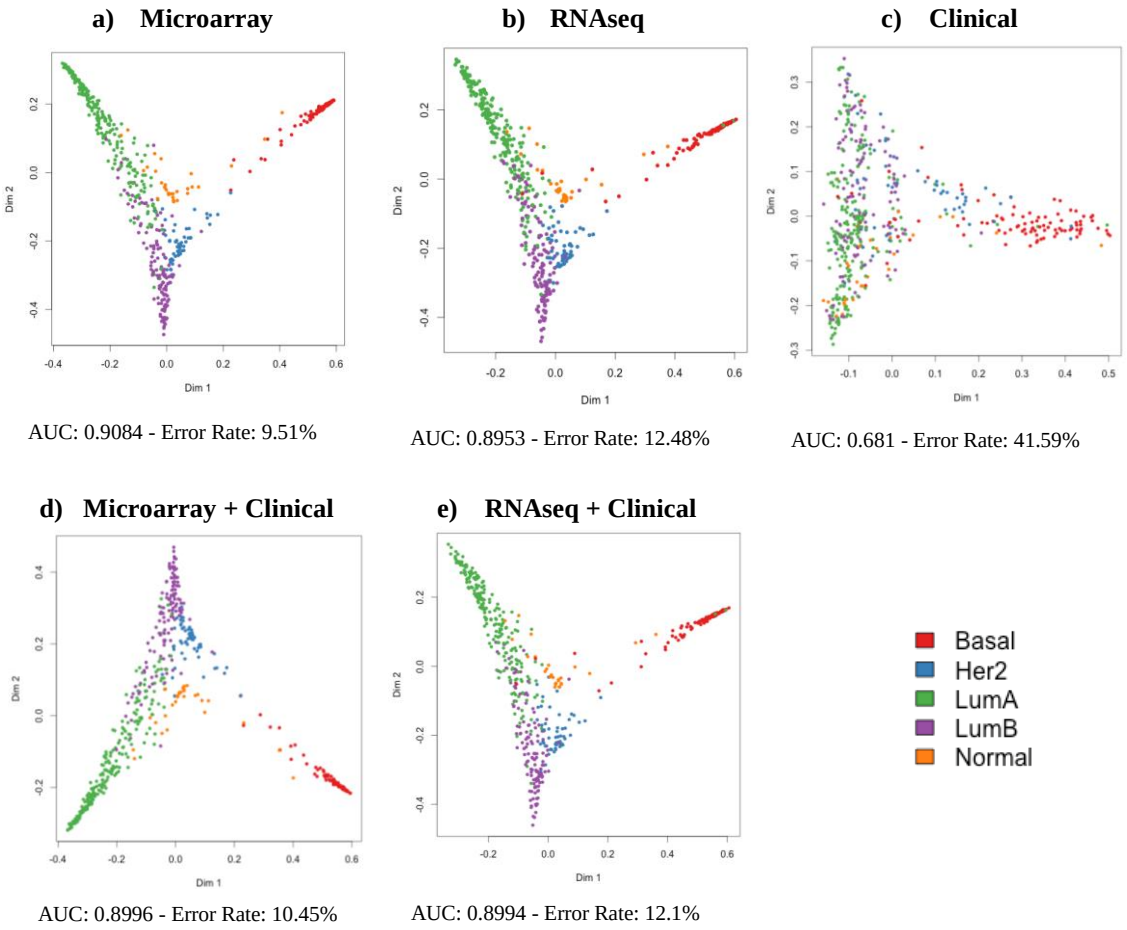


Figure 3-2. Multi-dimensional Scaling (MDS). Plot of proximity matrix from RFs for Microarray dataset (a), RNA-seq dataset (b), Clinical dataset (c), Microarray + Clinical datasets (d) and RNA-seq + Clinical datasets (e).

3.4.3 Biological Interpretation

Finally, to interpret the biological meaning of these results we further studied the 16 common genes among the top 30 relevant features obtained from both, microarray and RNA-seq Random Forest models. These overlapped genes showed a correlation of 0.827 showcasing the strong known association of the measurements from these platforms. The 16 common genes (FOXA1, FOXC1, ESR1, RGMA, THSD4, MIA, BCL11A, CRYAB, CMTM7, CENPK, IL17B, ERBB2, MASTL, STARD3, FGFR4, and DBNDD2) were analyzed through the tool available on the cBio website (http://www.cbioportal.org/) to evaluate the biological significance of our findings. The results were encouraging since it was found that these genes were considered altered with a z-score threshold of 2.0 folds in 98% and 95% of the samples for basal and HER2 subtypes respectively (See Figure 3-3). It is worth noting the downregulation mRNA that represents the FOXA1 gene to basal subtype, and upregulation for ERBB2 & STARD3 genes to HER2 subtype. Further biological experimentation on those features can be done to confirm association to breast cancer subtyping.

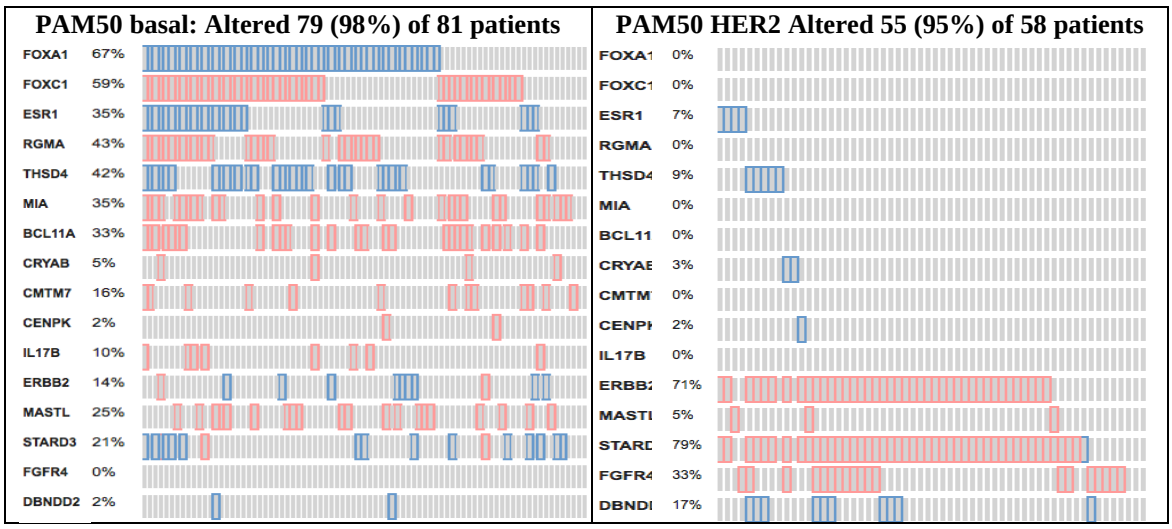


Figure 3-3. cBioPortal [74], [75] alterations plots for basal and HER2 subtype BC Genetic Alteration (mRNA Upregulation - mRNA Downregulation)

3.5 Conclusions

The findings observed in this work are very encouraging as they revealed high performance of feature selection methods in the selection of important variables. Genes such as FOXC1, ESR1 and FOXA1 have a proven track record in breast cancer because of its high impact in previous studies with this disease [76], [77]. These findings validate the prediction efficacy of our model and allows us to explore even further other genes such as THSD4, DBNDD2, CENPK and ANLN that have only few research studies related to cancer, especially that ANLN has a low relationship level with breast cancer subtyping. Further experimental studies of these genes behavior in breast cancer are recommended to better understand its causality effect. Future studies should explore other feature selection methods and classifiers to improve prediction performance through better tuning of its parameters. Also, integrating other types of genomic assays such as

methylation and protein expression levels could reveal deeper interactions important to understand the mechanisms of each breast cancer subtype especially those of aggressive behavior.

4 DATA-DRIVEN APPROACH TO EXTRACT MOLECULAR PATTERNS IN BREAST CANCER USING TRANSCRIPTOMIC, PROTEOMIC AND METHYLATION DATA

4.1 Introduction

Cancer genome data from research projects, such as The Cancer Genome Atlas (TCGA), can provide a wealth of information to better understand cancer biology if principal computational challenges are addressed. There is plenty of literature on microarrays to detect biomarkers oncogenes and tumor suppressors, but most are performed using data from similar platform technologies. The integration of heterogeneous data types is a very challenging task even more in cross omics analysis: scaling, mapping, missing values, sample size and the *curse of dimensionality* are among the issues related to integrating omics. In this project, we applied a multi-phase data-mining approach to integrate large data types (i.e. protein, gene expression and methylation) association with breast cancer subtypes and extract interactive pattern. The main purpose at this project is to deepen the understanding of each breast cancer subtype to optimize clinical treatment and improve patient outcomes. We aimed to address the high dimensionality challenge by applying several feature selection (FS) methods: Correlation-based feature selection (CFS) [22], Information gain (IG) [23], ReliefF [24], Fast clustering-based feature selection algorithm (FAST) [33] and Support Vector Machine based on Recursive Feature Elimination (SVM-RFE) [28], to reduced the feature space choosing feature groups with the best classification performance. We evaluated these FS methods using Random Forest (RF),

k-nearest neighbor (KNN) and support vector machine (SVM) classifiers measuring accuracy and Area Under the Curve (AUC).

Once we have reduced the dimensionality of our features for the different genomic assays, we integrated these to extract meaningful interactions that can distinguish the phenotypic outcome of study: breast cancer subtyping. In practice, there are diferents methods used to integrate different genomic datasets and model a response variable in terms of a diverse number of predictors with overall good performance [41]–[43], such as: Support Vector Machines (SVM) [35] and Random Forest (RFs) [36].

Finally, we performed to do an enrichment analysis, in order to extract biological insights from our gene sets, through the use of ClusterProfile [78] R package which uses commonly known pathway and functional repositories such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [79] and Gene Ontology Consortium (GO) [80]. The validation was evaluated using a visualization technique: heat maps, , for most relevant features found compared with other datasets from Gene Expression Omnibus (GEO) data repository. We aimed to integrate large amounts of heterogeneous data types to learn distinct and common profiles across different cancer types, specifically breast cancer subtype identification that enable researchers to deepen the understanding of cancer and could allow a more accurate diagnosis, early prognosis or even a personalized treatment.

4.2 Objective

To identify molecular patterns that can discriminate among breast cancer patients and their subtypes, by applying feature selection methods using gene, protein, methylation

and subtype information from over 560 breast tumor samples under The Cancer Genome Atlas.

4.3 Methodology

The methodology of this chapter can be summarized in the diagram shown in **Figure 4-1**, which consists of five main steps: data selection, feature selection, integration, biological interpretation and external validation.

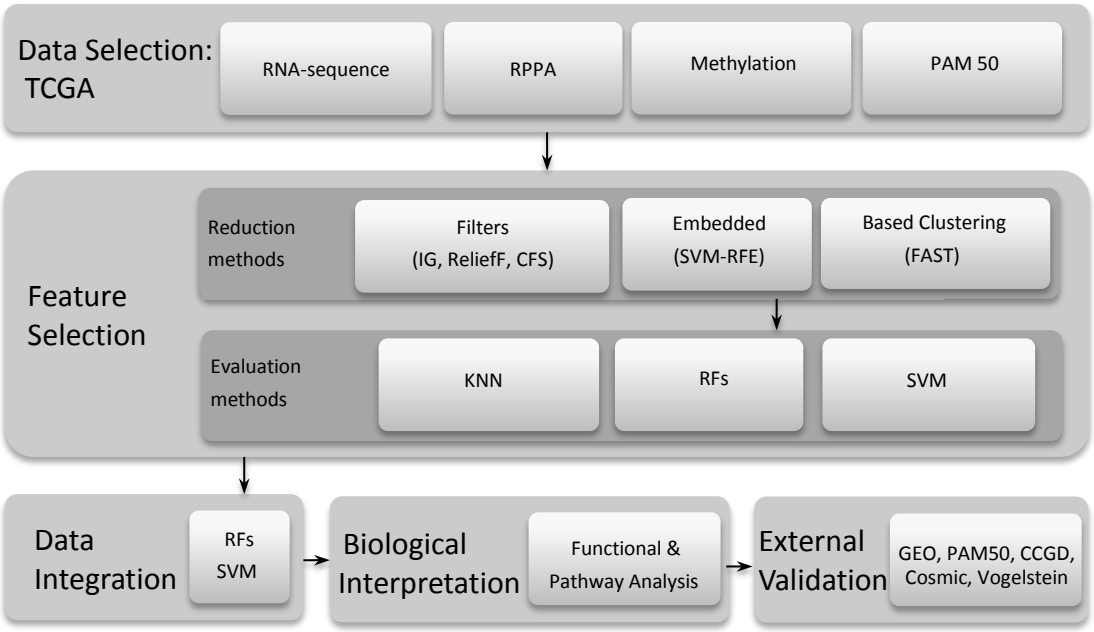


Figure 4-1. Phase 2 methodological framework. Consists in five important steps: data selection, feature selection, data integration, biological interpretations and external validation.

Initially we will focus on identifying relevant molecular patterns on each dataset to discriminate among cancer patients and their subtypes using feature selection techniques. To evaluate the performance of feature selection techniques, we will measure AUC and error rate through the implementation of three well-known classification methods: KNN, SVM and RFs. Then, we performed the integration of protein, methylation and gene

expression using relevant factors.Finally, for biological interpretation, we planned to assess the relevant results using functional and pathway analysis through known biological databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG).

4.3.1 Data Description

To build an integrative model to predict breast cancer subtypes, we gathered all gene expression, protein, methylation and subtype information from The Cancer Genome Atlas (TCGA) public repository (http://cancergenome.nih.gov/) using level 3 data. Table 4-1 contains a detailed description of the data used in this work.

Gene expression: TCGA provides microarray and RNA sequence (RNA-seq) to evaluate gene expression levels at the transcriptome level, however in a preliminary study [81] we concluded that the information from these two platforms are highly correlated and provide the same degree of information regarding the response variable. Hence we used RNA-seq platform data since it had no missing values.

RNA-sequence Version 2 Dataset (RNA-Seq V2): The information given by RNA-Seq technology is gene expression through the mapping of nucleotide sequences of mRNA to the human reference genome used (hg19). The RNA-Seq Version 2 uses MapSplice software for mapping RNA-seq to reference genome and RSEM (RNA-Seq by Expectation Maximization) to quantify gene expression. This dataset contains 1219 observations (patients) on 20531 variables (genes) where all variables are numerical,with values ranging from 0 to 20,656,039.

Reverse phase protein array (RPPA): Protein Array dataset provide expression levels and protein concentration. The platform used to extract this information comes from the M.D.

Anderson (MDA) Reverse Phase Protein Array Core. The dataset contains 633 observations on 187 variables where all variables are numerical.

Methylation: Changes in the expression of DNA can be associated with transcriptional inactivity when located in promoter regions and this type of data provides deeper understanding of the transcription or lack thereof. The methylation values obtained in this data included methylated and non-methylated probes from the CpG islands. TCGA used the Illumina Infinium Human DNA Methylation 450 platform and the dataset used contains 940 observations on 21986 numerical variables with 6.8% of missing information. We removed variables with missing values, yielding a total of 20486 attributes (CpG island targets).

Breast cancer subtype (PAM50): Subtype classification (basal, HER2, luminal A, luminal B, normal) of all samples is provided via PAM50 [82]. This is our response variable which is provided for 817 breast cancer patients.

Table 4-1. Description of RNA-seq, RPPA, methylation and response data

	Attributes			Response
Name	RNA-seq	RPPA	Methylation	Subtype breast cancer (PAM50)
Number Sample	1219	633	940	817
Number Attribute / Attribute Response	20531	187	21986	luminal A (415), luminal B (176), HER2 (65), basal (136), normal (25)
Missing Value	0	0	7206 (6.8%)	0

4.3.2 Feature Selection Implementation

To extract relevant variables to distinguish between different subtypes, we apply five different feature selection methods (IG, ReliefF, CFS, SVM-RFE and FAST) within the

filters, embedded and based clustering type using R packages and WEKA software [70]. The proposed methods were selected based on their ability to work with large data dimensionality and their relevance in the field of genomics.

For implementation of IG and ReliefF filters methods, we used FSelector [83] package in R with the following parameters, for instance, in IG the unit for computing entropy the default is "log"(unit = "log2"), in ReliefF function the number of neighbors to find for every sampled variables was 5 (neighbours.count = 5), and the number of variables to sample was 10 (sample.size = 10). For CFS filters method we used CfsSubsetEval as evaluator in Weka software, as well as cross validation (CV) was performed with a 10-fold CV form to improve the model error estimate. Furthermore, SVM-RFE embedded method was implemented using the svmrfeFeatureRankingForMulticlass function from the OmicsMarkeR R package. SVM model was implemented with one-against-one approach using linear function as the kernel and 10 percent of features removed during each iteration (perc.rem = 10). Lastly, the FAST method was implemented using a Java package in WEKA, in [84] we find a complete guide to its use.

All these feature selection methods were assessed through the following classifiers: k-nearest neighbor (KNN), support vector machine (SVM) and Random Forest (RF) as they are commonly known to have good overall performance in the context considered in this work. Variations of CV was performed in all three classifiers including leave-one-out and 10-Fold to improve the model error estimate. To implement the KNN method, we used leave-one-out cross validation and evaluated different number of neighbors (K) for each trained model. We used values of K= 7, 11, 31 and 17 which was calculated based on the

square root of the number of testing samples, a common estimator guideline. The Knn.cv R function was applied using the discussed parameters. For SVM implementation we selected the linear function as the kernel as well as a 10-fold cross validation using the SVM from the e1071 R library. Lastly, we used randomForest in R with default parameters except for number of trees. Furthermore, Random Forest models were implemented using equal-class sampling to reduce the effects of class imbalance. In addition, it also used 5000 trees in their ensemble, this number was estimated through initial tests until error estimation was stable.

For each classifier, we evaluated two metrics: Area Under the Curve (AUC) of Receiver Operating Characteristic and error rate. Based on the combinatorial experiments (feature selection - classifier) of these metrics we chose the best feature selection method and included the relevant attributes in the integrative model for breast cancer subtype prediction.

4.3.3 Data Integration Model

Once we have reduced the dimensionality of our features and selected the most important variables, we proceeded to evaluate the performance of integrating protein, methylation and gene expression. We proceeded to evaluate with these experiments whether this interaction has a significant effect to enhance subtype prediction. Lastly, we used the reduced list of important features to evaluate and infer its biological meaning.

4.3.4 Biological Interpretation

To accomplish the biological interpretation of the results we apply two different mechanisms. First, the partial dependence plots (PDP) tool, available on RandomForest R

package, can analyze the marginal contribution of a single predictor in a given class using a log-odds metric. This metric is currently implemented through graphical visualizations and can be used to infer the likelihood of being classified into a particular class based on a specific predictor value (see an example in Figure 4-2). The second mechanism is an enrichment analysis to identify a set of genes within known gene-groups by their functional or pathway categories . For this type of analysis, we used the enrichKEGG and enrichGO functions from clusterProfiler package [78] in R supported by the Bioconductor annotation database system. For both functions, we selected for the organism "hsa" (homo sapiens-human) and the method of adjustment of the p-value was "BH" (Benjamin-Hochberg). In the case of enrichment using KEGG, the parameters pvalueCutoff and qvalueCutoff specifies the maximum cutoff value for the p-value and q-value, respectively. Different values for these parameters were used to gather different levels of statistical confidence. For example we set the parameter values to 1 in order to find all known pathways in our 247 selected genes. For enrichment analysis using the gene ontology catalog, we focused on molecular function (ont="MF") and a maximum p-value of 0.01 (pvalueCutoff=0.01).

4.3.5 External Sources Validation

For validation purposes, we evaluate the classifying performance of our features extracted in the integrative step using two external datasets from GEO: GSE20685 with 327 breast cancer samples and GSE21653 with 266 early breast cancer samples. We evaluated the classifying performance of GEO variables using Random Forest methodology with default parameters for the minimum number of randomly sample

candidates at each split (mtry=sqrt(#variable)) and the size of terminal nodes (nodesize=1). Furthermore, Random Forest models were implemented using equal class sampling to reduce de effect of the class imbalance and it also used 17000 and 9000 trees in their ensemble for GSE20685 and GSE21653 data, respectively. These tree numbers were estimated through initial tests until error estimation was stable. Also, we evaluated how the extracted features overlap with known cancer lists. To accomplish this task we used three cancer lists: Candidate Cancer Gene Database (CCGD) [66] with a list of 7088 known genes, Cosmic [64] with an aggregated list of 594 genes, Vogelstein Science 2013 [65] with 255 genes and, lastly, PAM50 gene list [82]. Additionally, to complete this functional search we used GeneCard human genes database (www.genecards.org) [85] and PubMed search engine resource [86] as additional sources to support previous links to known diseases and its amount of published supporting work respectively. We expected to find the selected attributes in the list of commonly known genes in cancer. Also, we expected to find variables whose contribution has not been fully studied as well. Both are important findings to validate the presented methodology and to extend it to patterns not previously understood. These attributes can provide insights into better understanding how this disease is characterized.

4.4 Results

4.4.1 Feature Selection

The performances of all feature selection methods across all three datasets were similar. The model for the RNA-seq only data was slightly better than the models with

methylation and protein expression by itself. CFS yielded higher AUC values (85.06%, 81.75%) and error rates (8.93%, 23.36%) for RFs and KNN, respectively, within all the feature combinations evaluated. When applying SVM classifier, ReliefF resulted with highest AUC value (82.17%) and SVM-RFE yielded lower error rate (13.34%) nonetheless these were not better than those found using CFS. In fact, these values were very similar ranging (AUC: +0.01 | error rate: +0.02) from CFS method (See Table 4-2).

For RPPA dataset, among all evaluated feature selection methods, also CFS gave better AUC values across all three classification approaches with of 82.34%, 82.19% and 80.57% for RFs, SVM, and KNN, respectively (see Table 4-3). CFS also outperformed when evaluating error rate values (19.67%, 21.96%) for RFs and KNN models but ReliefF when applying SVM (19.92%). However, this error rate value was very similar to CFS method (ranging +/- 0.005).

For methylation data, CFS yielded better AUC values (82.73%, 81.08%) and error rate (23.49%, 26.67%) for RFs and KNN classifiers respectively across all feature selection methods. In contrast SVM-RFE gave better results (AUC: 81.43% | error rate: 12.01%) when applying SVM classifier (See Table 4.4).

Based on the performance metrics shown here and the parsimony principle, we selected the subset of 29 and 542 variables reported by CFS as the relevant variables for subtype identification for RPPA and RNA-seq data respectively. In terms of methylation probes we evaluated two sets of important features: the first from CFS with 158 variables and the second from SVM-RFE with 120 variables. These features were used later during out integration methodology steps.

Table 4-2. Results of evaluations feature selection methods for RNA-seq data

			RFs		SVM		KNN		
F. Selection	Threshold	# Features	AUC	Error Rate	AUC	Error Rate	*K**	AUC	Error Rate
IG	0.400	18	0.8169	0.1617	0.8168	0.1873	5	0.7694	0.2583
	0.350	71	0.8318	0.1515	0.8117	0.1700	11	0.7874	0.2656
	0.300	189	0.8473	0.0990	0.8101	0.1743	5	0.7964	0.2729
	0.250	430	0.8502	0.1004	0.8095	0.1604	7	0.7466	0.2570
	0.200	935	0.8464	0.1079	0.7951	0.1572	9	0.7355	0.2375
	0.150	1922	0.8420	0.1177	0.7875	0.1664	5	0.7640	0.2277
	0.130	2503	0.8378	0.1177	0.7826	0.1617	7	0.7641	0.2350
	0.100	3834	0.8310	0.1190	0.7828	0.1809	5	0.7567	0.2289
ReliefF	0.090	15	0.7224	0.2531	0.6839	0.2788	7	0.6110	0.3599
	0.700	83	0.7596	0.2077	0.7692	0.2057	5	0.7401	0.2338
	0.050	226	0.7823	0.1782	0.7893	0.1832	7	0.7203	0.2509
	0.030	1054	0.8277	0.1399	0.8126	0.1510	9	0.7121	0.2460
	0.020	2639	0.8247	0.1412	**0.8217**	0.1442	11	0.7339	0.2460
	0.017	3333	0.8248	0.1411	0.8217	0.1474	9	0.7197	0.2521
CFS	10 folds	127	0.8473	**0.0893**	0.7948	0.1619	11	**0.8175**	**0.2336**
	>= 5 folds	542	**0.8506**	0.0919	0.8020	0.1622	11	0.7602	0.2729
	>= 1 folds	1737	0.8445	0.1067	0.7774	0.1753	7	0.7388	0.2693
SVM-RFE	NA	30	0.8314	0.1311	0.7828	0.1802	9	0.7231	0.2375
		60	0.8357	0.1238	0.7954	0.1506	7	0.7096	0.2497
		120	0.8410	0.1227	0.7828	0.1395	5	0.6733	0.2693
		250	0.8438	0.1313	0.8058	0.1358	5	0.7170	0.2632
		500	0.8343	0.1203	0.7924	0.1580	9	0.6969	0.2938
		1000	0.8254	0.1326	0.8136	**0.1334**	7	0.6832	0.3035
		2000	0.8329	0.1240	0.8021	0.1491	9	0.6936	0.2938
		3000	0.8243	0.1252	0.8003	0.1363	11	0.6803	0.2925
FAST	NA	385	0.8228	0.1620	0.7926	0.1646	7	0.7904	0.2411

**K*: best number of nearest neighbor

Table 4-3. Results of evaluations feature selection methods for RPPA data

			RFs		SVM		KNN		
F. Selection	Threshold	# Features	AUC	Error Rate	AUC	Error Rate	*K**	AUC	Error Rate
IG	0.1	24	0.8144	0.203	0.8065	0.2091	11	0.7497	0.2243
	0.05	57	0.8195	0.1983	0.8016	0.2118	5	0.7711	0.2338
	0.04	79	0.8174	0.1983	0.8034	0.2115	11	0.7392	0.2385
	0	114	0.8175	0.1999	0.8098	0.2056	11	0.7439	0.2338
ReliefF	0.03	13	0.7939	0.2157	0.8032	0.2182	11	0.7715	0.2385
	0.02	30	0.8199	0.2125	0.801	0.2194	11	0.7183	0.2449
	0.015	55	0.8192	0.2014	0.8036	**0.1992**	9	0.731	0.237
	0.01	106	0.8157	0.211	0.8146	0.2088	7	0.7552	0.2306
CFS	10 folds	17	0.8068	0.2173	0.8115	0.2152	11	0.7249	0.2401
	>= 5 folds	29	**0.8234**	**0.1967**	0.8017	0.2116	11	**0.8057**	**0.2196**
	>= 1 folds	57	0.8202	0.203	**0.8219**	0.2043	9	0.7382	0.2433
SVM-RFE	NA	30	0.8124	0.2141	0.7729	0.2179	5	0.7094	0.2749
		60	0.8081	0.2094	0.8031	0.2101	5	0.7141	0.2449
		90	0.8177	0.2094	0.8041	0.2274	5	0.7278	0.2686
		120	0.816	0.2173	0.8078	0.2118	11	0.7417	0.2512
FAST	NA	8	0.7874	0.2348	0.7718	0.2286	7	0.782	0.2497

**K*: best number of nearest neighbor

Table 4-4. Results of evaluations feature selection methods for methylation data

			RFs		SVM		KNN		
F. Selection	Threshold	# Features	AUC	Error Rate	AUC	Error Rate	*K**	AUC	Error Rate
IG	0.25	24	0.6836	0.2943	0.6673	0.2761	11	0.6586	0.3244
	0.2	120	0.6633	0.3092	0.7261	0.2592	7	0.6572	0.3156
	0.15	344	0.6519	0.2989	0.7889	0.2424	5	0.686	0.323
	0.1	894	0.7942	0.293	0.7432	0.2386	5	0.6744	0.3126
	0.07	1648	0.8024	0.2901	0.7516	0.2337	7	0.6746	0.3185
	0.05	2664	0.7936	0.2901	0.752	0.2371	5	0.6702	0.3052
	0.04	3627	0.7904	0.2901	0.7543	0.2336	5	0.6707	0.3126
	0	6116	0.642	0.2857	0.7408	0.2399	5	0.6819	0.323
ReliefF	0.09	34	0.6467	0.3018	0.785	0.2356	11	0.6657	0.283
	0.08	66	0.656	0.2677	0.748	0.2268	5	0.6585	0.2993
	0.05	405	0.6507	0.2811	0.7536	0.2371	7	0.6742	0.3007
	0.03	1286	0.6537	0.2797	0.7712	0.2161	5	0.6766	0.2933
	0.02	2754	0.6497	0.2841	0.7526	0.2293	5	0.6795	0.2993
	0.01	3333	0.6494	0.2812	0.8	0..2221	5	0.6861	0.3081
CFS	10 folds	41	0.6928	0.2394	0.7625	0.2203	9	0.7242	0.2622
	>= 5 folds	158	**0.8273**	**0.2349**	0.7801	0.1853	9	**0.8108**	**0.2667**
	>= 1 folds	450	0.6842	0.2469	0.7814	0.1954	11	0.7612	0.2889
SVM-RFE	NA	30	0.6945	0.262	**0.8143**	0.1766	5	0.7632	0.277
		60	0.7104	0.2365	0.7971	0.1318	5	0.736	0.2563
		120	0.7027	0.2365	0.8066	**0.1201**	5	0.7155	0.2593
		250	0.6685	0.25	0.8012	0.1304	7	0.7966	0.2889
		500	0.6718	0.2396	0.8062	0.1284	11	0.7781	0.2963
		1000	0.674	0.2619	0.7873	0.1425	5	0.7095	0.2978
		2000	0.6572	0.2678	0.7693	0.1808	9	0.6687	0.3126
		3000	0.6506	0.2767	0.7563	0.2018	9	0.6687	0.3185
FAST	NA	112	0.638	0.2856	0.8126	0.2601	7	0.7028	0.3822

**K*: best number of nearest neighbor

4.4.2 Data Integration

We evaluated the integration of all variables reported by the feature selection methods as the relevant variables for subtype breast cancer identification. Two groups of important variables were evaluated through RFs and SVM classifiers, given that two subsets of variables for methylation were identified. The first group contained 729 features, as result of the sum of variables reported by CFS method for each dataset. Out of the 729 attributes, 542 were from RNA-seq, 29 were from RPPA and 158 were from methylation. The second set was composed of 691 features, the same 542 from RNA-seq and 29 from RPPA reported by CFS used in the first group but 120 methylation probes extracted by SVM-RFE method.

The results from first group yield better AUC values (85.90%, 81.75%) and error rate (9.48%, 17.83%) when compared to the second group (lower AUC values: 84.55%, 82.66% and higher error rates: 9.66%, 9.48%) for RFs and SVM respectively. Lastly, in both cases RFs yielded better metrics overall than SVM. Consequently, we decided to select the first group of 729 omics features, with RFs classifier to further reduce this set prior to integration.

Using these features, we obtained high performance metrics to distinguish between breast cancer subtypes. The results reflected improvements when compared with the models using the RNA-seq, proteomic and methylation datasets separately. This can be observed in the Multi-Dimensional Scaling (MDS) plots of the proximity matrix from the RFs model in Figure 4-3, where the integrative model showed better AUC values in contrast to the performance of separate models. In terms of error rates, the values between the

integrative and RNA-seq only model were comparable perhaps since most of the important features in the integrative model were from the RNA-seq platform. Figure 4-4 shows the top thirty important variables ranked by mean decrease in Gini Score for each RF model trained using the RNA-seq, RPPA dataset, methylation separate datasets and all important variables from each dataset used in the integrative model. The features from RNA-seq platform dominates the ranking of variables integrated (see Figure 4-4 (d)). However, it is noted than cg02643667 methylation and ER.alpha protein are within the top thirty relevant variables of the integrative group, this inclusion seems to slightly improve the sensitivity and specificity of the model as gathered through the AUC.

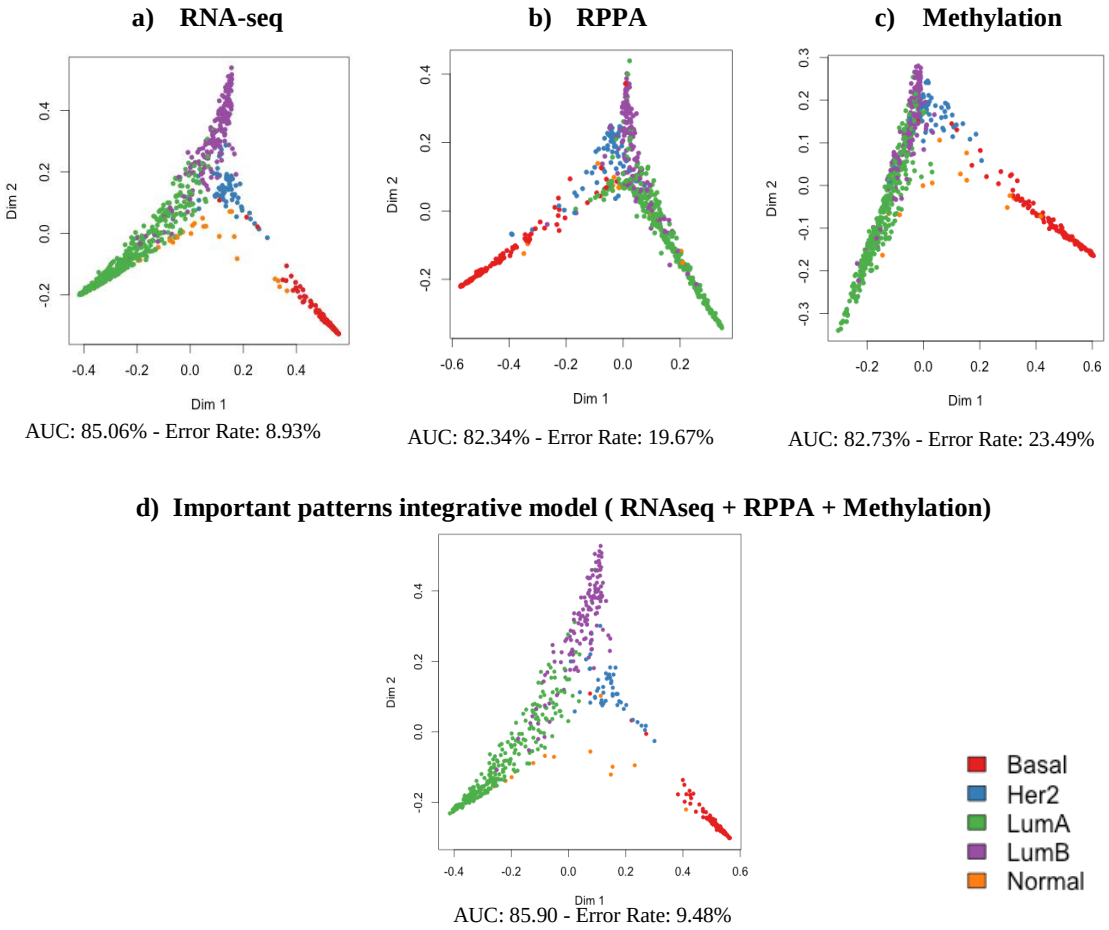


Figure 4-3. Multi-Dimensional Scaling (MDS). Plots of proximity matrix from RFs for: (a) RNA-seq, (b) RPPA, (c) methylation and (d) integrative model of important patters from RNA-seq, RPPA and methylation datasets. This integrative model showed better AUC values and lower error rates in contrast to the performance of separate models.

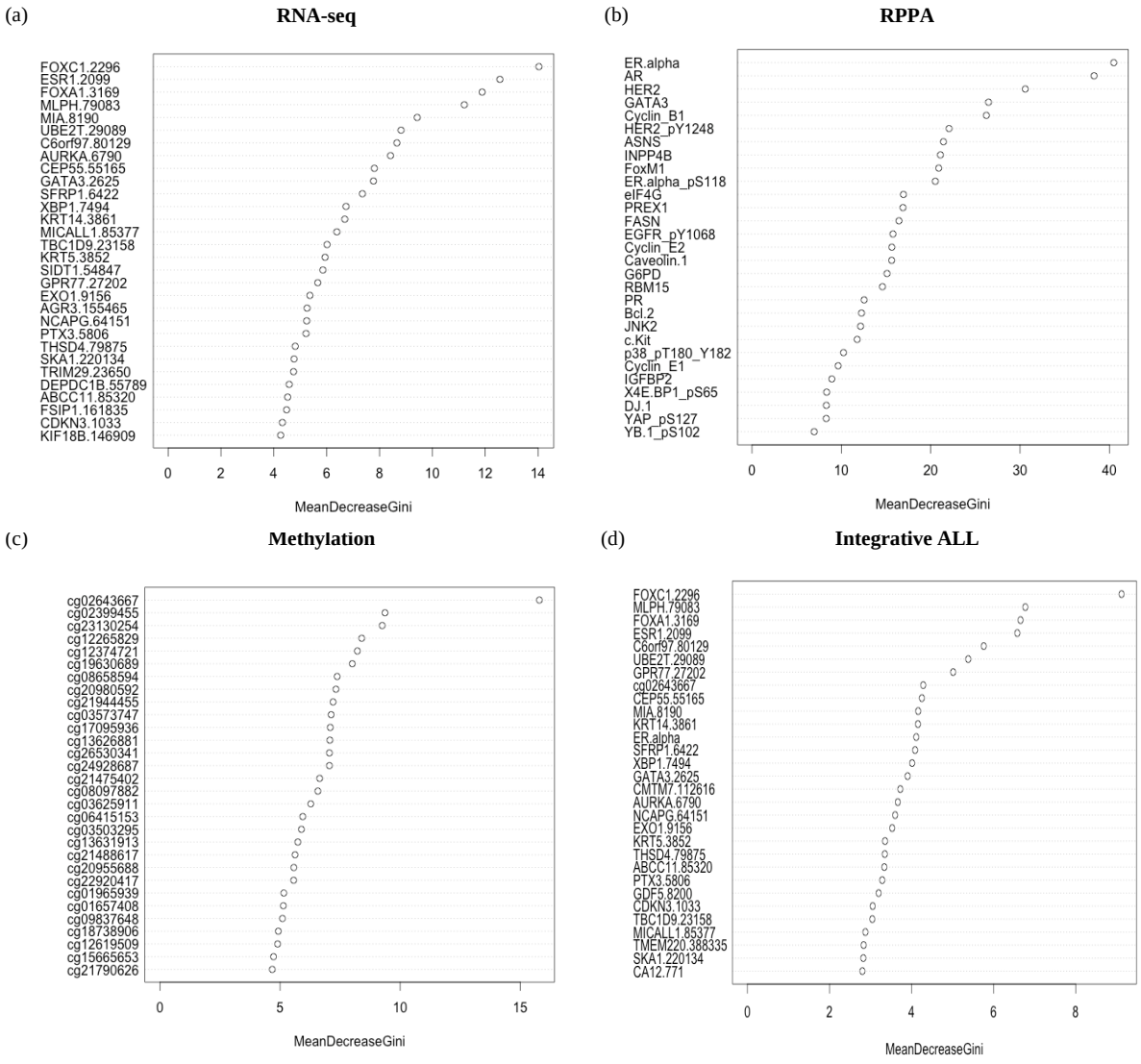


Figure 4-4. Variable Importance Plots (VIM). This plot visualizes the top thirty importance variables by mean decrease in Gini Score that RFs picked for: (a) RNA-seq dataset, (b) Methylation dataset, (c) RPPA dataset and (d) all integrative important variables selected for each dataset

Looking to further reduce the number of variables, we evaluated different thresholds based on the MDG scores of all 729 variables in the integrative model. The set of 477 variables yielded the highest AUC (0.8641) and lowest error rates (0.0877) as shown in

Table 4-5. However the groups of variables 328 and 247 were very similar ranging (AUC: +0.003 | error rate: +0.001). Finally based on these metrics and the parsimony principle, we selected the group of 247 variables to perform biological interpretation.

Table 4-5. Evaluations of different thresholds for important variables

Threshold MDG	# Features	RFs	
		AUC	Error Rate
4.00	10	0.8367	0.1053
3.00	26	0.8286	0.0929
2.00	46	0.8405	0.0999
1.00	114	0.8301	0.0947
0.80	144	0.8303	0.0929
0.50	212	0.8482	0.0912
0.40	247	**0.8607**	**0.0894**
0.30	328	0.8614	0.0877
0.20	**477**	**0.8641**	**0.0877**
0.10	711	0.8600	0.0912
All	729	0.8590	0.0948

4.4.3 Biological Interpretation

We started with around 42700 variables and reduced them to 247 relevant ones as described earlier. To interpret the biological meaning of these results, we analyzed the contribution of the seven most important variables (FOXC1, MLPH, FOXA1, C6orf97, ESR1, UBE2T, GPR77) according to the most representative jumps on the variables scores for Mean Decrease Gini (MDG) given by RFs as shown in Figure 4-4 (d). We constructed PDPs that shows the effect of these seven variables in the prediction of each one of the breast cancer subtypes (basal, HER2, luminal A, luminal B and normal) (see Figure 4-5). For instance, there is a higher likelihood of finding that a sample is basal subtype (see Figure 2 (a)) if FOXC1 is overexpressed and MLPH and FOXA1 are

underexpressed. The contrary, underexpression of FOXC1 and overexpression of MLPH/FOXA1 is found in all other subtypes with no distinctive pattern. Moreover, the underexpression of UBET2 likelihood is uniquely seen for luminal A samples (see Figure 2 (c)). Also, HER2 samples can be differentiated with luminal B samples by a combination of patterns. HER2 subtype is characterized by underexpression of C6orf97, ESR1 and GPR77 while luminal B samples shows overexpression of these genes when FOXC1 is underexpressed and MLPH, FOXA1, UBE2T are over expressed as shown in Figure 2 (b) and 2 (d). The partial dependence plots of these genes identify unique expression patterns for each breast cancer subtype.

Also, we collected the pathways of each of these seven most important variables: FOXC1, MLPH, FOXA1, C6orf97, ESR1, UBE2T, GPR77 (see Table 4-6). To gather the pathways, we made use of different sources such as: wikipathways, reactome and kegg. As results, we found that all these genes were in different pathways, except C6orf97, for which no specific link has been found.

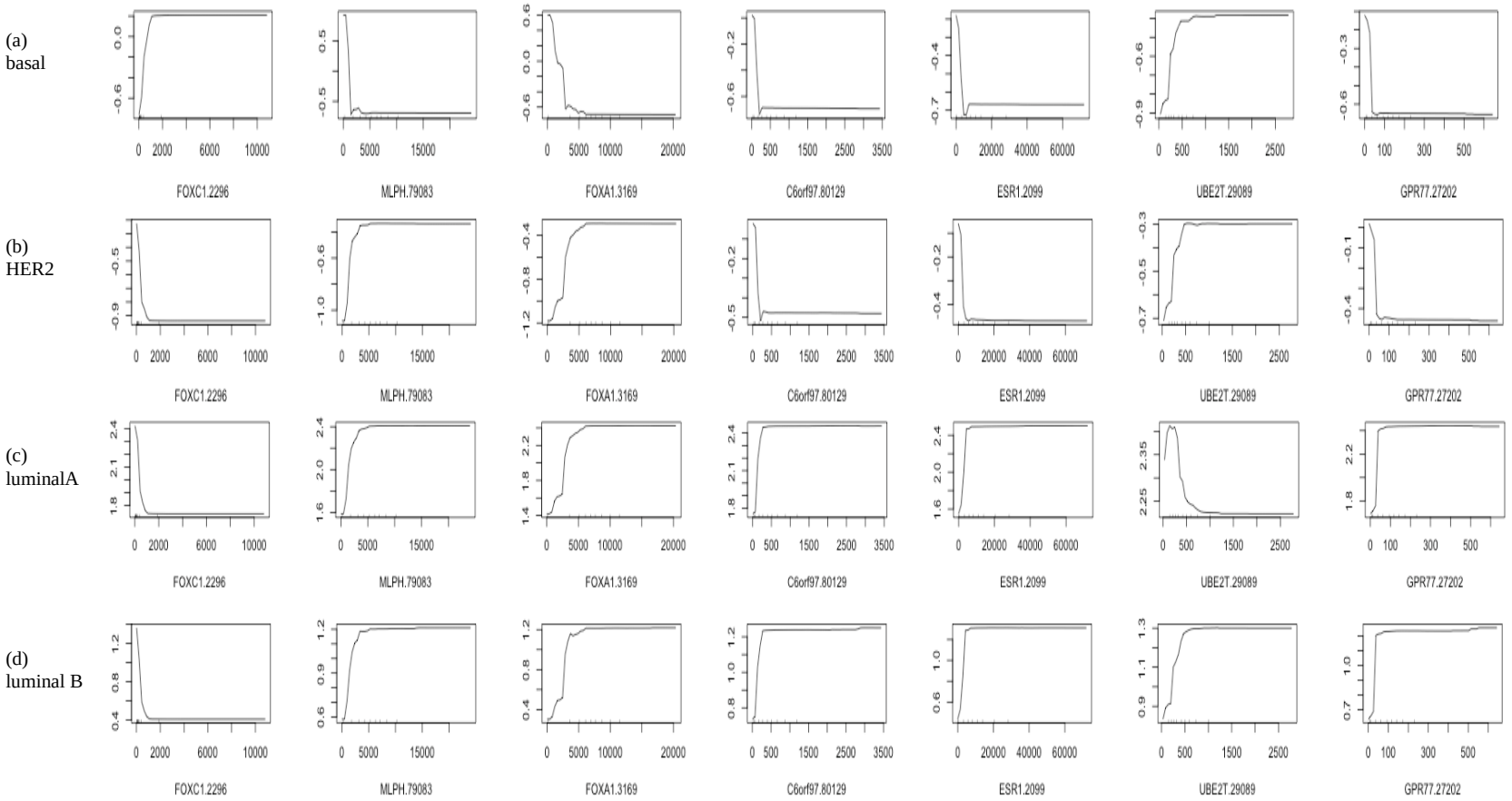


Figure 4-5. Partial dependency plots. Every graph gives a depiction of the marginal effect of a specific variable on the class probability: (a) basal, (b) HER2, (c) luminal A and (d) luminal B. For each of the 7 most important variables (FOXC1, MLPH, FOXA1, C6orf97, ESR1, UBE2T, GPR77) according to the variable important plot shown in Fig. 2 (d).

Table 4-6. Pathways for the seven most important variables [79]

GENE	PATHWAYS	
	SOURCE	NAME
FOXC1	wiki-pathways	Heart Development
		Mesodermal Commitment Pathway
MLPH	wiki-pathways	Deregulation of Rab and Rab Effector Genes in Bladder Cancer
FOXA1	Netpath	AndrogenReceptor
	wiki-pathways	Prostate Cancer
		Endoderm Differentiation
	pid	FOXA1 transcription factor network
		FOXA2 and FOXA3 transcription factor networks
		Direct p53 effectors
C6orf97		Null. Funtion: The function of this gene and its encoded protein is not known. Several genome-wide association studies have implicated the region around this gene to be involved in breast cancer and bone mineral density, but no link to this specific gene has been found
UBE2T	pid	Fanconi anemia pathway
	kegg	Fanconi anemia pathway - Homo sapiens (human)
	wiki-pathways	Gastric cancer network 2
	reactome	Fanconi Anemia pathway
		DNA Repair
	humancyc	protein ubiquitylation
GPR77	wiki-pathways	GPCRs, Other
		Human Complement System
		GPCRs, Class A Rhodopsin-like
	reactome	Signal Transduction
		Signaling by GPCR
		Class A/1 (Rhodopsin-like receptors)
		Peptide ligand-binding receptors
		GPCR ligand binding
ESR1	biocarta	carm1 and regulation of the estrogen receptor
		Estrogen responsive protein efp controls cell cycle and breast tumors growth
		role of erbb2 in signal transduction and oncology
		downregulated of mta-3 in er-negative breast tumors
		pelp1 modulation of estrogen receptor activity
		overview of telomerase protein component gene htert transcriptional regulation
	smpdb	Tamoxifen Action Pathway
		Tamoxifen Metabolism Pathway

GENE	PATHWAYS	
	SOURCE	NAME
ESR1	kegg	Estrogen signaling pathway - Homo sapiens (human)
		Prolactin signaling pathway - Homo sapiens (human)
		Thyroid hormone signaling pathway - Homo sapiens (human)
		Endocrine and other factor-regulated calcium reabsorption - Homo sapiens (human)
		Proteoglycans in cancer - Homo sapiens (human)
	reactome	Signaling by ERBB4
		Nuclear signaling by ERBB4
		Signal Transduction
		Generic Transcription Pathway
		Nuclear Receptor transcription pathway
		Gene Expression
	pharmgkb	Aromatase Inhibitor Pathway (Breast Cell), Pharmacodynamics
	pid	AP-1 transcription factor network
		ATF-2 transcription factor network
		Plasma membrane estrogen receptor signaling
		Validated nuclear estrogen receptor alpha network
		FOXM1 transcription factor network
		Signaling events mediated by HDAC Class II
		FOXA1 transcription factor network
		LKB1 signaling events
		Signaling mediated by p38-alpha and p38-beta
		Regulation of nuclear SMAD2/3 signaling
		Regulation of Telomerase
	wiki-pathways	Nuclear Receptors
		Integrated Breast Cancer Pathway
		miR-targeted genes in muscle cell - TarBase
		Leptin signaling pathway
		Integrated Pancreatic Cancer Pathway
		Aryl Hydrocarbon Receptor
		JAK-STAT
		Estrogen Receptor Pathway
		Nuclear Receptors Meta-Pathway
		Estrogen signaling pathway
	netpath	AndrogenReceptor
		Leptin
		Prolactin
		TGF_beta_Receptor

An enrichment analysis was implemented as a second mechanism to explore the biological meaning of our results. This was executed using the enrichKEGG function in R with the parameter specified in section 42. The enrichment analysis attempts to identify a set of genes within known gene-groups by their functional or pathway categories to interpret their possible biological impact. This analysis found 97 genes out of the 247 features analyzed to be associated with 167 pathways (Appendix 1) when the cutoff values for the statistical parameters (pvalue, qvalue) were set to 1 to detect all represented pathways in our features.

Figure 4-6 shows a total of 13 pathways that contain the largest gene sets and most statistically represented groups when adjusting p-values< 0.1 and q-value < 0.1 as our cutoff thresholds. The "Pathway in cancer" resulted as the most statistically enriched pathway (p-value<0.05) exhibiting 15 genes (FGF2, IGF1R, CCNE2, ADCY4, GSTP1, AR, ERBB2, ADCY9, BCL2, CCNE1, EGFR, CCND1, FZD10, STAT5A, ZBTB16). Pathway in cancer was followed by "Oocyte meiosis" with 10 represented genes, "Proteoglycans in cancer" and "Prostate cancer" both with nine genes.

In other hand, to extend the biological meaning of our results, we used the enrichGO function from ClusterProfiler R package with the parameters specified in section 4.3. This analysis found that our 247 genes in study have protein, enzyme and identical protein binding (see Table 4-7), leading us to think that its biological impact can be strong. These results validate the efficacy of our integrative model to extract important variables which can help us understand cancer behavior.

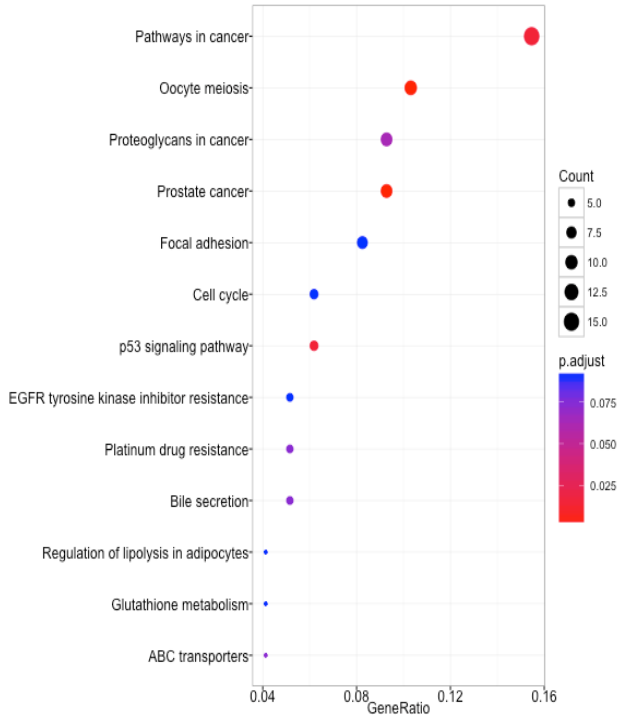


Figure 4-6. Dot plot enrichment analysis. This plot visualizes the pathways (x axis) that contain the largest genes number, from 247 relevant genes. The color dots in the plot are related on their corresponding p-values and the dot sizes is based on the number of genes. The largest is the one that contains the highest amount of genes [78].

Table 4-7. Gene Ontology (GO) analysis for the 247 relevant genes [80]

ID	Description	GeneRatio	BgRatio	P-value	p.adjust	q-value
GO:0003674	molecular_function	203/203	16619/18679	4.35E-11	2.30E-08	1.89E-08
GO:0005488	binding	180/203	13927/18679	4.16E-07	1.10E-04	9.05E-05
GO:0005515	protein binding	144/203	10431/18679	6.36E-06	1.12E-03	9.22E-04
GO:0019899	enzyme binding	37/203	1636/18679	1.44E-05	1.90E-03	1.56E-03
GO:0042802	identical protein binding	28/203	1171/18679	7.09E-05	7.49E-03	6.17E-03
GO:0004716	receptor signaling protein tyrosine kinase activity	3/203	9/18679	1.01E-04	8.91E-03	7.34E-03

4.4.4 Further Gene-Set Validation

For validation purposes, we used two external datasets: GSE20685 and GSE21653 where 211 out of the 247 features extracted in the integrative step were found in these datasets. There exist certain experimental limitations that do not allow for whole-gene profiling as is the case of these two datasets. We evaluated the classifying performance of these 211 variables using the external GEO datasets and Random Forest classifier as described in the methodology (section 4.3.53.3).

With these 211 relevant variables we achieved good performance metrics to distinguish between subtypes. The results from GSE20685 and GSE21653 yielded high AUC values: 94.69% and 84.92 %, as well as low error rates: 12.23% and 15.79% for each (see Table 4-8). Note that on GSE20685 data, the error rate for the normal subtype appears as NA since there were no samples available for this subtype. These results allow us to corroborate that the 211 variables selected in our integrative model can discriminate very well between each breast cancer subtypes across external datasets with comparable results to those obtained earlier with the TCGA data.

Table 4-8. Evaluations of error rate for GSE Data

Dataset	# Samples	AUC	Out-Of-Bag Error Rate (%)					
			Overall Error	Basal	HER2	luminal A	luminal B	Normal
TCGA	547	84.83	8.58	3.09	15.68	1.89	14.59	91.66
GSE21653	266	84.92	15.79	6.67	33.33	5.62	22.45	44.83
GSE20685	327	94.69	12.23	0.00	12.00	7.53	15.57	NA*

*NA: Not Available

To validate specific patterns found with the integrative model in these two external datasets we observed the overlap of the most important variables among all three datasets (original TCGA and the two external ones). Figure 4-7, shows the top 30 important

variables by RFs mean decrease in Gini Score for GSE21653, GSE20685 and TCGA model integrative. We established a threshold (see red lines in Figure 4-7) in each dataset to select the most important among those thirty resulting in 17 and 18 relevant genes for GSE21653 and GSE20685 respectively. Lastly, we compared the overlap among these gene sets and the five top genes from TCGA which the union of the three sets yielded a 25-gene set. From the five genes already established as important for TCGA, we found all to be important in the GSE21653 and GSE20685 except ESR1 which was not found in in the top 18 of GSE20685.

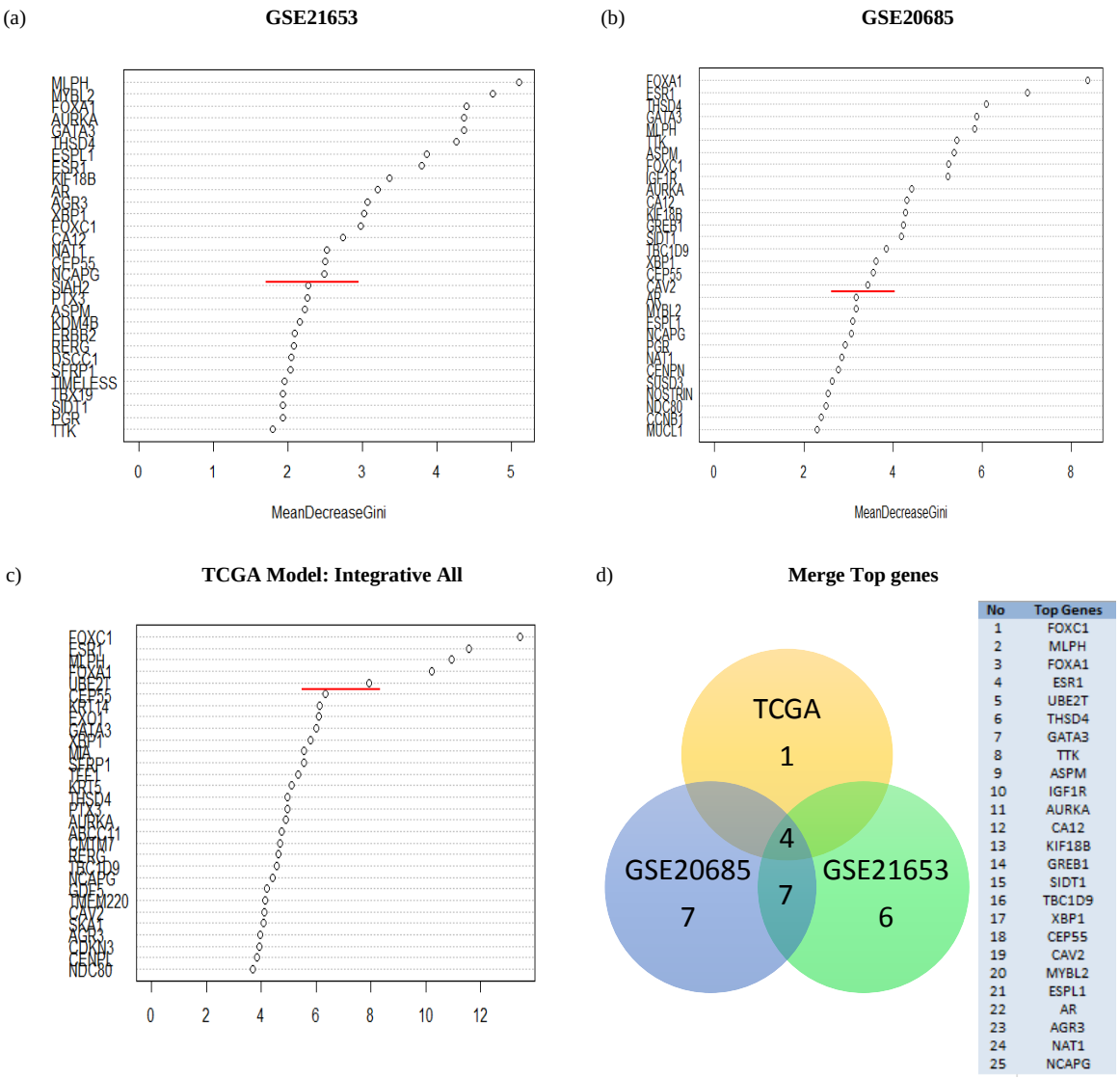


Figure 4-7. Variable importance plots. This plot visualizes the top thirty important variables by mean decrease in Gini Score that RFs picked for: (a) GSE21653 dataset with 17 top genes above red line, (b) GSE20685 with 18 top genes above red line, (c) TCGA all integrative important variables model with 5 top genes above red line, and (d) merge top genes for TCGA, GSE21653, GSE20685 data: 25 results as important.

To visualize the specific patterns of the resulting 25-gene set among all three datasets, we created a series of heatmaps to validate the expression patterns found to be key to differentiate breast cancer subtypes. Figure 4-8 shows the heatmaps for these 25 genes

across the TCGA, GSE20685, and GSE21653 datasets where the rows of each heatmap image correspond to genes and columns correspond to samples. In this figure, we can visualize the same gene expression patterns over all genes across all three datasets. For example, the basal subtype is clearly defined by the block of genes with red color: FOXA1, SIDT1, AR, THSD4, GREB1, MLPH, TBCID9, CA12, GATA3, XBP1, ESR1, NAT1, AGR3, these in average, have a scaled value below the norm (less expressed). Similarly, luminal A subtype is clearly defined by the block of genes with red color (negative values): ESPL1, KIF18B, NCAPG, TTK, CEP55, UBE2T, ASPM, AURKA, and MYBL2.

These results strongly support the findings of our integrative model and we can conclude that the selected variables can effectively discriminate breast cancer subtypes. Furthermore, we found an overlap between 247 selected variables and known cancer gene lists: CCGD, Cosmic and Vogelstein, revealed 6, 13 and 75 genes, from 255, 594 and 7088 respectively for each list. AR, BCL2, EGFR, ERBB2 and CCND1 were common genes in all lists. Similarly, we revised if these 247 selected variables contained the original 50-gene list signature previously constructed in the PAM50 with an overlap of 26 genes out of the 50 genes associated with this signature.

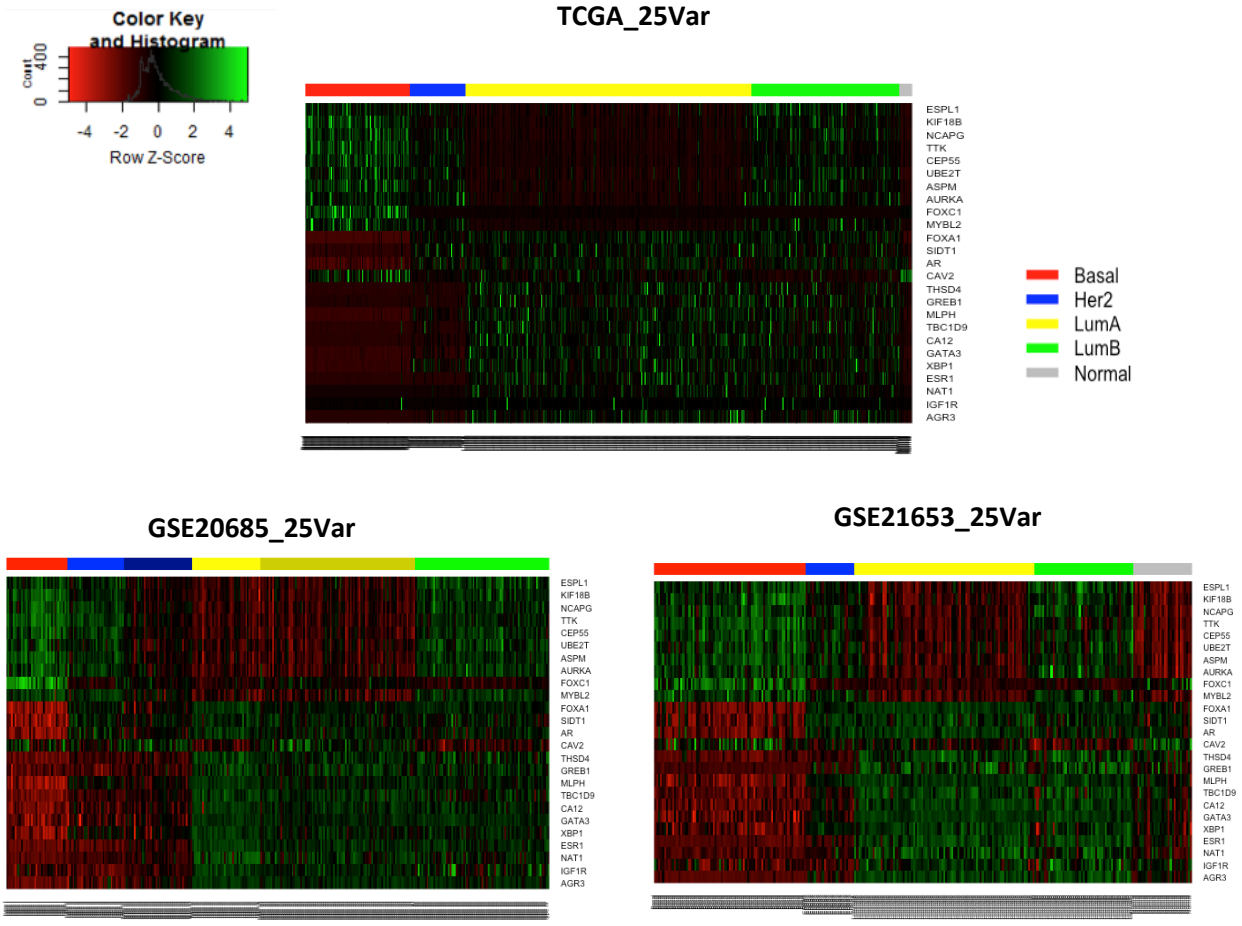


Figure 4-8. Heat map plots. This plot visualizes the heat map of the chosen 25 top important variables for (a) TCGA, (b) GSE20685, and (c) GSE21653 datasets. Rows of each heat map correspond to genes, ordered according to hierarchical clustering from GSE20685 dataset. Columns of each heat map correspond to samples, ordered by breast cancer subtype. The color of pixel indicates the expression of one gene, where red means low expression and green means high expression

Finally, we search for relevant information (i.e. related pathways, associated diseases and supporting literature articles) of 73 common genes in the three evaluated datasets during the interaction studies in Phase 3 (see Chapter 5); 15 of those are listed in **Error! Reference source not found.** (see Appendix 2 for the entire list). According to the PubMed engine search database, we found that 42 out of the 73 genes (57.53%) have

more than six published articles related to breast cancer. For instance, the androgen receptor gene, AR, has over 2000 published papers associated with breast cancer. Also, eight genes (FOXA1, MUCL1, GREB1, TFF1, ESR1, AR, BCL2, GRB7) are directly associated with breast cancer according to GeneCard database [85]. This result supports the sensitivity of our methodology to detect genes that are currently known to play a key role in breast cancer. Furthermore, we found genes with little or no publishable track based on our PubMed search. Nine of those genes (CENPL, RERGL, TBX19, KCMF1, ADCY4, NOSTRIN, CMTM7, SCCPDH and DSCC1) did not show at all in our search, whereas 22 of them have between 1 and 5 published studies linked to breast cancer as of May 16, 2017 PubMed search hits [86]. These genes are clearly strong candidates for more in depth explorations of their implications to the disease we studied in this work.

Table 4-9. Top fifteen genes biological insights

Gene	Search Engine Source			
	GeneCards [85]			PubMed [86]
	Related pathways	Associated diseases	Associated with breast cancer?	* Number of published scientific articles
MLPH	Deregulation of Rab and Rab Effector Genes in Bladder Cancer	Griscelli Syndrome, Type 3 and Osteogenesis Imperfecta, Type Xv.	No	3
FOXA1	Embryonic and Induced Pluripotent Stem Cell Differentiation Pathways and Lineage-specific Markers and FOXA1 transcription factor network.	Estrogen-Receptor Positive Breast Cancer and Luminal Breast Carcinoma.	Yes	221
SIDT1	No data available	No data available	No	1
CEP55	Cytoskeletal Signaling and DNA Damage.	No data available	No	7
ASPM	No data available	Microcephaly 5, Primary, Autosomal Recessive and Autosomal Recessive Primary Microcephaly. upregulated in several types of cancer: in particular, brain tumors.	No. But associated with cancer	8
CENPL	Mitotic Metaphase and Anaphase and Cell Cycle, Mitotic	Seckel Syndrome 1	No	0
AURKA	Integrated Breast Cancer Pathway and Regulation of PLK1 Activity at G2/M Transition	Colorectal Cancer and Colorectal Adenocarcinoma	No. But associated with cancer	177
ESPL1	Mitotic Metaphase and Anaphase and Cell Cycle, Mitotic.	Fallopian Tube Disease and Salpingitis.	No	9
TTK	**RB in Cancer** and DNA Damage.	Chronic Polyneuropathy.	No. But associated with cancer	70
UBE2T	Fanconi anemia pathway and Metabolism of proteins	Fanconi Anemia, Complementation Group T and Ube2t-Related Fanconi Anemia.	No	4
NCAPG	Cell cycle_Chromosome condensation in prometaphase and Aurora B signaling	No data available	No	1
GMPS	Metabolism and purine nucleotides de novo biosynthesis	Leukemia, Acute Myeloid	No	7
NDC80	Mitotic Metaphase and Anaphase and Aurora B signaling	Female Reproductive Organ Cancer.	No	9
MYBL2	HTLV-I infection and EGFR1 Signaling Pathway	Paraneoplastic Cerebellar Degeneration.	No	38
KIF18B	Vesicle-mediated transport and Factors involved in megakaryocyte development and platelet production	No data available	No	1

* Entry Search in PubMed page as follows: “name of gene AND breast cancer”

4.5 Conclusions

We were able to computationally integrate multiple heterogeneous and highly dimensional datasets to discriminate breast cancer subtypes by reducing the feature space through feature selection techniques. The feature selection methods evaluated in this work exhibited high predictive performance (AUC: ~85.9%, accuracy: ~90.6%) in the selection of important omics variables. Here, we found that7 out of 9 times CFS outperformedall other assessed methods (i.e. Information Gain, ReliefF, SVM-RFE and FAST Clustering based) in terms of accuracy and AUC. This method was the mostappropriate for extracting significant features from gene expression, protein and methylation data.

The extracted features yielded main encouraging results with useful biological meaning. First, the integrated model revealed that gene expression variables were more important to predict breast cancer subtypes than protein and methylation. Historically, subtypes have been defined by mRNA expression; therefore it is not surprising that features from RNA-seq were the most significant. Among the top selected features, the best ranks belong to following genes: FOXC1, MLPH, FOXA1, C6orf97, ESR1, UBE2T, and GPR77. These results agree with those obtained by List et. al.[43], where in their integrated model of gene expression with methylation, found that gene expression variables were superior in the combined model. They found the following top genes: ESR1, FOXA1, MLPH and FOXC1, which our model extracted as important as well. Also, our integrative model was able to detect cg02643667 and ER.alpha to have a critical role in breast cancer subtype classification. These were ranked in eighth and

twelfth place respectively. This inclusion seems to slightly improve the sensitivity and specificity of the model as gathered through the AUC. In previous studies cg02643667 methylation probe had strong implications in research related with breast cancer. Dedeurwaerder *et al* [87] established this methylation as part of a set of 86 CpGs found highly associated with breast tumors prognostics. Additional research provided a target gene set where this methylation was included for prediction, prognosis, diagnosis and therapy of breast cancer[88]. List et. al. [43] also found to the methylation probe, cg02643667 (TTF1), but only in the top most important features of their methylation model. Hence, features from protein and methylation should be further explored and future studies should investigate the interaction of these important variables in depth. Finally, our integration model yielded slightly higher accuracy (~91%) and less number of features used (211 genes) when compared to List et al model (~88% accuracy; 275 genes) [43].

Second, in the enrichment analysis we found 97 genes out of the 247 features analyzed to be associated with 167 pathways, and the "Pathway in cancer" resulted as the most statistically enriched pathway (p-value<0.05) exhibiting 15 genes (FGF2, IGF1R, CCNE2, ADCY4, GSTP1, AR, ERBB2, ADCY9, BCL2, CCNE1, EGFR, CCND1, FZD10, STAT5A, ZBTB16). Also, we found that our 247 genes have protein, enzyme and identical protein binding, leading us to think that its biological impact can be strong.

On other hand, despite of the inclusion of protein and methylation data improve our integrative model, even the error rate is not small enough (8.94%). We consider, that this error can be explained by the unbalanced nature of the subtype information available in

the datasets used. About 40% of this error was concentrated in normal and HER2 subtypes which together contribute only 11.22% of the all samples. The model also has problems in classifying between luminal A and luminal B patients. In reality, these two types tend to be transitional subtypes therefore they are more difficult to differentiate. This can be seen in the similarity on its transcriptomic patterns as shown in the partial dependence plots in Figure 4-5 in which all genes behave similarly, except for UBET2 for which the expression behavior is opposite for luminal A (identify less expressed) and luminal B (identify highly expressed).

Lastly, nine genes (CENPL, RERGL, TBX19, KCMF1, ADCY4, NOSTRIN, CMTM7, SCCPDH and DSCC1) extracted as important by our model do not have any reported literature related to breast cancer based on our search in the PubMed database. Therefore, it is imperative to study their biological impact in breast cancer further. These results validate the efficacy of our integrative model to extract important variables based on their genomic characterization which can help us to understand the cancer behavior.

5 MEASURING INTERACTIONS IMPORTANCE USING RANDOM FOREST

5.1 Introduction

Breast cancer is a heterogeneous disease and detecting interactions patterns that could lead to new understandings of biological mechanisms in cancer is of great need. Detecting gene interactions is a difficult problem due to dimensionality of genomic data and the infinite number of possibilities. This creates a computational barrier to evaluate all possible interactions at various degrees for whole genome information which increases drastically as the number of genes increases. Nonetheless, there exists several techniques used to detect interactions; these ranged from traditional linear models to more computationally complex machine learning methods. Some of the most renowned machine learning techniques proven to perform well in detecting gene-gene interactions are Neural Networks (NNs) [54]–[56], Support Vector Machine (SVM)[28], [89]–[91], and Random Forests (RFs) [56].

Many of these heuristics methods have the ability to classify complex classification problems. For example, in the case of NNs, the method focuses on mimicking the brain's ability to solve problems by connecting large number of neurons [92]. Though NNs have done well in certain applications its black box nature and computational load makes it unattractive for application with biological data which aim to uncover new biological meaning from the model. SVM is another extensively studied model that achieves high performance metrics (i.e. accuracy, AUC) using hyperplanes and non-probabilistic binary

linear classifier. SVM have a proven record to work very well to classify complex biological data and in most cases its response is more interpretable if compared to other methods such as multifactor dimensionality reduction (MDS). However the output of SVM can be affected when working with genetic heterogeneity [9]. In another hand, RFs is very attractive to study gene interactions since they have several intrinsic characteristics that fit very well with the requirements of molecular dataset [56]. This type of ensemble model can model diverse types of variables with no restrictions on distributional assumptions using a nonlinear approach that can act as a feature selection mechanism to reduce complexity. Moreover, RFs can be highly interpretable because it can rank the most significant features through the estimation of Variable Importance Measures (VIM) and can evaluate the marginal effect of a feature in a given class through the partial dependency plots (PDPs). Because of these advantages and the capacity of modeling over different random subsets created for each tree [93], it becomes a solid candidate to discover interactions at the molecular level.

Therefore, this work uses the ensemble methodology of random forest to model breast cancer subtype across thousands of gene expression profiles to focus on measuring those detected interactions using a new metric. As discussed earlier, there are several methods to detect interactions but not many include metrics to assess which of those interactions are most important. The assessment of those interactions is critical to interpret its biological meaning, expand current knowledge, and design further experiments to validate the effects of those significant patterns. To the best of our knowledge, there are not many metrics to measure interaction. RFs implementations results in metrics focusing

solely in the marginal contribution of one feature even though its model structure is highly interactive. When two features or variables have many levels it is computationally demanding to calculate all possible rules with its integrative contributions in the model. Jones and Linder extended the implementation of partial dependency plots from marginal contribution of a feature to estimate the a marginal combination of a pair of features including visualization aid [94]. Nonetheless, to be able to implement their algorithm the user must know in advance which pairs of features to analyze. Therefore, a ranking metric is needed to evaluate promising interactions of higher order.

In this phase, we aim to develop a new metric called Importance Between Features through Random Forest (IBF-RF) capable to assess the most important interactions extracted by RFs. We defined variable interactions as the capacity of variable sets to describe a class, in this case specifically a breast cancer subtype. We expected that if two or more variables have strong links between them, they should be frequently appear in the branches of different trees defining a specific class (subtype).

5.2 Objective

To develop a new metric and implement an algorithm capable of assessing the interaction between relevant features resulting from the integration explored in the second phase of this thesis and implemented using RFs classifier.

5.3 Methodology

We proposed a metric called Interaction Between Features through Random Forest *(IBF-RF)*. The metric to be implemented will measure the prevalence of a set of features through their recurrence in the forest created in using Random Forest methodology. The interaction importance will be assessed based in the recurrence of a branch (set of features) through all the trees of the forest toward a class x (see Figure 5-1).

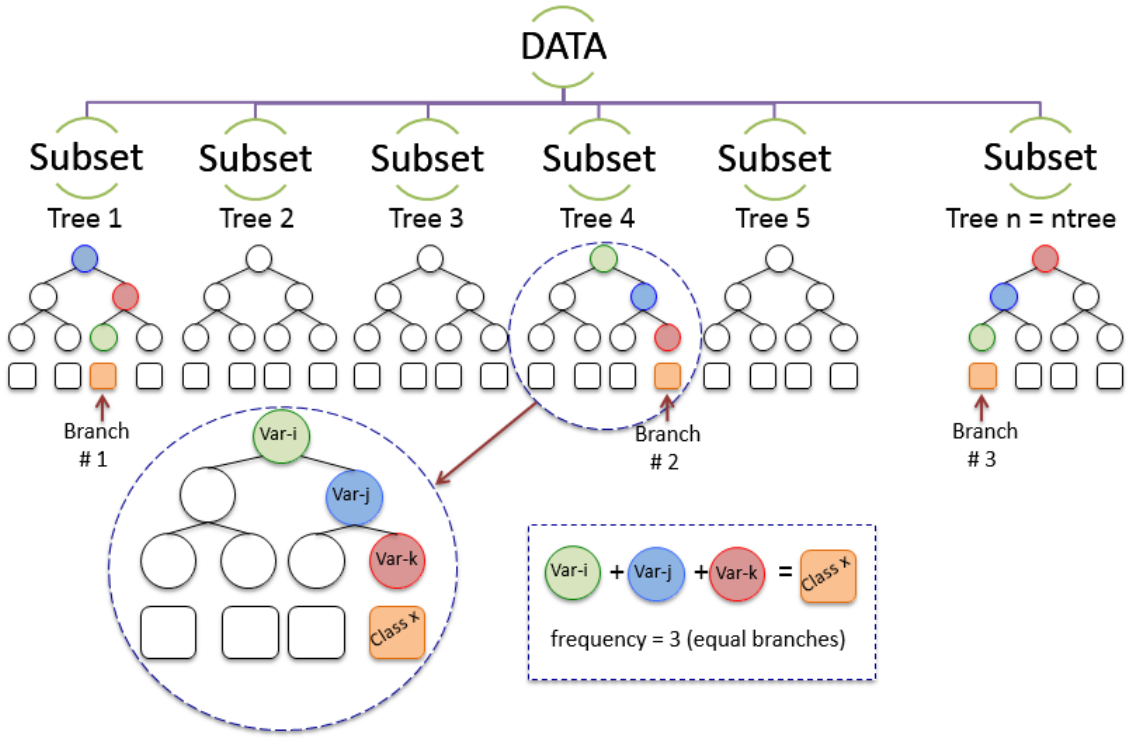


Figure 5-1. Overview of Interaction Between Features through Random Forest (IBF-RF). This plot visualizes one example the recurrence of a branches (set of features) through all the trees of the forest toward a class x. According to the example the *i*, *j*, *k* variables are present (order no matters) in three different trees. The metric counts and ranks the prevalence of a set of features (rules) through their recurrence in the forest created in using Random Forest methodology.

5.3.1 Interaction Between Features through Random Forest (IBF-RF):

IBF-RF metric, allows to assess the interactions between set of features. For achieve this, an algorithm extract the rules constructed in the natural process of Random forest classifier, and ranks the frequency of each rules. The algorithm 1 describes the pseudo code for IBF-RF metric, in which the *BootstrapSampling* function returns a sample that has been taken from **N** variables with replacement of the full set. The returned sample will be used to build the set of decision trees (See section 2.2.3). The function called *BuildsRandomForest* runs a random forest classifier consisting of a set of trees, each constructed on a bootstrap sample set (i.e. samples from *BootstrapSampling*). These trees are grown and each predictive values is averaged across all trees. Later these trees are translated into classification rules for specific classes. Finally, variables and combinations of variables are tallied across all trees to measure its prevalence in the forest model.

Algorithm 1. Pseudo code:

Interaction Between Features through Random Forest (IBF-RF)

Input

$Conjunto\ \mathbb{D} = \{(X_n, Y_n) | \ \boldsymbol{n} = 1, 2, \dots, N, X_n \in feature\ selection\ variables, Y_n \in levels\ response\ variable\}$

$Conjunto\ C_i = 1, 2, \dots, m. \qquad C_i = Set\ of\ features\ (rules)$

$Number\ of\ trees\ K \mid k = 1, 2, \dots, K$

$Number\ of\ rules\ R\ in\ each\ tree\ k \mid r = 1, 2, \dots, R$

$Frequency\ C_i = 0$

$RF_t \coloneqq BuildsRandomForest\ (D_t, K)$

$D_t \coloneqq BootstrapSampling\ (\mathbb{D})$

For $C_i \coloneqq 1$ **to** m **do** $(m = all\ possible\ rules\ )$

  For $k \coloneqq 1$ **to** K **do**

    For r $\coloneqq 1$ **to** R **do**

      if C_i *is equal to rule* R, **then**

        $Frequency\ C_i = 1$

      else

        $Frequency\ C_i = 0$

      End if

      $Total\ frequency\ C_i = Total\ frequency\ C_i + frequency\ C_i$

    End for

  End for

End for

Output: *Ranking* $C_i\ frequency$

In order to extract the rules of a tree, we validated and used an unpublished code shown in Appendix 2 [95]. This code defines three functions: getConds, PrevCond and Collapse. In general, these functions store the rules of each tree, which can be obtained through the getTree function available in the radomForest package [73] of R software. The getTree

function, through parameter *k* (*k= tree to extract*), allows us to extract the rules of a specific tree from a forest. The code presented in Appendix 2, presents outputs on the specific rules generated in tree; a hypothetical example is presented in Figure 5-2 to illustrate its structure.

[[1]][*k*] Var.j<value & Var.i<value => Class.x

[[2]][*k*] Var.j<value & Var.i>value & Var.k<value => Class.x

[[3]][*k*] Var.i>value & Var.j>value & Var.l<value => Class.x

…

[[r]][*k*] Var.l>value & Var.j>value & Var.k<value & Var.m<value => Class.

Figure 5-2. Extracting the rules of a tree. This plot visualizes one example of the rules extracted from a specific tree, through a code available in Stack Overflow (http://stackoverflow.com) (see Appendix 2), this code defines three functions: getConds, PrevCond and Collapse, which store the rules of each tree of forest.

Where,

$k = 1, 2, 3, \dots, K.$ $K = Number\ of\ trees\ (K = ntree)$

$r = 1, 2, 3, \dots, R.$ $R = Number\ of\ rules\ in\ each\ tree\ k$

$\text{Var. i}, \text{Var. j}, \text{Var. k} \in X_n.$ $X_n = feature\ selection\ variables$

$\text{Class. x} \in Y_n.$ $Y_n = levels\ response\ variable$

valor $= \forall\ levels\ of\ X_n.$

Once all rules are extracted then each rule was split into their individual factors of information to be stored for later use as shown in
Figure 5-3a. Then, all variables that belong to the same tree and to the same rule are compiled. This variable set is checked to eliminate the existence of duplicates and are concatenated alphabetically. Finally, this set of ordered variables are stored is a new table (see

Figure 5-3b), which also indicates the number of the tree and the prediction of rule. With this table the frequency per set of genes per subtype will be extracted, which will be ordered from highest to lowest to identify the ones with the highest recurrence. The general output of metric looks like
Figure 5-3c, and the code is present in Appendix 3.

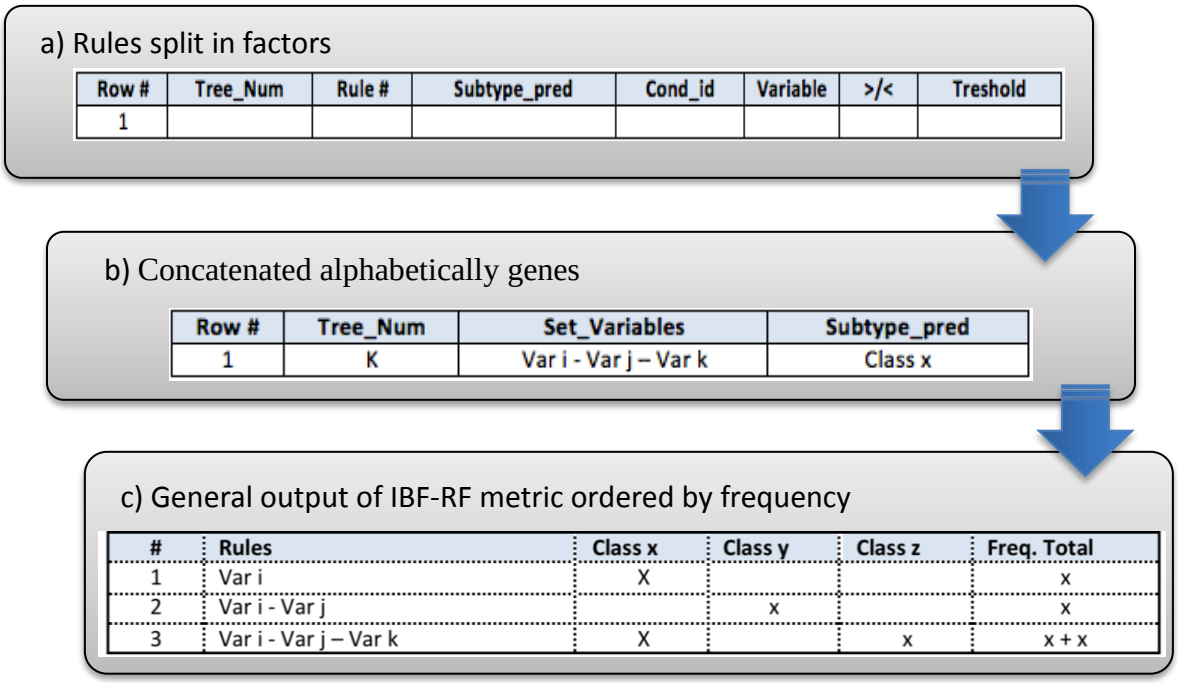


Figure 5-3. Step diagram of IBF-RF metric. This diagram visualizes the outputs at every steps of the IBF-RF metric: a) first the rules are splits in factors, b) the genes are concatenated alphabetically, and c) the frequency total of the rules is calculated.

5.4 Results and Discussion

To assess the interaction through IBF-RF metric, we used the 211 important genes found in the integrative model from Phase 2. This metric was implemented across three gene expression datasets TCGA, GSE20685 and GSE21653 (see Chapter 4). Random forest was implemented and its parameters were tuned as follows. The number of trees were estimated through initial tests until error estimation was stable, yielding 5000, 17000 and

9000 trees in their ensemble for TCGA, GSE20685 and GSE21653 data respectively. We observed that the run time of IBF-RF metric increases as the number of trees used to build the forest (ntree) increases. Intuitively, the larger the number of trees considered, the longer it took to complete the subroutine as shown in Table 5-1. Due to the large number of trees (ntree=17000) necessary to reach the desired stability of GSE20685 the original implementation strategy was not possible. Hence, we ran the metric algorithm in a parallel scheme, meaning that it was necessary to run by parts the IBF-RF metric (each 1000 trees) without loss of information. This parallel implementation took about 100 hours to run completely.

IBF-RF metric allowed us to extract a total of 154312, 190481 and 463917 rules for TCGA, GSE20685 and GSE21653 data respectively. All rules were ordered from highest to lowest frequency to identify those with the highest recurrence (see Appendix 4 for frequent rules in each one datasets).

Table 5-1. Evaluations of IBF-RF metric using three datasets

Description	Datasets		
	TCGA	GSE20685	GSE21653
Ntree	5000	17000	9000
Run time (hr)	12	100	48
Rules extracted	154312	463917	190481

The extracted results in TCGA data shows MLPH and FOXA1 rules as most important, being only one gene sufficient to differentiate between subtypes. Both genes clearly discriminated the basal subtype found in the forest with the highest frequencies: 237 and 109 times respectively. Similarly, GSE20685 data shows a rule conformed by only one

gene (FOXC1) as the most important predicting the basal subtype with 107 frequencies. For both datasets (TCGA and GSE20685), besides of their top rules, we found several rules conformed by two genes. In the case of GSE21653 data the tops rules were generated by two genes that lead the prediction of basal and normal subtypes. The first rule with a single gene, FOXA1, appears in the thirteenth position to predict the basal subtype validating the results from TCGA and GSE20685. An initial overview, allowed us to observe that in at least 2 out of 3 databases, the rules with highest frequency are those of second and third order (i.e. 2 and 3 genes in a rule) (See Table 5-2). Although we obtained rules of higher order (more than 4 genes in rule) that were less frequent, they are still important because they enable the prediction of basal subtype as well as the other types of breast cancer.

Table 5-2. Frequency according rules order

Rules order	Total frequency
1	9
2	523
3	1211
4	331
5	29
6	4
Total rules	2107

We hypothesized that those results give us important rules to discriminate the breast cancer subtypes and if they are valid they must be found in all three databases (TCGA, GSE20685, and GSE21653). Consequently, we found the most common rules between these datasets, resulted in 156 common rules where the top 20 of those are listed in Table

5-3 (See Appendix 5 for the entire list). Other rules were found important but only in two out of three datasets where the most rules in common were between both GSE datasets with 1438 rules in common, followed by 597 rules between TCGA and GSE20685 datasets, and lastly 384 rules between TCGA and GSE21653 datasets as shown in Figure 5-4.

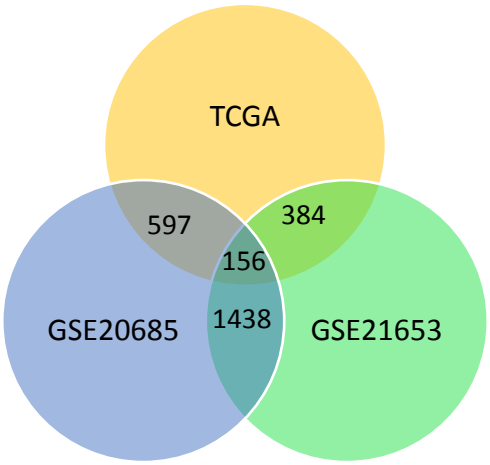
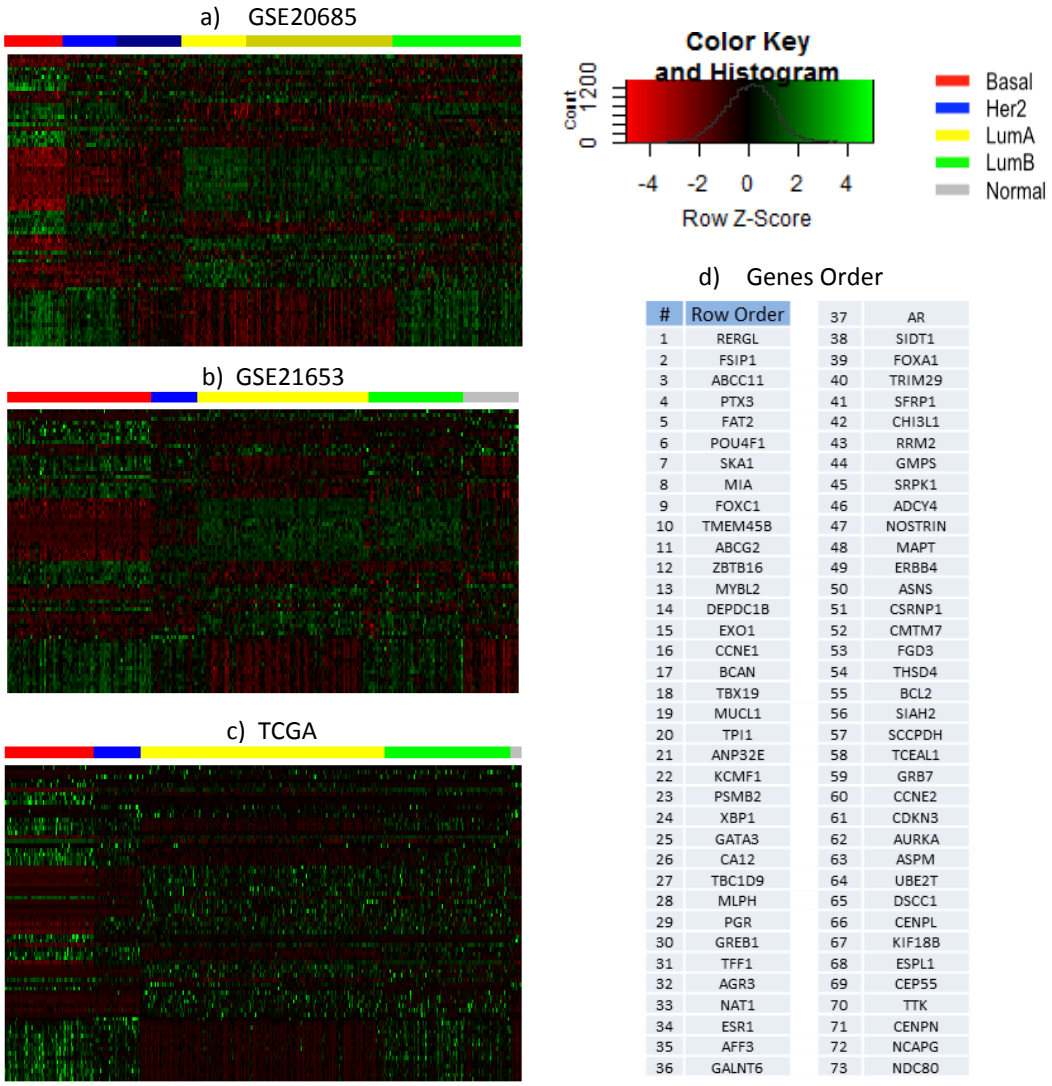


Figure 5-4. Venn diagram: common rules. Venn diagram visualizes the number of common (intersections) rules between TCGA, GSE20685 and GSE21653 datasets, extracted through IBF-RF metric. 156 rules are common in all three datasets, 597 rules between TCGA and GSE20685 datasets, 384 rules between TCGA and GSE21653 datasets, and lastly 1438 rules are common between GSE21653 and GSE20685 datasets.

Table 5-3. Top 20 common rules between TCGA, GSE21653 and GSE 20685 datasets, extracted through IBF-RF metric.

#	Rules	TCGA						GSE21653						GSE20685							Grand Total
														basal	HER2		LumB	LumA		Total	
		Basal	HER2	LumA	LumB	normal	Total	Basal	HER2	LumA	LumB	normal	Total	typeI	typeII	typeIII	typeIV	typeV	typeVI		
1	MLPH	237					237	8					8	56						56	301
2	FOXA1	109					109	22					22	8						8	139
3	CEP55-FOXA1	3					3	35				19	54	41		14				55	112
4	FOXC1-THSD4	11					11					2	2	76		3			16	95	108
5	FOXA1-TTK	9				2	11	25				20	45	38		8				46	102
6	MLPH-NOSTRIN	1					1	2					2	54	2	28	3			87	90
7	ASPM-FOXA1	2					2	27				20	47	33		6				39	88
8	CENPL-FOXA1	5					5	6				5	11	40		31				71	87
9	AURKA-FOXA1	4				2	6	34				9	43	18		6				24	73
10	MLPH-TTK	2			1		3	10				15	25	29		16				45	73
11	ESPL1-FOXA1	7				5	12	25				8	33	10		8				18	63
12	ASPM-FOXC1	2		1			3	5				5	10	36					10	46	59
13	FOXA1-GMPS	5				5	10	9					9	31		7				38	57
14	CEP55-MLPH	4			1		5	12				12	24	16		9				25	54
15	FOXC1-KIF18B	6		4			10	2				2	4	31					7	38	52
16	FOXC1-NOSTRIN	1					1	2					2	35		1			11	47	50
17	SIDT1	1					1	14					14	35						35	50
18	FOXA1-UBE2T	9				7	16	11				6	17	10		6				16	49
19	FOXA1-NCAPG	4				3	7	24				11	35	2		3				5	47
20	FOXA1-NDC80	2					2	14					14	26	1	2	2			31	47

In these 156 common rules, we can see MLPH and FOXA1 rules at the top list focusing in discriminating basals from other subtypes. Nonetheless, several rules conformed by two genes appeared quickly after the most frequent rules which consisted on single genes. Many of these rules included MLPH or FOXA1 with other genes as an interacting pattern with specific expression regions characterizing many subtypes differently (See Table 5-4). These 156 common rules generated in all three datasets are made up by the combination of 73 genes in total. Their expression behavior across samples was depicted using a series of heatmaps across the different five subtypes as shown in Figure 5-5. In these heatmaps, we can validate the predictive importance of the 156 common rules since the expression patterns of those 73 genes are similar across the two validation datasets. For example, MLPH and FOXA1 are very significant to differentiate basal subtype because of their distinct expression levels which are always less expressed (bright red) for the basals than for any other subtype. Additionally, we see that the third rule: CEP55-FOXA1, characterizes three subtypes: basal, normal and HER2 with distinguishable combinations of expression Table 5-4). In Figure 5-5, we find that these predictions are given using normalized expression scores as follows: basal occurs when CEP55 is expressed between 3 to 6 (bright green) and FOXA1 between -5 to -3 (bright red). But for normal samples, CEP55 is expressed between -5 to -3 (bright red) and FOXA1 between 0 to -3 (opaque green) while HER2 occurs when both CEP55 and FOXA1 are expressed between -3 to 0 (opaque red).



d) Genes Order

#	Row Order	#	Row Order
1	RERGL	37	AR
2	FSIP1	38	SIDT1
3	ABCC11	39	FOXA1
4	PTX3	40	TRIM29
5	FAT2	41	SFRP1
6	POU4F1	42	CHI3L1
7	SKA1	43	RRM2
8	MIA	44	GMPS
9	FOXC1	45	SRPK1
10	TMEM45B	46	ADCY4
11	ABCG2	47	NOSTRIN
12	ZBTB16	48	MAPT
13	MYBL2	49	ERBB4
14	DEPDC1B	50	ASNS
15	EXO1	51	CSRNP1
16	CCNE1	52	CMTM7
17	BCAN	53	FGD3
18	TBX19	54	THSD4
19	MUCL1	55	BCL2
20	TPI1	56	SIAH2
21	ANP32E	57	SCCPDH
22	KCMF1	58	TCEAL1
23	PSMB2	59	GRB7
24	XBP1	60	CCNE2
25	GATA3	61	CDKN3
26	CA12	62	AURKA
27	TBC1D9	63	ASPM
28	MLPH	64	UBE2T
29	PGR	65	DSCC1
30	GREB1	66	CENPL
31	TFF1	67	KIF18B
32	AGR3	68	ESPL1
33	NAT1	69	CEP55
34	ESR1	70	TTK
35	AFF3	71	CENPN
36	GALNT6	72	NCAPG
		73	NDC80

Figure 5-5. Heat map plots. This plot visualizes the heat map for: (a) GSE20685, (b) GSE21653 and (c) TCGA datasets. Rows of each heat map correspond to 73 genes (d) that make up the 156 common rules between all three datasets extracted through IBF-RF metric. Columns of each heat map correspond to samples, ordered by breast cancer subtype.

In Table 5-4 we observe two things: 1) two set of genes (i.e. blocks) have similar behavior (see Table 5-5), and 2) the rules are formed by the combination between blocks and not within same block. To corroborate the similar behavior, we studied the

correlation of two blocks of genes, the first formed by: FOXA1, MLPH and SIDT1 genes, and the second by the genes: CEP55, ASPM, CENPL, AURKA, ESPL1, TTK, UBE2T, NCAPG, GMPS, NDC80, MYBL2, KIF18B and EXO1 as shown in Table 5-5. These genes were plotted in Figure 5-6 over all three datasets (i.e. TCGA, GSE21653 and GSE20685) showing evident expression differences across subtypes. The first block is evidently less expressed across all basal samples with a significant change in values when compared with other subtypes. Similarly, the second group had a very similar behavior among themselves but showing less expression in luminal A subtype than the other subtypes.

Table 5-4. Interpretation of heatmap plots

# Rule	Gene	Basal	Her2	LumA	LumB	Normal
2	FOXA1					
3	FOXA1 - CEP55					
5	FOXA1 - TTK					
7	FOXA1 - ASPM					
8	FOXA1 - CENPL					
9	FOXA1 - AURKA					
11	FOXA1 - ESPL1					
13	FOXA1 - GMPS					
18	FOXA1 - UBE2T					
19	FOXA1 - NCAPG					
29	FOXA1 - KIF18B					
23	FOXA1 - EXO1					
20	FOXA1 - NDC80					
25	FOXA1 - MYBL2					
26	FOXA1 - ZBTB16					
1	MLPH					
14	MLPH - CEP55					
10	MLPH - TTK					
21	MLPH - CENPL					
30	MLPH - AURKA					
18	MLPH - UBE2T					
24	MLPH - KIF18B					
28	MLPH - EXO1					
6	MLPH - NOSTRIN					
4	FOXC1 - THSD4					
22	FOXC1 - FOXA1					
22	FOXC1 - AR					
16	FOXC1 - NOSTRIN					
12	FOXC1 - ASPM					
15	FOXC1 - KIF18B					
32	FOXC1 - AURKA					
17	SIDT1					

Table 5-5. Blocks of highly correlated genes

BLOCK 1 GENES	MLPH	FOXA1	SIDT1									
BLOCK 2 GENES	CEP55	ASPM	CENPL	AURKA	ESPL1	TTK	UBE2T	NCAPG	GMPS	NDC80	MYBL2	KIF18B	EXO1

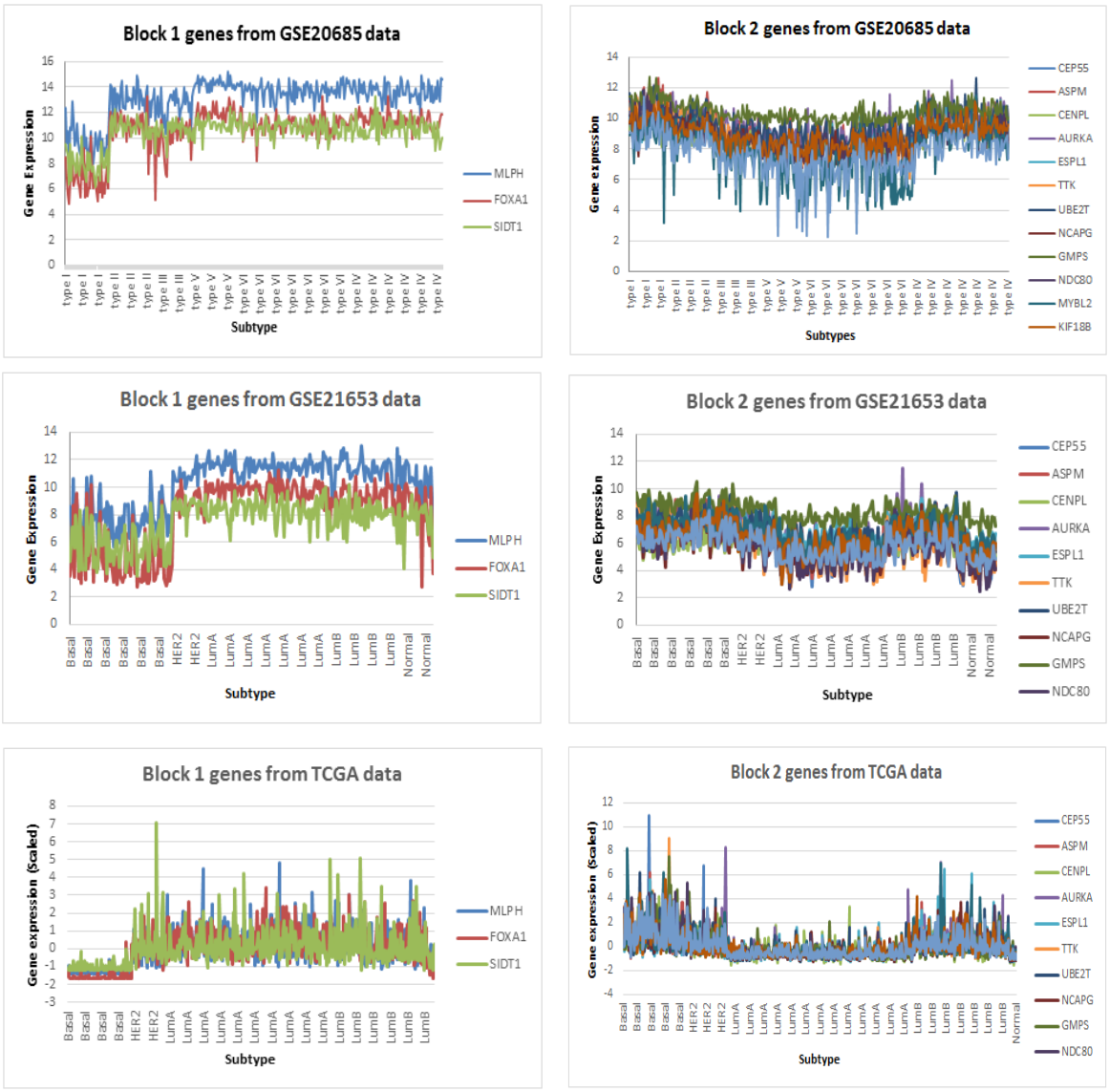


Figure 5-6. Scatter plots for gene blocks: The first block formed by the genes: FOXA1, MLPH and SIDT1. The second block formed by the genes: CEP55, ASPM, CENPL, AURKA, ESPL1, TTK, UBE2T, NCAPG, GMPS, NDC80, MYBL2, KIF18B, and EXO1. For each of the three databases under study: TCGA, GSE21653 and GSE20685 plots.

Furthermore, to estimate the high expression correlation perceived visually through the scatter plots on Figure 5-6 we calculated Pearson and Spearman correlation metrics for genes within the same blocks (see Appendix 6). For genes in the block1, the pair of genes with the highest Pearson Correlation are MLPH and FOXA1 with 0.860, 0.892 and 0.6715 for GSE20685, GSE21653 and TCGA respectively. The SIDT1 gene with MLPH or FOXA1 shows a good correlation for the GSE datasets (~ 0.74), nonetheless lower values for the TCGA (average around 0.35). Similarly for block 2, the results of the correlation index between the genes were highest for the GSE data, obtaining on average ~ 0.80 while TCGA showed values around 0.65. These results corroborate the high correlation between some genes (i.e. genes blocks) and allow us to suggest that the number of important variables found in the phase 2 can be reduced considerably, since between blocks, the genes exhibit similar behavior and maybe provide the same degree of information regarding the response variable.

5.5 Further Inferences and Visualization

For validation purposes, we used two different tools, one offered by Leo Breiman and Adele Cluter [96], and the other by Jones et. al. [94]. First, a method of Leo Breiman and Adele Cutler through Random Forest where offered the possibility of detecting variables interactions. They defined interaction between two variables as the correlation between them, in the sense, that highly correlated variables will have interacting scores [96]. This concept differs from our definition of interaction of variables, which is the ability of a set

of variables to describe a class in a joined manner where we cannot describe a class without one or the other. However, we observed genes that presented similar expression patterns (high correlation) in our analysis of rules extracted by IBF-RF metric.

We applied the code available in Random Forest web page [96], for all three database in study (TCGA, GSE06 and GSE), then we looked for the highest interaction scores to find highly correlated variables according to Breiman (see results in Table 5-6).

Table 5-6. Results of variables interaction according Random Forest code

TCGA			
#	Interaction Ranking	Var 1	Var 2
1	107	MLPH	FOXA1
2	99	FOXC1	FOXA1
3	94	ER.alpha	ESR1
4	90	KRT5	KRT14
5	84	CEP55	UBE2T
6	81	GPR77	ESR1
7	79	CDKN3	NDC80
8	79	KRT17	KRT14
9	78	C6orf97	ESR1
10	74	DEPDC1B	CEP55
11	72	EXO1	UBE2T
12	69	MIA	KRT14
13	68	EXO1	CEP55
14	68	AURKA	CEP55
15	67	AGR3	ESR1
16	66	ASPM	CEP55
17	64	AURKA	UBE2T
18	63	KRT5	MIA
19	63	MLPH	FOXC1
20	63	XBP1	FOXA1

GSE20685			
#	Interaction Ranking	Var 1	Var 2
1	112	GATA3	ESR1
2	105	THSD4	GATA3
3	104	CA12	GATA3
4	103	GREB1	ESR1
5	97	ASPM	KIF18B
6	94	ASPM	AURKA
7	91	THSD4	ESR1
8	89	FOXA1	CAV2
9	82	ASPM	CEP55
10	78	CA12	ESR1
11	78	GREB1	GATA3
12	77	ASPM	ESPL1
13	77	ASPM	MYBL2
14	74	IGF1R	ESR1
15	71	KIF18B	AURKA
16	70	IGF1R	THSD4
17	69	CA12	THSD4
18	67	MLPH	FOXA1
19	63	NCAPG	ASPM
20	63	PTX3	CAV2

GSE21653			
#	Interaction Ranking	Var 1	Var 2
1	83	GATA3	ESR1
2	79	GATA3	CA12
3	76	MYBL2	AURKA
4	67	MLPH	FOXA1
5	60	MYBL2	ESPL1
6	55	KIF18B	AURKA
7	51	NAT1	GATA3
8	50	MYBL2	CENPN
9	48	AGR3	GATA3
10	47	ESPL1	AURKA
11	46	MYBL2	DSCC1
12	45	AGR3	ESR1
13	45	TBC1D9	GATA3
14	45	NCAPG	MYBL2
15	44	AR	MLPH
16	43	NCAPG	AURKA
17	43	NAT1	ESR1
18	41	CA12	ESR1
19	41	TBC1D9	ESR1
20	40	TIMELESS	MYBL2

Once we had the results of important variables according to Breiman, we compared those highly correlated genes with genes from the two blocks from Table 5-5 extracted as important from our integrative model. We found that the interactions resulting from the Breiman code were indeed genes also found within our defined genes blocks, corroborating the strong correlation between them. This is an interesting finding of this work, where highly correlated variables (i.e. genes) can be extracted as important and then the random forest ensemble model can randomly select any of them and generate rules with different genes but same patterns. For example, FOXA1-CEP55 and FOXA1-TTK interactions shows similar expression behavior across all subtypes as seen in Table 5-4 revealing CEP55 and TTK to be highly correlated variables based on their expression but the random forest yields them as two different interactions when they could be counted as one representative rule. The rules extracted through IBF-RF metric are a result of a combinatorial process perform by random forest to generate the best splits at every node. Due to the highly correlated nature of some genes (grouped by blocks) and random sampling process of selecting and evaluating variables (i.e. genes) at each split we have observed that any gene within the same block (i.e. CEP55 or TTK) can be selected as important with respect to other specific gene (i.e. FOXA1) and still be considered as two different rules.

The second tool used to validate our results is proposed by Jones et. Al. in [94] to generate a modified partial dependence plots from Random Forest to visualize interactions between pairs of variables. This implementation was developed using Edarf R package for exploratory data analysis using Random Forests. To extract the marginal

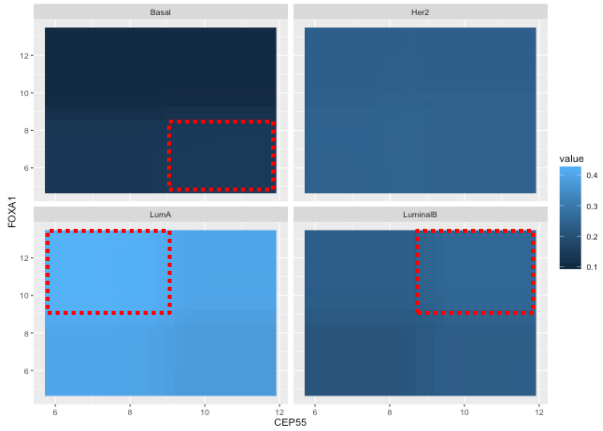
effect of specific rules we used partial dependency plots and evaluated the behavior of three rules extracted by our metrics: FOXA1-CEP55, FOXC1-THSD4 and MLPH-NOSTRIN. We wanted to validate whether the interaction results of these rules for each subtype were similar to those shown in Table 5-4. In Figure 5-7 we can visualize the interaction for the rule FOXA1-CEP55 calculated through plot_pd functions of Edarf R package, for each datasets in study: (a) GSE20685, (b) GSE21653 and (c) TCGA. For instance, in Figure 5-7a, y axis is FOXA1 gene and x axis is CEP55 gene, indicating that: 1) basal occurs when FOXA1 is lowly expressed and CEP55 is highly expressed, 2) luminal A occurs when FOXA1 is highly expressed and CEP55 is lowly expressed and 3) luminal B occurs when both FOXA1 and CEP55 are highly expressed. Similarly, the results for FOXC1-THSD4 and MLPH-NOSTRIN are showed in Figure 5-8 and Figure 5-9, respectively.

These results corroborated the interaction conclusions shown in Table 5-4 and lead us to validate that the IBF-RF metric can rank important interactive patterns considering all possible rules granting the opportunity to further explore the biological mechanism of these interactions at the experimental level.

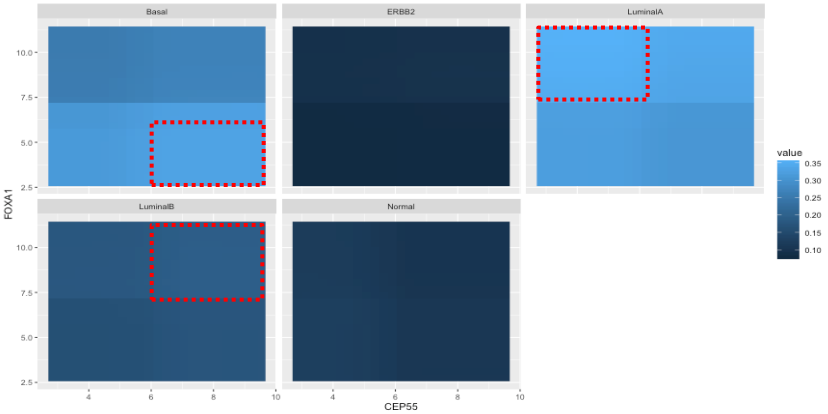
Rule: FOXA1- CEP55

# Rule	Gene	Basal	Her2	LumA	LumB	Normal
3	FOXA1 - CEP55					

a) GSE20685



b) GSE21653



c) TCGA

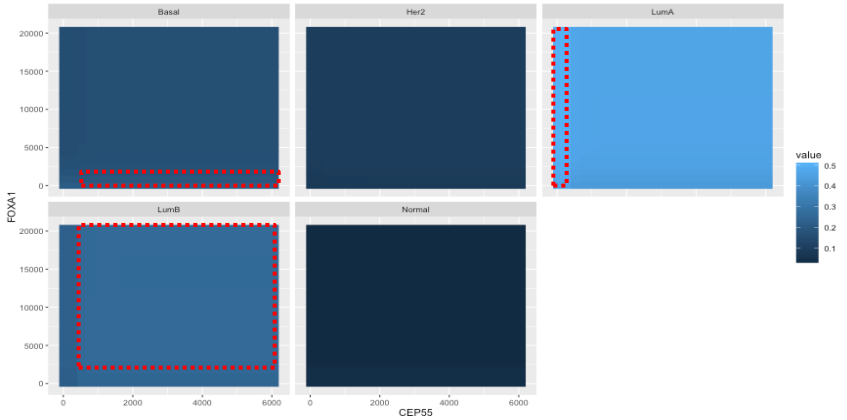
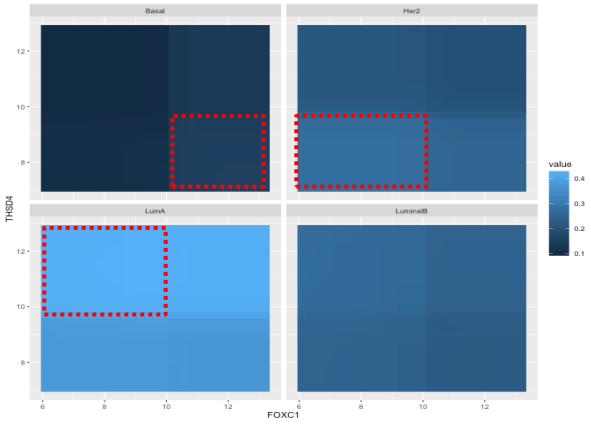


Figure 5-7. Plot partial dependence from Random Forests: These plots visualize interactions for the rule FOXA1-CEP55 calculated through partial dependence, for each one datasets in study: (a) GSE20685, (b) GSE21653 and (c) TCGA. The red squares highlight the area with the necessary expression of genes in interaction for the prediction of a specific subtype.

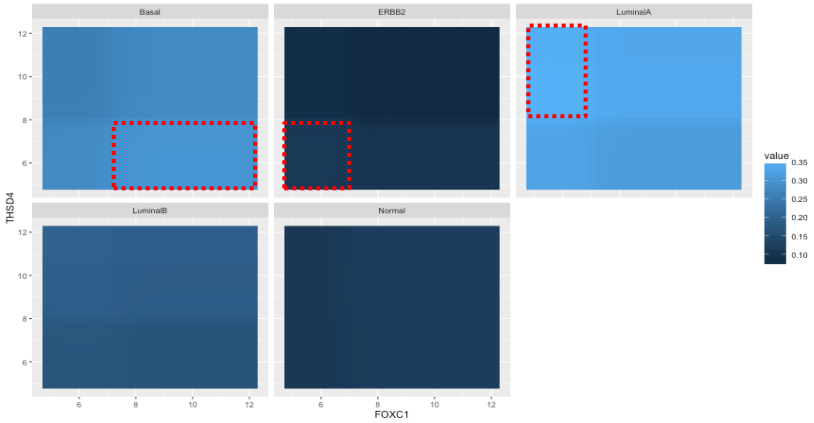
Rule: FOXC1- THSD4

# Rule	Gene	Basal	Her2	LumA	LumB	Normal
4	FOXC1 - THSD4	-	-	-	-	

a) GSE20685



b) GSE21653



c) TCGA

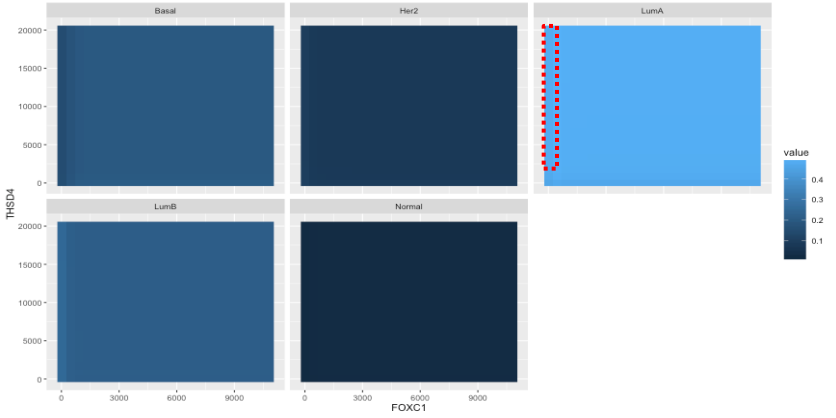
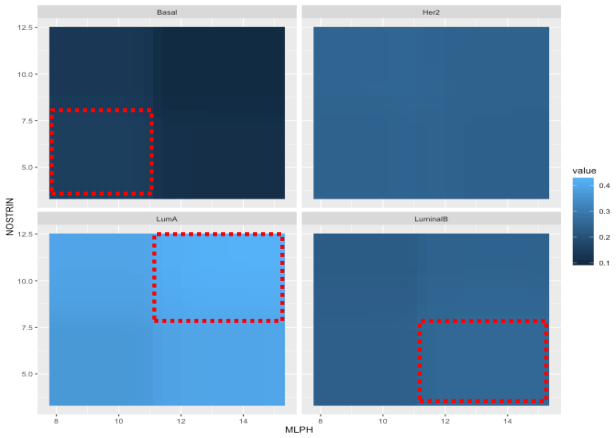


Figure 5-8. Plot partial dependence from Random Forests: These plots visualize interactions for the rule FOXC1- THSD4 calculated through partial dependence, for each one datasets in study: (a) GSE20685, (b) GSE21653 and (c) TCGA. The red squares highlight the area with the necessary expression of genes in interaction for the prediction of a specific subtype.

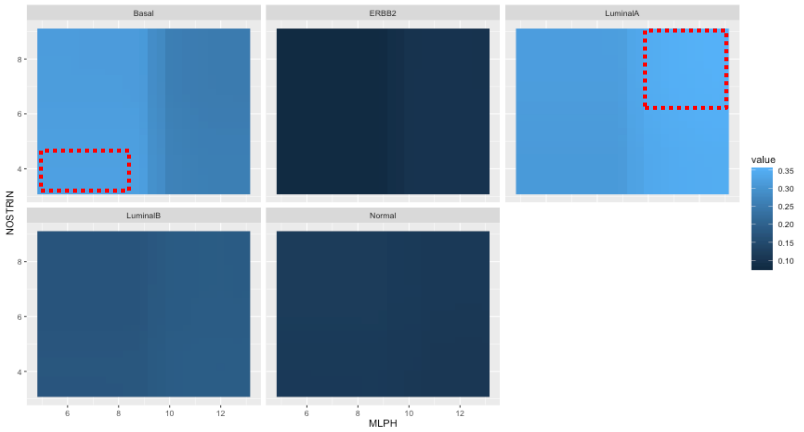
Rule: MLPH- NOSTRIN

# Rule	Gene	Basal	Her2	LumA	LumB	Normal
6	MLPH - NOSTRIN					

a) GSE20685



b) GSE21653



c) TCGA

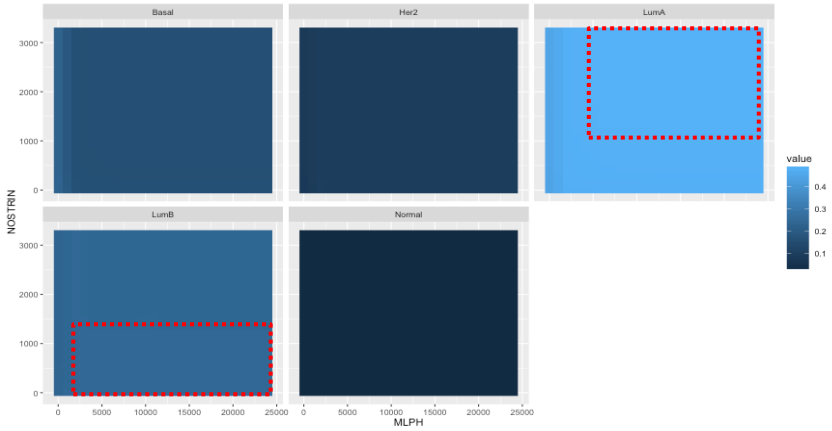


Figure 5-9. Plot partial dependence from Random Forests: These plots visualize interactions for the rule MLPH- NOSTRIN calculated through partial dependence, for each one datasets in study: (a) GSE20685, (b) GSE21653 and (c) TCGA. The red squares highlight the area with the necessary expression of genes in interaction for the prediction of a specific subtype.

5.6 Conclusions

The results strongly support the importance of the rules extracted by the IBF-RF metric and we can conclude that ranked top rules demonstrate significant interactions that can discriminate breast cancer subtypes. IBF-RF metric is a great contribution because it provides a tool capable to assess interaction importance in a holistic manner without any prior knowledge onto which features combinations are most important. Also, our metric can rank rules considering all possible ones. To the best of our knowledge, there is not a metric that can evaluate higher degree interactions for random forest models and that can search across all possibilities. Therefore, this grants the opportunity to pinpoint the most important interactions and explore the biological meaning of these interactions at the experimental level. However, the computational complexity and run time of IBF-RF metric increases as the number of trees used to build the forest (ntree) increases, hence, for high number of ensemble trees (>10000) is necessary to run the metric algorithm in a parallel scheme. We suggest to parallelize the code to make it computationally efficient.

The reason why we did not obtain a very high frequency of rules compared to the number of trees (ntree) used in each dataset can be explained by the random nature of the Random Forest classifier. Therefore, some source of normalization scheme should be incorporated into the metric.

Also, not many common rules were found since many rules were composed on highly correlated genes and our current metric is considering them as unique rules when perhaps all these could be combined as they describe the same subtype process. As observed in the dispersion graphs and in the Pearson correlation index (Figure 5-6), there are many

genes with high correlation therefore when selecting a variable to generate a rule; this variable can be easily replaced by another that has high correlation. For our analysis, we created two blocks of genes with high correlation (See **Table 5-5**) and we saw that many of the rules were formed from a combinatorial process between these two blocks and not within the same block. Then, future studies can focus at the feature selection stage to extract these correlated blocks of features.

On the other hand, IBF-RF metric extracts only the genes from rules formed in the construction process of Random Forest algorithm. The metric no take into account the values used on each split of the branch, let it than an extracted rule classify more than one class at the same time. Therefore, we suggest that future studies considers different values regions defined by Random Forest in branches construction. .

6 CONCLUSIONS AND FUTURE WORK

This thesis enabled a methodology to integrate large and heterogeneous data types to extract interactions that can deepen the current understanding of how breast cancer subtypes are characterized through the implementation of a multi-stage data mining approach. It also includes a new metric that can capture in a holistic manner which interactive patterns are most significant. Consequently, to achieve these objectives we defined three principal phases: 1) integration of transcriptomic and clinical data, 2) integration of transcriptomic, proteomic and methylation data, and 3) ranking of interaction patterns to discriminate among subtypes.

In the first phase, we evaluated the performance of two transcriptomic platforms (i.e. microarray and RNA-sequence) using three feature selection methods (CFS, IG and ReliefF) and three classification methods (KNN, SVM and RFs). Although, we hypothesized that integrating transcriptomic and clinical data would improve prediction of breast cancer subtypes the results of this thesis does not supports a significant improvement in the metrics when compared with the models using transcriptomic and clinical data separately. However, the integrative model achieved high accuracy and AUC values mainly described by transcriptomic predictors (AUC: 90.84%, Accuracy: 90.49%). Another by product of the first phase was that we can concluded that the information from two microarray and RNA-seq platforms are highly correlated and provides the same degree of information regarding the response variable which has been

commonly discussed in the literature. Hence, we decided to use RNA-seq platform for further genomic integration since it had no missing values and it is becoming a new standard for gene expression.

The second phase of this thesis gathered large-scale datasets (gene expression, protein and methylation) to build an integrative model to predict breast cancer subtypes. We applied five different feature selection methods (CFS, Information Gain, ReliefF, SVM-RFE and FAST Clustering based) to reduce the feature space by choosing feature groups with the best classification performance. The reduction of features permitted the development and computational implementation of an integrative model with heterogeneous genomics variables. The integrative approach revealed ~250 relevant features to discriminate breast cancer subtypes yielding good performance metrics (AUC: 85.9%, Accuracy: 90.6%). Although most extracted top variables were transcriptomic, one methylation (cg02643667) was found among the ten most significant which is linked to TFF1, an estrogen-regulated protein, strongly associated to breast cancer. Also, this thesis revealed nine genes (CENPL, RERGL, TBX19, KCMF1, ADCY4, NOSTRIN, CMTM7, SCCPDH and DSCC1) as strong candidates for future experiments since there is no literature support on breast cancer per our PubMed database search on May 16, 2017.

In the third phase, we develop IBF-FR, a new metric capable of assessing in a holistic manner the interaction between relevant features (rules) without any prior knowledge of important features combinations. Also, this metric can rank the rules according to their frequencies. Finally, thanks to the evaluation of the IBF-FR results, we defined two sets

of genes that have a similar behavior and we were able to infer that several distinct rules are formed by the combination between these set of genes. The first set formed by: FOXA1, MLPH and SIDT1 genes, and the second by the genes: CEP55, ASPM, CENPL, AURKA, ESPL1, TTK, UBE2T, NCAPG, GMPS, NDC80, MYBL2, KIF18B and EXO1. These results are encouraging since we extracted important patterns and addressed the computational complexity existing when high dimensional data is studied. Furthermore, we developed IBF-RF, a metric capable of assessing and rank the interactions of higher order. This metric can help in the exploration of the effects of significant patterns and expand current knowledge of the behavior of the each breast cancer subtype.

6.1 Future Work

The milestones reached in this thesis were significantly important to characterize the behavior of each breast cancer subtypes. This work can be further improved by addressing the following limitations. Further experimental studies of the important variables extracted from the integrative models are recommended to better understand their causality effects to breast cancer mechanisms. Especially, those correlated blocks of genes found in Phase 3 can reveal new knowledge on how this disease manifests at the subtype level.

Additionally, the inclusion of more complete and balanced clinical data could expose important results to understand the mechanisms of each breast cancer subtype especially those of aggressive behavior that tend to have poorer survival rates. The clinical

information available in this work was, perhaps, not sufficient to extract links with breast cancer subtype and survival.

Finally, the computational complexity of IBF-RF metric can be improved by parallelizing great parts of the programming algorithm, since the run can take many days, depending of the number of trees used in the RFs construction, for instance, for ntree = 15000, the time run can take up to 5 days. Additionally, the metric value for a given rule increased as the number of trees in the random forest classifier increase. Some source of normalization scheme should be incorporated into the metric to reduce the impact of this parameter (number of trees).

7 REFERENCES

[1] "Registro central de cancer de Puerto Rico," 2017. [Online]. Available: http://www.rcpr.org/Datos-de-Cancer/Estadisticas.

[2] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours.," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.

[3] J. Makki, "Diversity of breast carcinoma: Histological subtypes and clinical relevance," *Clin. Med. Insights Pathol.*, vol. 8, no. 1, pp. 23–31, 2015.

[4] A. Prat, C. Cruz, K. A. Hoadley, O. Díez, C. M. Perou, and J. Balmaña, "Molecular features of the basal-like breast cancer subtype based on BRCA1 mutation status," *Breast Cancer Res. Treat.*, vol. 147, no. 1, pp. 185–191, 2014.

[5] O. Yersal and S. Barutca, "Biological subtypes of breast cancer: Prognostic and therapeutic implications.," *World J. Clin. Oncol.*, vol. 5, no. 3, pp. 412–24, 2014.

[6] F. Ades, D. Zardavas, I. Bozovic-Spasojevic, L. Pugliano, D. Fumagalli, E. De Azambuja, G. Viale, C. Sotiriou, and M. Piccart, "Luminal B breast cancer: Molecular characterization, clinical management, and future perspectives," *J. Clin. Oncol.*, vol. 32, no. 25, pp. 2794–2803, 2014.

[7] S. Liu, H. Wang, L. Zhang, C. Tang, L. Jones, H. Ye, L. Ban, A. Wang, Z. Liu, F. Lou, D. Zhang, H. Sun, H. Dong, G. Zhang, Z. Dong, B. Guo, H. Yan, C. Yan, L. Wang, Z. Su, Y. Li, X. F. Huang, S.-Y. Chen, and T. Zhou, "Rapid detection of genetic mutations in individual breast cancer patients by next-generation DNA sequencing.," *Hum. Genomics*, vol. 9, p. 2, 2015.

[8] B. A. McKinney, D. M. Reif, M. D. Ritchie, and J. H. Moore, "Machine learning for detecting gene-gene interactions: A review," *Appl. Bioinformatics*, vol. 5, no. 2, pp. 77–88, 2006.

[9] C. L. Koo, M. J. Liew, M. S. Mohamad, and A. Mohamed Salleh, "A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology," *Biomed Res. Int.*, 2013.

[10] E. R. Dougherty, "Small sample issues for microarray-based classification," *Comp. Funct. Genomics*, vol. 2, pp. 28–34, 2001.

[11] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci. (Ny).*, vol. 250, pp. 113–141, 2013.

[12] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 4, pp. 463–484, 2012.

[13] J. a. Sáez, J. Luengo, and F. Herrera, “Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification,” *Pattern Recognit.*, vol. 46, pp. 355–364, 2013.

[14] “Gene Set Enrichment Analysis (GSEA).” [Online]. Available: http://software.broadinstitute.org/gsea/index.jsp. [Accessed: 18-Dec-2015].

[15] “cBioPortal for Cancer Genomics.” [Online]. Available: http://www.cbioportal.org/. [Accessed: 18-Dec-2015].

[16] et al Guyon, I., “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[17] et al Fan, R., “Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases.,” *Genet. Epidemiol.*, vol. 35, no. 7, pp. 706–21, Nov. 2011.

[18] L. Yu and H. Liu, “Redundancy based feature selection for microarray data,” *Proc. tenth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 737–742, 2004.

[19] et al Liu, H., “A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns,” *Genome informatics*, 2002.

[20] J. Moore and B. White, “Tuning ReliefF for genome-wide genetic analysis,” *Evol. Comput. Mach. Learn. data Min. Bioinforma.*, pp. 166–175, 2007.

[21] et al Zhang, Y., “Gene selection algorithm by combining reliefF and mRMR,” *BMC Genomics*, 2008.

[22] M. Hall, “Correlation-based feature selection for machine learning,” *Diss. Univ. Waikato*, no. April, 1999.

[23] M. a. Hall and L. a. Smith, “Practical feature subset selection for machine learning,” *Comput. Sci.*, vol. 98, pp. 181–191, 1998.

[24] I. Kononenko, “Estimating attributes: analysis and extensions of RELIEF,” *Mach. Learn. ECML-94*, 1994.

[25] C.-S. Y. C.-S. Yang, L.-Y. C. L.-Y. Chuang, Y.-J. C. Y.-J. Chen, and C.-H. Y. C.-H. Yang, “Feature Selection Using Memetic Algorithms,” *2008 Third Int. Conf. Converg. Hybrid Inf. Technol.*, vol. 1, pp. 416–423, 2008.

[26] R. Kohavi and R. Kohavi, “Wrappers for feature subset selection,” *Artif. Intell.*, vol. 97, no. 97, pp. 273–324, 1997.

[27] et al Bolón-Canedo, V., “A review of microarray datasets and applied feature selection methods,” *Inf. Sci. (Ny).*, vol. 282, pp. 111–135, 2014.

[28] I. GUYON, S. WESTON, and J. BARNHILL, “Gene Selection for Cancer Classification using Support Vector Machines,” *Barnhill Bioinformatics, Savannah, Georg. USA*, pp. 389–422, 2002.

[29] at al Duan, K., “Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data,” *IEEE Trans. Nanobioscience*, vol. 4, no. 3, pp. 228–234,

Sep. 2005.

[30] et al Tang, Y., "Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis.," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 4, no. 3, pp. 365–81, Jan. 2007.

[31] P. A. Mundra and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection.," *IEEE Trans. Nanobioscience*, vol. 9, no. 1, pp. 31–7, Mar. 2010.

[32] Y. Ding and D. Wilkins, "Improving the performance of SVM-RFE to select genes in microarray data.," *BMC Bioinformatics*, vol. 7 Suppl 2, no. Suppl 2, p. S12, Jan. 2006.

[33] S. Qinbao, N. Jingjie, and W. Guangtao, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 1–14, Jan. 2013.

[34] K. N. Stevens, T. M. Cover, and P. E. Hart, "Nearest neighbor pattern classification," vol. I, 1967.

[35] C. Cortes and V. Vapnik, "Support-Vector Networks," vol. 297, pp. 273–297, 1995.

[36] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[37] S. Hanash, "Integrated global profiling of cancer.," *Nat. Rev. Cancer*, vol. 4, no. 8, pp. 638–44, Aug. 2004.

[38] D. Kim, J.-G. Joung, K.-A. Sohn, H. Shin, Y. R. Park, M. D. Ritchie, and J. H. Kim, "Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction.," *J. Am. Med. Inform. Assoc.*, vol. 22, no. 1, pp. 109–20, Jan. 2015.

[39] and J. C. Luis Martín, Alberto Anguita, Víctor Maojo, *Integration of Omics Data for Cancer Research*. Springer, 2010.

[40] D. Kim, R. Li, S. M. Dudek, and M. D. Ritchie, "Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer.," *J. Biomed. Inform.*, vol. 56, pp. 220–8, Aug. 2015.

[41] W. Zhou, "Machine learning methods for omics data integration," Iowa State University, 2011.

[42] D. M. Reif, A. A. Motsinger, B. A. McKinney, J. E. Crowe, and J. H. Moore, "Feature Selection using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types," *2006 IEEE Symp. Comput. Intell. Bioinforma. Comput. Biol.*, pp. 1–8, 2006.

[43] M. List, A.-C. Hauschild4, Q. Tan, T. A. Kruse, J. Mollenhauer, J. Baumbach, and R. Batra, "Classification of Breast Cancer Subtypes by combining Gene Expression and DNA Methylation Data," *J. Integr. Bioinform.*, vol. 11(2), p. 236, 2014.

[44] A. G. Heidema, J. M. A. Boer, N. Nagelkerke, E. C. M. Mariman, D. L. van der A, and E. J. M. Feskens, "The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases.," *BMC Genet.*, vol. 7, no. 1, p. 23, Jan. 2006.

[45] J. H. Moore, "The ubiquitous nature of epistasis in determining susceptibility to common human diseases," *Hum. Hered.*, vol. 56, no. 1–3, pp. 73–82, 2003.

[46] B. A. Goldstein, A. E. Hubbard, A. Cutler, and L. F. Barcellos, "An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings.," *BMC Genet.*, vol. 11, no. 1, p. 49, Jan. 2010.

[47] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh, "Identifying SNPs predictive of phenotype using random forests.," *Genet. Epidemiol.*, vol. 28, no. 2, pp. 171–82, Feb. 2005.

[48] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B (...*, 1996.

[49] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.," *Am. J. Hum. Genet.*, vol. 69, no. 1, pp. 138–47, Jul. 2001.

[50] P. R. Lucek and J. Ott, "Neural network analysis of complex traits.," *Genet. Epidemiol.*, vol. 14, no. 6, pp. 1101–6, Jan. 1997.

[51] D. R. Velez, B. C. White, A. A. Motsinger, W. S. Bush, M. D. Ritchie, S. M. Williams, and J. H. Moore, "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction.," *Genet. Epidemiol.*, vol. 31, no. 4, pp. 306–15, May 2007.

[52] J. Namkung, R. C. Elston, J.-M. Yang, and T. Park, "Identification of gene-gene interactions in the presence of missing data using the multifactor dimensionality reduction method.," *Genet. Epidemiol.*, vol. 33, no. 7, pp. 646–56, Nov. 2009.

[53] S. Y. Lee, Y. Chung, R. C. Elston, Y. Kim, and T. Park, "Log-linear model-based multifactor dimensionality reduction method to detect gene gene interactions.," *Bioinformatics*, vol. 23, no. 19, pp. 2589–95, Oct. 2007.

[54] S. Bicciato, M. Pandin, G. Didonè, and C. Di Bello, "Pattern identification and classification in gene expression data using an autoassociative neural network model.," *Biotechnol. Bioeng.*, vol. 81, no. 5, pp. 594–606, Mar. 2003.

[55] T.-C. Hsia, H.-C. Chiang, D. Chiang, L.-W. Hang, F.-J. Tsai, and W.-C. Chen, "Prediction of survival in surgical unresectable lung cancer by artificial neural networks including genetic polymorphisms and clinical parameters.," *J. Clin. Lab. Anal.*, vol. 17, no. 6, pp. 229–34, Jan. 2003.

[56] A. Sherriff and J. Ott, "20 Applications of neural networks for gene finding," *Adv. Genet.*, vol. 42, pp. 287–297, 2001.

[57] N. R. Cook, R. Y. L. Zee, and P. M. Ridker, “Tree and spline based association analysis of gene-gene interaction models for ischemic stroke.,” *Stat. Med.*, vol. 23, no. 9, pp. 1439–53, May 2004.

[58] K. L. Lunetta, L. B. Hayward, J. Segal, and P. Van Eerdewegh, “Screening large-scale association study data: exploiting interactions using random forests.,” *BMC Genet.*, vol. 5, no. 1, p. 32, Jan. 2004.

[59] S. J. Winham, C. L. Colby, R. R. Freimuth, X. Wang, M. de Andrade, M. Huebner, and J. M. Biernacka, “SNP interaction detection with Random Forests in high-dimensional genetic data.,” *BMC Bioinformatics*, vol. 13, no. 1, p. 164, Jan. 2012.

[60] L. Breiman, “Manual on setting up, using, and understanding random forests v3. 1,” *Tech. Report, http//oz.berkeley.edu/users/breiman, Stat. Dep. Univ. Calif. Berkeley, …*, p. 29, 2002.

[61] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, “Understanding variable importances in forests of randomized trees,” *Neural Inf. Process. Syst.*, pp. 1–9, 2013.

[62] C. Niel, C. Sinoquet, C. Dina, and G. Rocheleau, “A survey about methods dedicated to epistasis detection.,” *Front. Genet.*, vol. 6, p. 285, Jan. 2015.

[63] A. Bureau, J. Dupuis, B. Hayward, K. Falls, and P. Van Eerdewegh, “Mapping complex traits using Random Forests.,” *BMC Genet.*, vol. 4 Suppl 1, no. Suppl 1, p. S64, Jan. 2003.

[64] S. Forbes and D. Beare, “COSMIC: exploring the world’s knowledge of somatic mutations in human cancer,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D805–D811, 2015.

[65] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. J. Diaz, and M. W. K. and Kenneth, “Cancer genome landscapes.,” *Science*, vol. 339, no. 6127, pp. 1546–58, Mar. 2013.

[66] K. L. Abbott, E. T. Nyre, J. Abrahante, Y.-Y. Ho, R. Isaksson Vogel, and T. K. Starr, “The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice.,” *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D844-8, Jan. 2015.

[67] M. Templ, A. Kowarik, and P. Filzmoser, “Iterative stepwise regression imputation using standard and robust methods,” *Comput. Stat. Data Anal.*, vol. 55, no. 10, pp. 2793–2806, 2011.

[68] S. Verboven, K. Vanden Branden, and P. Goos, “Sequential imputation for missing values,” *Comput. Biol. Chem.*, vol. 31, no. 5, pp. 320–327, 2007.

[69] V. Todorov, “Scalable Robust Estimators with High Breakdown Point for Incomplete Data.” R package version 0.4-9, 2016.

[70] S. Singhal and M. Jena, “A Study on WEKA Tool for Data Preprocessing , Classification and Clustering,” no. 6, pp. 250–253, 2013.

[71] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, Fourth Edi. New York, 2002.

[72] D. Meyer, E. Dimitriadou, K. Hornik, W. Andreas, and F. Leisch, "e1071: Misc Functions of the Department of Statistics, Probability Theory Group." R package version 1.6-7, 2015.

[73] A. Liaw and M. Wiener, "Classification and Regression by randomForest." R News 2(3), 18--22, 2002.

[74] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.," *Sci. Signal.*, vol. 6, no. 269, p. pl1, Apr. 2013.

[75] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz, "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.," *Cancer Discov.*, vol. 2, no. 5, pp. 401–4, May 2012.

[76] X. Wang, J., Ray, P. S., Sim, M. S., Zhou, X. Z., Lu, K. P., Lee, A. V., ... & Cui, "FOXC1 regulates the functions of human basal-like breast cancer cells by activating NF-κB signaling," *Oncogene*, vol. 31, no. 45, pp. 4798–4802, 2012.

[77] R. A. Bernardo, G. M., Bebek, G., Ginther, C. L., Sizemore, S. T., Lozada, K. L., Miedler, J. D., ... & Keri, "FOXA1 represses the molecular phenotype of basal breast cancer cells," *Oncogene*, vol. 32, no. 5, pp. 554–563, 2013.

[78] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters," *Omi. a J. Integr. Biol.*, vol. 16(5), pp. 284–287, 2012.

[79] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[80] M. Ashburner, C. A. Ball, J. A. Blake, and D. Botstein, "Gene ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25–9, 2000.

[81] I. Narvaez-bandera and W. Torres-garcia, "Data-driven Approach to Extract Molecular Patterns in Breast Cancer using Transcriptomic and Clinical Data," *Proc. 2016 Ind. Syst. Eng. Res. Conf. - Unpubl.*, p. 6, 2016.

[82] J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, T. David Voduc, A. Vickery, S. Davies, C. Fauron, Z. H. Xiaping He, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *J. Clin. Oncol.*, vol. 27, no. 8, pp. 1160–1167, 2009.

[83] P. Romanski and L. Kotthoff, "Package FSelector: Selecting Attributes." R package version 0.21., p. 18, 2016.

[84] Q. Song, “XI’AN JIAOTONG UNIVERSITY,” 2013. [Online]. Available: http://gr.xjtu.edu.cn/c/document_library/get_file?folderId=1851177&name=DLFE-36278.pdf.

[85] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, “GeneCards: A novel functional genomics compendium with automated data mining and query reformulation support,” *Bioinformatics*, vol. 14, no. 8, pp. 656–664, 1998.

[86] (US). Information Bethesda (MD) National Center for Biotechnology, “PubMed [Internet].” [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK3827/. [Accessed: 09-May-2017].

[87] S. Dedeurwaerder, C. Desmedt, E. Calonne, S. K. Singhal, B. Haibe-Kains, M. Defrance, S. Michiels, M. Volkmar, R. Deplus, J. Luciani, F. Lallemand, D. Larsimont, J. Toussaint, S. Haussy, F. Rothé, G. Rouas, O. Metzger, S. Majjaj, K. Saini, P. Putmans, G. Hames, N. van Baren, P. G. Coulie, M. Piccart, C. Sotiriou, and F. Fuks, “DNA methylation profiling reveals a predominant immune component in breast cancers,” *EMBO Mol. Med.*, vol. 3, no. 12, pp. 726–741, 2011.

[88] F. Fuks, S. Dedeurwaerder, C. Sotiriou, and C. Desmedt, “Epigenetic portraits of human breast cancers,” 2012.

[89] J. Listgarten, S. Damaraju, B. Poulin, L. Cook, J. Dufour, A. Driga, J. Mackey, D. Wishart, R. Greiner, and B. Zanke, “Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms.,” *Clin. Cancer Res.*, vol. 10, no. 8, pp. 2725–37, 2004.

[90] H. Schwender, M. Zucknick, K. Ickstadt, H. M. Bolt, C. Justenhoven, H. Brauch, B. Pesch, V. Harth, U. Hamann, T. Brüning, and Y. Ko, “A pilot study on the application of statistical classification procedures to molecular epidemiological data,” *Toxicol. Lett.*, vol. 151, no. 1, pp. 291–299, 2004.

[91] S. Yuanyuan, L. Zhe, and J. Ott, “Detecting gene-gene interactions using support vector machines with L<inf>1</inf> penalty,” *Bioinforma. Biomed. Work. (BIBMW), 2010 IEEE Int. Conf.*, pp. 309–311, 2010.

[92] D. M. Skapura, *Building Neural Networks*. Addison-Wesley Professional, 1996.

[93] X. Chen and H. Ishwaran, “Random Forests for Genomic Data Analysis,” *Genomics*, vol. 99, no. 6, pp. 323–329, 2013.

[94] Z. Jones and F. Linder, “Exploratory Data Analysis using Random Forests,” *73rd Annu. MPSA Conf. - Unpubl.*, 2015.

[95] “Random Forest Codes.” [Online]. Available: https://stats.stackexchange.com/questions/41443/how-to-actually-plot-a-sample-tree-from-randomforestgettree. [Accessed: 17-May-2017].

[96] L. Breiman and A. Cutler, “Random Forests.” [Online]. Available: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

8 APPENDICES

Appendix 1

Table 8-1. Enrichment analysis using enrichKEGG: set of genes within known gene-groups by their functional or pathway categories

ID	Description	Gene Ratio	p.adjust	qvalue	*geneID*	Count
hsa05215	Prostate cancer	9/97	0.0005014	0.0004267	IGF1R/CCNE2/GSTP1/AR/ERBB2/BCL2/CCNE1/EGFR/CCND1	9
hsa04114	Oocyte meiosis	10/97	0.0005076	0.000432	AURKA/ESPL1/IGF1R/CCNE2/ADCY4/AR/ADCY9/CCNE1/PGR/CCNB1	10
hsa05200	Pathways in cáncer	15/97	0.0141064	0.0120036	FGF2/IGF1R/CCNE2/ADCY4/GSTP1/AR/ERBB2/ADCY9/BCL2/CCNE1/EGFR/CCND1/FZD10/STAT5A/ZBTB16	15
hsa04115	p53 signaling pathway	6/97	0.0141064	0.0120036	CCNE2/RRM2/PMAIP1/CCNE1/CCNB1/CCND1	6
hsa05205	Proteoglycans in cáncer	9/97	0.0630812	0.0536777	ESR1/CAV2/FGF2/IGF1R/ERBB2/ERBB4/EGFR/CCND1/FZD10	9
hsa04976	Bile secretion	5/97	0.0734797	0.062526	ADCY4/ADCY9/EPHX1/ABCC2/ABCG2	5
hsa02010	ABC transporters	4/97	0.0734797	0.062526	ABCC11/ABCA12/ABCC2/ABCG2	4
hsa01524	Platinum drug resistance	5/97	0.0734797	0.062526	GSTP1/ERBB2/PMAIP1/BCL2/ABCC2	5
hsa01521	EGFR tyrosine kinase inhibitor resistance	5/97	0.0904471	0.0769641	FGF2/IGF1R/ERBB2/BCL2/EGFR	5
hsa00480	Glutathione metabolism	4/97	0.0904471	0.0769641	GSTP1/RRM2/G6PD/GGCT	4
hsa04510	Focal adhesion	8/97	0.0905525	0.0770538	CAV2/IGF1R/ERBB2/BCL2/EGFR/CCND1/ITGB8/MYLK4	8
hsa04110	Cell cycle	6/97	0.0905525	0.0770538	ESPL1/TTK/CCNE2/CCNE1/CCNB1/CCND1	6
hsa04923	Regulation of lipolysis in adipocytes	4/97	0.0905525	0.0770538	ADCY4/ADCY9/ADORA1/NPY1R	4
hsa04914	Progesterone-mediated oocyte maturation	5/97	0.1292018	0.1099416	IGF1R/ADCY4/ADCY9/PGR/CCNB1	5
hsa05218	Melanoma	4/97	0.1777828	0.1512807	FGF2/IGF1R/EGFR/CCND1	4
hsa05219	Bladder cancer	3/97	0.1801119	0.1532626	ERBB2/EGFR/CCND1	3
hsa04971	Gastric acid secretion	4/97	0.1801119	0.1532626	ADCY4/ADCY9/KCNJ16/MYLK4	4
hsa04921	Oxytocin signaling pathway	6/97	0.1943842	0.1654073	ADCY4/ADCY9/EGFR/CCND1/MYLK4/CACNB2	6
hsa05166	HTLV-I infection	8/97	0.2210527	0.1881003	XBP1/ADCY4/ADCY9/MYBL2/CCND1/MYBL1/FZD10/STAT5A	8
hsa05222	Small cell lung cancer	4/97	0.2404299	0.2045889	CCNE2/BCL2/CCNE1/CCND1	4
hsa04913	Ovarian steroidogenesis	3/97	0.2404299	0.2045889	IGF1R/ADCY4/ADCY9	3
hsa04012	ErbB signaling pathway	4/97	0.2404299	0.2045889	ERBB2/ERBB4/EGFR/STAT5A	4

hsa05213	Endometrial cancer	3/97	0.2404299	0.2045889	ERBB2/EGFR/CCND1	3
hsa05414	Dilated cardiomyopathy	4/97	0.2404299	0.2045889	ADCY4/ADCY9/ITGB8/CACNB2	4
hsa04020	Calcium signaling pathway	6/97	0.244221	0.2078149	ADCY4/ERBB2/ADCY9/ERBB4/EGFR/MYLK4	6
hsa03430	Mismatch repair	2/97	0.2443439	0.2079195	EXO1/RFC4	2
hsa04211	Longevity regulating pathway	4/97	0.2443439	0.2079195	IGF1R/ADCY4/ADIPOQ/ADCY9	4
hsa05223	Non-small cell lung cancer	3/97	0.2443439	0.2079195	ERBB2/EGFR/CCND1	3
hsa05221	Acute myeloid leukemia	3/97	0.2466245	0.2098601	CCND1/STAT5A/ZBTB16	3
hsa04915	Estrogen signaling pathway	4/97	0.2584271	0.2199033	ESR1/ADCY4/ADCY9/EGFR	4
hsa04933	AGE-RAGE signaling pathway in diabetic complications	4/97	0.2584271	0.2199033	BCL2/CCND1/STAT5A/F3	4
hsa05161	Hepatitis B	5/97	0.2584271	0.2199033	CCNE2/BCL2/CCNE1/CCND1/STAT5A	5
hsa04066	HIF-1 signaling pathway	4/97	0.2650179	0.2255116	IGF1R/ERBB2/BCL2/EGFR	4
hsa04213	Longevity regulating pathway - multiple species	3/97	0.2799484	0.2382164	IGF1R/ADCY4/ADCY9	3
hsa05214	Glioma	3/97	0.2823757	0.2402818	IGF1R/EGFR/CCND1	3
hsa05212	Pancreatic cancer	3/97	0.2848498	0.2423872	ERBB2/EGFR/CCND1	3
hsa05230	Central carbon metabolism in cancer	3/97	0.2873652	0.2445276	ERBB2/G6PD/EGFR	3
hsa04918	Thyroid hormone synthesis	3/97	0.3077277	0.2618547	ADCY4/ADCY9/IYD	3
hsa00051	Fructose and mannose metabolism	2/97	0.3077277	0.2618547	SORD/TPI1	2
hsa04215	Apoptosis - multiple species	2/97	0.3077277	0.2618547	PMAIP1/BCL2	2
hsa04917	Prolactin signaling pathway	3/97	0.3077277	0.2618547	ESR1/CCND1/STAT5A	3
hsa04520	Adherens junction	3/97	0.3203195	0.2725694	IGF1R/ERBB2/EGFR	3
hsa04610	Complement and coagulation cascades	3/97	0.3638123	0.3095787	F7/F3/CD59	3
hsa04151	PI3K-Akt signaling pathway	8/97	0.3638123	0.3095787	FGF2/IGF1R/CCNE2/BCL2/CCNE1/EGFR/CCND1/ITGB8	8
hsa05204	Chemical carcinogenesis	3/97	0.3784245	0.3220126	GSTP1/NAT1/EPHX1	3
hsa04068	FoxO signaling pathway	4/97	0.403965	0.3437458	IGF1R/EGFR/CCNB1/CCND1	4
hsa05162	Measles	4/97	0.41148	0.3501406	CCNE2/CCNE1/CCND1/STAT5A	4
hsa04540	Gap junction	3/97	0.4152732	0.3533683	ADCY4/ADCY9/EGFR	3
hsa04210	Apoptosis	4/97	0.4182539	0.3559047	PMAIP1/BCL2/PARP2/DAB2IP	4
hsa04912	GnRH signaling pathway	3/97	0.4182539	0.3559047	ADCY4/ADCY9/EGFR	3

hsa05032	Morphine addiction	3/97	0.4182539	0.3559047	ADCY4/ADCY9/ADORA1	3
hsa00983	Drug metabolism - other enzymes	2/97	0.4182539	0.3559047	NAT1/GMPS	2
hsa04340	Hedgehog signaling pathway	2/97	0.4182539	0.3559047	BCL2/CCND1	2
hsa04961	Endocrine and other factor-regulated calcium reabsorption	2/97	0.4182539	0.3559047	ESR1/ADCY9	2
hsa04261	Adrenergic signaling in cardiomyocytes	4/97	0.4457018	0.3792609	ADCY4/ADCY9/BCL2/CACNB2	4
hsa05203	Viral carcinogenesis	5/97	0.4457018	0.3792609	CCNE2/PMAIP1/CCNE1/CCND1/STAT5A	5
hsa04916	Melanogenesis	3/97	0.4601918	0.3915909	ADCY4/ADCY9/FZD10	3
hsa04390	Hippo signaling pathway	4/97	0.4601918	0.3915909	GDF5/CCND1/FZD10/DLG3	4
hsa04015	Rap1 signaling pathway	5/97	0.4601918	0.3915909	FGF2/IGF1R/ADCY4/ADCY9/EGFR	5
hsa04630	Jak-STAT signaling pathway	4/97	0.4697547	0.3997283	GFAP/BCL2/CCND1/STAT5A	4
hsa04810	Regulation of actin cytoskeleton	5/97	0.4697547	0.3997283	FGF2/EGFR/ITGB8/MYLK4/FGD3	5
hsa04725	Cholinergic synapse	3/97	0.5177031	0.440529	ADCY4/ADCY9/BCL2	3
hsa04022	cGMP-PKG signaling pathway	4/97	0.5177031	0.440529	ADCY4/ADCY9/ADORA1/MYLK4	4
hsa01200	Carbon metabolism	3/97	0.5256682	0.4473067	G6PD/MTHFR/TPI1	3
hsa05210	Colorectal cancer	2/97	0.5293595	0.4504477	BCL2/CCND1	2
hsa00910	Nitrogen metabolism	1/97	0.5293595	0.4504477	CA12	1
hsa00230	Purine metabolism	4/97	0.5433799	0.4623781	ADCY4/RRM2/ADCY9/GMPS	4
hsa05206	MicroRNAs in cancer	6/97	0.5433799	0.4623781	CCNE2/ERBB2/BCL2/CCNE1/EGFR/CCND1	6
hsa04270	Vascular smooth muscle contraction	3/97	0.5481454	0.4664332	ADCY4/ADCY9/MYLK4	3
hsa04611	Platelet activation	3/97	0.5576402	0.4745126	ADCY4/ADCY9/MYLK4	3
hsa00670	One carbon pool by folate	1/97	0.5675995	0.4829872	MTHFR	1
hsa04152	AMPK signaling pathway	3/97	0.5675995	0.4829872	IGF1R/ADIPOQ/CCND1	3
hsa00900	Terpenoid backbone biosynthesis	1/97	0.5827396	0.4958704	PDSS1	1
hsa00980	Metabolism of xenobiotics by cytochrome P450	2/97	0.5827396	0.4958704	GSTP1/EPHX1	2
hsa05220	Chronic myeloid leukemia	2/97	0.5827396	0.4958704	CCND1/STAT5A	2
hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	2/97	0.5827396	0.4958704	ITGB8/CACNB2	2

hsa04010	MAPK signaling pathway	5/97	0.5827396	0.4958704	FGF2/NTRK2/MAPT/EGFR/CACNB2	5
hsa01040	Biosynthesis of unsaturated fatty acids	1/97	0.5827396	0.4958704	ELOVL2	1
hsa04977	Vitamin digestion and absorption	1/97	0.596505	0.5075838	CUBN	1
hsa04024	cAMP signaling pathway	4/97	0.5993503	0.510005	ADCY4/ADCY9/ADORA1/NPY1R	4
hsa00062	Fatty acid elongation	1/97	0.5993503	0.510005	ELOVL2	1
hsa04925	Aldosterone synthesis and secretion	2/97	0.5993503	0.510005	ADCY4/ADCY9	2
hsa04550	Signaling pathways regulating pluripotency of stem cells	3/97	0.5993503	0.510005	FGF2/IGF1R/FZD10	3
hsa04310	Wnt signaling pathway	3/97	0.5993503	0.510005	SFRP1/CCND1/FZD10	3
hsa04742	Taste transduction	2/97	0.5993503	0.510005	ADCY4/TAS2R13	2
hsa05410	Hypertrophic cardiomyopathy (HCM)	2/97	0.5993503	0.510005	ITGB8/CACNB2	2
hsa04072	Phospholipase D signaling pathway	3/97	0.5993503	0.510005	ADCY4/ADCY9/EGFR	3
hsa04350	TGF-beta signaling pathway	2/97	0.5993503	0.510005	GDF5/CHRD	2
hsa04320	Dorso-ventral axis formation	1/97	0.5993503	0.510005	EGFR	1
hsa04911	Insulin secretion	2/97	0.5993503	0.510005	ADCY4/ADCY9	2
hsa00030	Pentose phosphate pathway	1/97	0.5993503	0.510005	G6PD	1
hsa05216	Thyroid cancer	1/97	0.5993503	0.510005	CCND1	1
hsa04727	GABAergic synapse	2/97	0.6029862	0.5130989	ADCY4/ADCY9	2
hsa04970	Salivary secretion	2/97	0.6029862	0.5130989	ADCY4/ADCY9	2
hsa00512	Mucin type O-Glycan biosynthesis	1/97	0.6029862	0.5130989	GALNT6	1
hsa04710	Circadian rhythm	1/97	0.6029862	0.5130989	CRY2	1
hsa04974	Protein digestion and absorption	2/97	0.6029862	0.5130989	MEP1A/COL27A1	2
hsa03410	Base excision repair	1/97	0.6236099	0.5306482	PARP2	1
hsa04713	Circadian entrainment	2/97	0.6236099	0.5306482	ADCY4/ADCY9	2
hsa04972	Pancreatic secretion	2/97	0.6236099	0.5306482	ADCY4/ADCY9	2
hsa00250	Alanine, aspartate and glutamate metabolism	1/97	0.6236099	0.5306482	ASNS	1
hsa00350	Tyrosine metabolism	1/97	0.6236099	0.5306482	PNMT	1
hsa04750	Inflammatory mediator regulation of TRP channels	2/97	0.6236099	0.5306482	ADCY4/ADCY9	2

hsa00040	Pentose and glucuronate interconversions	1/97	0.6236099	0.5306482	SORD	1
hsa03030	DNA replication	1/97	0.6236099	0.5306482	RFC4	1
hsa04723	Retrograde endocannabinoid signaling	2/97	0.6309377	0.5368837	ADCY4/ADCY9	2
hsa05231	Choline metabolism in cancer	2/97	0.6309377	0.5368837	SLC44A4/EGFR	2
hsa00240	Pyrimidine metabolism	2/97	0.6509745	0.5539335	RRM2/CTPS1	2
hsa00260	Glycine, serine and threonine metabolism	1/97	0.6509745	0.5539335	GNMT	1
hsa03050	Proteasome	1/97	0.6810774	0.579549	PSMB2	1
hsa04962	Vasopressin-regulated water reabsorption	1/97	0.6810774	0.579549	ADCY9	1
hsa04724	Glutamatergic synapse	2/97	0.6810774	0.579549	ADCY4/ADCY9	2
hsa04062	Chemokine signaling pathway	3/97	0.6810774	0.579549	ADCY4/ADCY9/PREX1	3
hsa03420	Nucleotide excision repair	1/97	0.6810774	0.579549	RFC4	1
hsa04144	Endocytosis	4/97	0.6810774	0.579549	CAV2/IGF1R/ERBB4/EGFR	4
hsa04919	Thyroid hormone signaling pathway	2/97	0.6810774	0.579549	ESR1/CCND1	2
hsa01212	Fatty acid metabolism	1/97	0.6810774	0.579549	ELOVL2	1
hsa04330	Notch signaling pathway	1/97	0.6810774	0.579549	APH1B	1
hsa04930	Type II diabetes mellitus	1/97	0.6810774	0.579549	ADIPOQ	1
hsa04722	Neurotrophin signaling pathway	2/97	0.6849385	0.5828345	NTRK2/BCL2	2
hsa04071	Sphingolipid signaling pathway	2/97	0.6853568	0.5831905	BCL2/ADORA1	2
hsa05014	Amyotrophic lateral sclerosis (ALS)	1/97	0.6874452	0.5849675	BCL2	1
hsa05110	Vibrio cholerae infection	1/97	0.6874452	0.5849675	ADCY9	1
hsa04978	Mineral absorption	1/97	0.6910739	0.5880553	CLCN2	1
hsa03460	Fanconi anemia pathway	1/97	0.7064525	0.6011414	UBE2T	1
hsa05217	basal cell carcinoma	1/97	0.7064525	0.6011414	FZD10	1
hsa05150	Staphylococcus aureus infection	1/97	0.7177272	0.6107354	DSG1	1
hsa00561	Glycerolipid metabolism	1/97	0.725004	0.6169275	MBOAT1	1
hsa04730	Long-term depression	1/97	0.725004	0.6169275	IGF1R	1
hsa05416	Viral myocarditis	1/97	0.725004	0.6169275	CCND1	1
hsa04924	Renin secretion	1/97	0.7425537	0.631861	ADORA1	1

hsa05321	Inflammatory bowel disease (IBD)	1/97	0.7425537	0.631861	GATA3	1
hsa00010	Glycolysis / Gluconeogenesis	1/97	0.7425537	0.631861	TPI1	1
hsa04014	Ras signaling pathway	3/97	0.7425537	0.631861	FGF2/IGF1R/EGFR	3
hsa05120	Epithelial cell signaling in Helicobacter pylori infection	1/97	0.7425537	0.631861	EGFR	1
hsa00982	Drug metabolism - cytochrome P450	1/97	0.7425537	0.631861	GSTP1	1
hsa04932	Non-alcoholic fatty liver disease (NAFLD)	2/97	0.7425537	0.631861	XBP1/ADIPOQ	2
hsa04622	RIG-I-like receptor signaling pathway	1/97	0.7425537	0.631861	TANK	1
hsa04920	Adipocytokine signaling pathway	1/97	0.7425537	0.631861	ADIPOQ	1
hsa00562	Inositol phosphate metabolism	1/97	0.7425537	0.631861	TPI1	1
hsa04150	mTOR signaling pathway	2/97	0.7425537	0.631861	IGF1R/FZD10	2
hsa03320	PPAR signaling pathway	1/97	0.7425537	0.631861	ADIPOQ	1
hsa01230	Biosynthesis of amino acids	1/97	0.7549872	0.642441	TPI1	1
hsa03018	RNA degradation	1/97	0.7560378	0.643335	MPHOSPH6	1
hsa04260	Cardiac muscle contraction	1/97	0.7560378	0.643335	CACNB2	1
hsa05100	Bacterial invasion of epithelial cells	1/97	0.7560378	0.643335	CAV2	1
hsa04141	Protein processing in endoplasmic reticulum	2/97	0.7595592	0.6463315	XBP1/BCL2	2
hsa05010	Alzheimer's disease	2/97	0.7614368	0.6479292	MAPT/APH1B	2
hsa04512	ECM-receptor interaction	1/97	0.7614368	0.6479292	ITGB8	1
hsa04640	Hematopoietic cell lineage	1/97	0.7850408	0.6680146	CD59	1
hsa05202	Transcriptional misregulation in cancer	2/97	0.7869707	0.6696568	IGF1R/ZBTB16	2
hsa04064	NF-kappa B signaling pathway	1/97	0.7965688	0.6778241	BCL2	1
hsa00564	Glycerophospholipid metabolism	1/97	0.7996365	0.6804345	MBOAT1	1
hsa05016	Huntington's disease	2/97	0.8103437	0.6895455	DNALI1/DNAH5	2
hsa04668	TNF signaling pathway	1/97	0.8438695	0.7180737	DAB2IP	1
hsa05145	Toxoplasmosis	1/97	0.8659659	0.7368762	BCL2	1
hsa04142	Lysosome	1/97	0.8715435	0.7416223	ARSG	1
hsa05160	Hepatitis C	1/97	0.8910478	0.758219	EGFR	1

hsa04514	Cell adhesion molecules (CAMs)	1/97	0.9130019	0.7769005	ITGB8	1
hsa04060	Cytokine-cytokine receptor interaction	2/97	0.9223335	0.7848409	GDF5/EGFR	2
hsa04080	Neuroactive ligand-receptor interaction	2/97	0.9330761	0.7939822	ADORA1/NPY1R	2
hsa03013	RNA transport	1/97	0.9347422	0.7953999	ELAC1	1
hsa05034	Alcoholism	1/97	0.9347422	0.7953999	NTRK2	1
hsa05152	Tuberculosis	1/97	0.9347422	0.7953999	BCL2	1
hsa05168	Herpes simplex infection	1/97	0.9369124	0.7972466	SRPK1	1
hsa01100	Metabolic pathways	12/97	0.9418448	0.8014437	RRM2/G6PD/MBOAT1/NAT1/SORD/CTPS1/PNMT/GMPS/ASNS/MTHFR/GALNT6/TPI1	12
hsa05169	Epstein-Barr virus infection	1/97	0.942435	0.8019459	BCL2	1

Appendix 2

Table 8-2. **A**nalysis using GeneCard and PubMed

Gene	Source:			
	GeneCards			PubMed
	Related pathways	**Diseases associated**	**Associated with breast cancer**	*** Number of Scientific articles published**
MLPH	Deregulation of Rab and Rab Effector Genes in Bladder Cancer	Griscelli Syndrome, Type 3 and Osteogenesis Imperfecta, Type Xv.	No	3
FOXA1	Embryonic and Induced Pluripotent Stem Cell Differentiation Pathways and Lineage-specific Markers and FOXA1 transcription factor network.	Estrogen-Receptor Positive Breast Cancer and Luminal Breast Carcinoma.	Yes	221
SIDT1	No data available	No data available	No	1
CEP55	Cytoskeletal Signaling and DNA Damage.	No data available	No	7
ASPM	No data available	Microcephaly 5, Primary, Autosomal Recessive and Autosomal Recessive Primary Microcephaly. upregulated in several types of cancer: in particular, brain tumors.	No. But Associated with cancer	8
CENPL	Mitotic Metaphase and Anaphase and Cell Cycle, Mitotic	Seckel Syndrome 1	No	0
AURKA	Integrated Breast Cancer Pathway and Regulation of PLK1 Activity at G2/M Transition	Colorectal Cancer and Colorectal Adenocarcinoma	No. But Associated with cancer	177
ESPL1	Mitotic Metaphase and Anaphase and Cell Cycle, Mitotic.	Fallopian Tube Disease and Salpingitis.	No	9
TTK	**RB in Cancer** and DNA Damage.	Chronic Polyneuropathy.	No. But Associated with cancer	70
UBE2T	Fanconi anemia pathway and Metabolism of proteins	Fanconi Anemia, Complementation Group T and Ube2t-Related Fanconi Anemia.	No	4
NCAPG	Cell cycle_Chromosome condensation in prometaphase and Aurora B signaling	No data available	No	1
GMPS	Metabolism and purine nucleotides de novo biosynthesis	Leukemia, Acute Myeloid	No	7
NDC80	Mitotic Metaphase and Anaphase and Aurora B signaling	Female Reproductive Organ Cancer.	No	9

MYBL2	HTLV-I infection and EGFR1 Signaling Pathway	Paraneoplastic Cerebellar Degeneration.	No	38
KIF18B	Vesicle-mediated transport and Factors involved in megakaryocyte development and platelet production	No data available	No	1
EXO1	Cell Cycle Checkpoints and Mismatch repair	Chilblain Lupus and Xeroderma Pigmentosum, Group G.	No	15
RERGL	No data available	No data available	No	0
FSIP1	No data available	Chromosome 3Q29 Duplication Syndrome.	No	2
ABCC11	Regulation of activated PAK-2p34 by proteasome mediated degradation and Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds	Apocrine gland secretion, variation in	No	22
PTX3	Immune System	Infectious Myocarditis and Hyper-Igd Syndrome.	No	13
FAT2	No data available	Skin Squamous Cell Carcinoma and Spinal Canal And Spinal Cord Meningioma	No. But Associated with cancer	3
POU4F1	Regulation of TP53 Activity and Gene Expression.	Cervical Cancer, Somatic and Papilloma.	No. But Associated with cancer	3
SKA1	Mitotic Metaphase and Anaphase and Cell Cycle, Mitotic.	No data available	No	2
MIA	Neural Crest Differentiation.	Skin Melanoma and Uveal Melanoma.	No	112
FOXC1	Transcriptional Regulatory Network in Embryonic Stem Cell and Heart Development.	Axenfeld-Rieger Syndrome, Type 3 and Iridogoniodysgenesis, Type 1.	No	43
TMEM45B	No data available	No data available	No	1
ABCG2	Metabolism and Statin Pathway - Generalized, Pharmacokinetics.	Erythroplakia and Placental Choriocarcinoma.	No	1369
ZBTB16	Immune System and Pathways in cancer	Skeletal Defects, Genital Hypoplasia, And Mental Retardation and Leukemia, Acute Promyelocytic, Somatic.	No	6
DEPDC1B	Signaling by GPCR and p75 NTR receptor-mediated signalling	No data available	No	1
CCNE1	Regulation of retinoblastoma protein and E2F mediated regulation of DNA replication.	Chronic Endophthalmitis and Facial Dermatosis.	No	90
BCAN	Cell adhesion_Cell-matrix glycoconjugates and Metabolism.	No data available	No	2

TBX19	Adrenocorticotropic Hormone Deficiency and Acth Deficiency.	Corticotropin-releasing hormone.	No	0
MUCL1	Immune System and HIV Life Cycle.	Diseases associated with MUCL1 include Breast Cancer.	Yes	12
TPI1	Metabolism and Glucose metabolism.	Hemolytic Anemia Due To Triosephosphate Isomerase Deficiency and Giardiasis.	No	3
ANP32E	No data available	No data available	No	5
KCMF1	Sweet Taste Signaling and Immune System	No data available	No	0
PSMB2	RET signaling and Regulation of activated PAK-2p34 by proteasome mediated degradation.	No data available	No	1
XBP1	HTLV-I infection and IgA-Producing B Cells in the Intestine.	Major Affective Disorder-7 and Bipolar Disorder.	No	60
GATA3	IL27-mediated signaling events and Regulation of nuclear SMAD2/3 signaling.	Hypoparathyroidism, Sensorineural Deafness, And Renal Dysplasia and Renal Dysplasia.	No	271
CA12	Metabolism and Nitrogen metabolism.	Hyperchlorhidrosis, Isolated and Renal Cell Carcinoma.	No	19
TBC1D9	No data available	No data available	No	3
PGR	Oocyte meiosis and Gene Expression.	Progesterone Resistance and Myoma.	No	1846
GREB1	No data available	Breast Cancer.	Yes	67
TFF1	Adhesion and Integrated Pancreatic Cancer Pathway.	Breast Cancer and Gastric Cancer.	Yes	487
AGR3	No data available	Breast Abscess.	No	8
NAT1	Drug metabolism - cytochrome P450 and Metabolism.	Ascending Cholangitis and Colorectal Adenoma.	No	70
ESR1	Regulation of nuclear SMAD2/3 signaling and **Integrated Breast Cancer Pathway.**	Estrogen Resistance and Migraine With Or Without Aura 1	Yes	581
AFF3	No data available	Fibular Aplasia.	No	4
GALNT6	Metabolism of proteins and Mucin type O-glycan biosynthesis.	No data available	No	7
AR	Regulation of nuclear SMAD2/3 signaling and Integrated Breast Cancer Pathway.	Androgen Insensitivity, Partial, With Or Without Breast Cancer and Androgen Insensitivity.	Yes	2116
TRIM29	Interferon gamma signaling and Immune System	Ataxia-Telangiectasia.	No	10

SFRP1	Wnt Signaling Pathways: beta-Catenin-dependent Wnt Signaling and Wnt Signaling Pathway and Pluripotency.	Glucocorticoid-Induced Osteoporosis and Meningothelial Meningioma.	No	73
CHI3L1	Immune System.	Asthma-Related Traits 7 and Schizophrenia.	No	43
RRM2	E2F mediated regulation of DNA replication and superpathway of pyrimidine deoxyribonucleotides de novo biosynthesis.	Choriocarcinoma and Pancreas Adenocarcinoma.	No	29
SRPK1	mRNA Splicing - Major Pathway and Influenza A	No data available	No	8
ADCY4	Signaling by GPCR and DAG and IP3 signaling.	No data available	No	0
NOSTRIN	Metabolism and eNOS activation and regulation.	Eclampsia	No	0
MAPT	Regulation of activated PAK-2p34 by proteasome mediated degradation and EphB-EphrinB Signaling.	Pick Disease and Dementia, Frontotemporal.	No	28
ERBB4	RET signaling and Activation of cAMP-Dependent PKA.	Amyotrophic Lateral Sclerosis 19 and Erbb4-Related Amyotrophic Lateral Sclerosis.	No	33
ASNS	Metabolism and Amino acid synthesis and interconversion (transamination).	Asparagine Synthetase Deficiency and Acute Lymphoblastic Leukemia, Childhood.	No	6
CSRNP1	No data available	No data available	No	1
CMTM7	No data available	No data available	No	0
FGD3	Signaling by GPCR and p75 NTR receptor-mediated signalling	Aarskog-Scott Syndrome.	No	1
THSD4	O-glycosylation of TSR domain-containing proteins and HIV Life Cycle.	No data available	No	1
BCL2	Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pathways and **Integrated Breast Cancer Pathway.**	Follicular Lymphoma 1 and Follicular Lymphoma.	Yes	579
SIAH2	Immune System and Class I MHC mediated antigen processing and presentation.	No data available	No	18
SCCPDH	Response to elevated platelet cytosolic Ca2+.	No data available	No	0
TCEAL1	No data available	No data available	No	2
GRB7	RET signaling and Cell surface interactions at the vascular wall.	Breast Cancer.	Yes	97
CCNE2	Mitotic G1-G1/S phases and GPCR Pathway.	No data available	No	29

CDKN3	No data available	Hepatocellular Carcinoma and Bannayan-Riley-Ruvalcaba Syndrome.	No	8
DSCC1	Gastric cancer network 2.	No data available	No	0
CENPN	Mitotic Metaphase and Anaphase and Cell Cycle, Mitotic.	No data available	No	1

Appendix 3

Code: Extract rules [95]

```
#*************************
#return the rules of a tree
#*************************
getConds<-function(tree){
  #store all conditions into a list
  conds<-list()
  #start by the terminal nodes and find previous conditions
  id.leafs<-which(tree$status==-1)
         j<-0
         for(i in id.leafs){
              j<-j+1
              prevConds<-prevCond(tree,i)
              conds[[j]]<-prevConds$cond
              while(prevConds$id>1){
                prevConds<-prevCond(tree,prevConds$id)
                conds[[j]]<-paste(conds[[j]]," & ",prevConds$cond)
                if(prevConds$id==1){
                     conds[[j]]<-paste(conds[[j]]," => ",tree$prediction[i])
       break()
      }
    }
  }
  return(conds)
}
#*************************
#find the previous conditions in the tree
#*************************
prevCond<-function(tree,i){
  if(i %in% tree$right_daughter){
              id<-which(tree$right_daughter==i)
              cond<-paste(tree$split_var[id],">",tree$split_point[id])
         }
         if(i %in% tree$left_daughter){
    id<-which(tree$left_daughter==i)
              cond<-paste(tree$split_var[id],"<",tree$split_point[id])
  }
  return(list(cond=cond,id=id))
}
#remove spaces in a word
collapse<-function(x){
  x<-sub(" ","_",x)
  return(x)
}
data(data_name)
require(randomForest)
mod.rf <- randomForest(Species ~ ., data=data_name)
tree<-getTree(mod.rf, k=1, labelVar=TRUE)
#rename the name of the column
colnames(tree)<-sapply(colnames(tree),collapse)
rules<-getConds(tree)
print(rules)
```

Appendix 4

Code: IBF-RF metric

```
ntree = 5
# for save Results
mod.rf <- randomForest(Species~.,data= iris, ntree=5)

for (t in 1:ntree)
{
  tree<-getTree(mod.rf, k=t, labelVar=TRUE)
  #rename the name of the column
  colnames(tree)<-sapply(colnames(tree),collapse)
  rules<-getConds(tree)
  assign(paste("rules",sep="_",t),rules)
}

matrix_tree = matrix(ncol=7,nrow=1, dimnames = list( c("row1"), c("Tree_Num",
"Rules", "Subtype_pred", "Cond_id", "Gene", ">/<", "Treshold" )))
maux=matrix(ncol=7,nrow=1)

vecaux=matrix(ncol=3,nrow=1)
vecsort=matrix(ncol=1,nrow=1)
vecsort2=matrix(ncol=1,nrow=1)

for(x in 1:ntree){
    aux_rules_for=get(paste("rules",sep="_",x))
  inter_subttype = as.data.frame(strsplit(as.character(aux_rules_for), "=>"))

  for (q in 1:ncol(inter_subttype)){
    num_ampersand = nchar(as.character(inter_subttype[1,q])) -
nchar(as.character(gsub("&","",inter_subttype[1,q])))
    num_ampersand = num_ampersand+1
    inter_genes = as.data.frame(strsplit(as.character(inter_subttype[1,q]), "&
"))
    for(w in 1:num_ampersand){
      inter_conditions =
as.data.frame(strsplit(as.character(inter_genes[w,1]), " "))

maux=c(x,q,as.character(inter_subttype[2,q]),w,as.character(inter_conditions[1
,1]),as.character(inter_conditions[2,1]),as.character(inter_conditions[3,1]))
      matrix_tree=rbind(matrix_tree,maux)
      maux=matrix(ncol=7,nrow=1)
      vecsort = rbind(vecsort, as.character(inter_conditions[1,1]))
    }
    vecsort2 = sort(vecsort[-1,])
    vecsort2
    a = matrix(ncol=1,nrow=1)

    for (r in 1:nrow(as.data.frame(vecsort2))){
```

```
      m=duplicated(as.data.frame(vecsort2))
      if (m[r] == FALSE){
        a = cbind(a,r)
      }
    }
    vecsort2 = vecsort2[a[,-1]]
    a = matrix(ncol=1,nrow=1)

    cbind(x,paste(vecsort2, collapse="-"))
    vecaux =rbind(vecaux, cbind(x, paste(vecsort2, collapse="-
"),as.character(inter_subttype[2,q])))
    vecsort=matrix(ncol=1,nrow=1)
    vecsort2=matrix(ncol=1,nrow=1)

  }
}
```

Appendix 5

Table 8-3. Results IBF-RF metric for TCGA data (211 important variables)

#	Row Labels	Basal	HER2	LumA	LumB	normal	Frec. Total
1	MLPH	237					237
2	FOXA1	109					109
3	FOXA1-UBE2T	9				7	16
4	C22orf23-MLPH	7			7		14
5	DEPDC1B-MLPH	7			6		13
6	AGR3-MLPH	11			2		13
7	ABCC11-FOXA1	10				3	13
8	FOXC1-MLPH	11		1			12
9	ESPL1-FOXA1	7				5	12
10	ABCC11-FOXC1	12					12
11	FOXA1-TTK	9				2	11
12	FOXC1-THSD4	11					11
13	ABCC11-XBP1	11					11
14	FOXC1-KIF18B	6		4			10
15	MICALL1-TFF1	5			5		10
16	FOXA1-MYBL2	5				5	10
17	FOXA1-GMPS	5				5	10
18	FOXC1	10					10
19	FOXA1-KIF18B	5				5	10
20	FOXA1-TCEAL1	8				1	9
21	FOXA1-PSMB2	5				4	9
22	ESR1-MLPH	9					9
23	CCNE1-FOXA1	6				3	9
24	ANP32E-FOXC1	6		2	1		9
25	GSTP1-TFF1	5			3		8
26	FOXA1-TBC1D9	5	2			1	8
27	MIA-MLPH	6			2		8
28	MLPH-SFRP1	6			2		8
29	FOXC1-PPP1R14C	4	4				8
30	MLPH-SRPK1	4			4		8

Table 8-4. Results IBF-RF metric for GSE20685 data (211 important variables)

#	Row Labels	Basal	HER2		LumB	LumA		Frec. Total
		type I	type II	type III	type IV	type V	type VI	
1	FOXC1	107						107
2	FOXC1-THSD4	76		3			16	95
3	MLPH-NOSTRIN	54	2	28	3			87
4	CENPL-FOXA1	40		31				71
5	ASPM-MLPH	46		15				61
6	ADCY4-MLPH	37		20				57
7	MLPH	56						56
8	CEP55-FOXA1	41		14				55
9	CA12-FOXC1	37	1				13	51
10	AFF3-FOXC1	30					18	48
11	FOXC1-NOSTRIN	35		1			11	47
12	ARSG-SIDT1	38	3	5				46
13	ASPM-FOXC1	36					10	46
14	FOXA1-TTK	38		8				46
15	MLPH-PDSS1	30		15			1	46
16	MLPH-TTK	29		16				45
17	FOXA1-FOXC1	27					15	42
18	FOXC1-GREB1	28					13	41
19	ASPM-FOXA1	33		6				39
20	FOXA1-GMPS	31		7				38
21	FOXC1-KIF18B	31					7	38
22	FOXC1-MAPT	20					17	37
23	GMPS-MLPH	29		8				37
24	TBC1D9	37						37
25	SIDT1	35						35
26	FOXA1-ZBTB16	14		20				34
27	AR	33						33
28	AR-FOXC1	20		2	1		10	33
29	CENPL-MLPH	21		12				33
30	CENPN-FOXC1	24					8	32

Table 8-5. Results IBF-RF metric for GSE21653 data (211 important variables)

#	Row Labels	basal	HER2	LumA	LumB	normal	Frec. Total
1	CEP55-FOXA1	35				19	54
2	C22orf23-FOXA1	30				17	47
3	ASPM-FOXA1	27				20	47
4	FOXA1-TTK	25				20	45
5	AURKA-FOXA1	34				9	43
6	FOXA1-NCAPG	24				11	35
7	ESPL1-FOXA1	25				8	33
8	FOXA1-SIAH2	18				9	27
9	FOXA1-MYBL2	23				3	26
10	MLPH-TTK	10				15	25
11	CEP55-MLPH	12				12	24
12	DSCC1-MLPH	6				18	24
13	FOXA1	22					22
14	FOXA1-TPI1	16				5	21
15	KCMF1-MLPH	9				12	21
16	EXO1-FOXA1	12				7	19
17	MLPH-PSMB2	9				10	19
18	FOXA1-PPP1R14C	14				4	18
19	MLPH-MYBL2	14				3	17
20	AURKA-MLPH	13				4	17
21	FOXA1-UBE2T	11				6	17
22	FOXA1-KCMF1	8				8	16
23	CCNB1-FOXA1	12				3	15
24	FOXA1-NDC80	14					14
25	SIDT1	14					14
26	MLPH-TPI1	6				8	14
27	CCNE1-MLPH	11				2	13
28	FOXA1-SKA1	10				3	13
29	MLPH-RRM2	9				4	13
30	MLPH-SIAH2	6				7	13

Appendix 6

Table 8-6. Common rules between TCGA, GSE21653 and GSE 20685 datasets, extracted through IBF-RF metric.

#	Rules	TCGA						GSE21653						GSE20685							Grand Total
														basal	HER2		LumB	LumA		Total	
		Basal	HER2	LumA	LumB	normal	Total	basal	HER2	LumA	LumB	normal	Total	typeI	typeII	typeIII	typeIV	typeV	typeVI		
1	MLPH	237					237	8					8	56						56	301
2	FOXA1	109					109	22					22	8						8	139
3	CEP55-FOXA1	3					3	35				19	54	41		14				55	112
4	FOXC1-THSD4	11					11					2	2	76		3			16	95	108
5	FOXA1-TTK	9				2	11	25				20	45	38		8				46	102
6	MLPH-NOSTRIN	1					1	2					2	54	2	28	3			87	90
7	ASPM-FOXA1	2					2	27				20	47	33		6				39	88
8	CENPL-FOXA1	5					5	6				5	11	40		31				71	87
9	AURKA-FOXA1	4				2	6	34				9	43	18		6				24	73
10	MLPH-TTK	2			1		3	10				15	25	29		16				45	73
11	ESPL1-FOXA1	7				5	12	25				8	33	10		8				18	63
12	ASPM-FOXC1	2		1			3	5				5	10	36					10	46	59
13	FOXA1-GMPS	5				5	10	9					9	31		7				38	57
14	CEP55-MLPH	4			1		5	12				12	24	16		9				25	54
15	FOXC1-KIF18B	6		4			10	2				2	4	31					7	38	52
16	FOXC1-NOSTRIN	1					1	2					2	35		1			11	47	50
17	SIDT1	1					1	14					14	35						35	50
18	FOXA1-UBE2T	9				7	16	11				6	17	10		6				16	49
19	FOXA1-NCAPG	4				3	7	24				11	35	2		3				5	47

20	FOXA1-NDC80	2					2	14					14	26	1	2	2			31	47
21	CENPL-MLPH	2			2		4	6				2	8	21		12				33	45
22	FOXA1-FOXC1	1					1	1				1	2	27					15	42	45
23	EXO1-FOXA1	4					4	12				7	19	13		7				20	43
24	KIF18B-MLPH	2			1		3	5				2	7	20		10				30	40
25	FOXA1-MYBL2	5				5	10	23				3	26	3						3	39
26	FOXA1-ZBTB16	1					1	4					4	14		20				34	39
27	AR-FOXC1	3					3	2					2	20		2	1		10	33	38
28	EXO1-MLPH	5			2		7	3				5	8	15		8				23	38
29	FOXA1-KIF18B	5				5	10	3				1	4	17		7				24	38
30	AURKA-MLPH	5					5	13				4	17	11		3				14	36
31	MLPH-SIDT1	5					5					2	2	18	3	8				29	36
32	AURKA-FOXC1	3		1			4	2				1	3	24		2			2	28	35
33	MLPH-SRPK1	4			4		8	1				2	3	13	2	9				24	35
34	CDKN3-FOXA1	6				2	8	3				4	7	6		13				19	34
35	FOXC1-TFF1	4			2		6	1					1	15					12	27	34
36	FOXA1-TBC1D9	5	2			1	8	2					2	14	3	6				23	33
37	FOXA1-SIAH2	2				2	4	18				9	27	1						1	32
38	MLPH-UBE2T	2				2	4	4				2	6	11		11				22	32
39	AR-MLPH	3					3	4			6	1	11	11		6				17	31
40	MLPH-TCEAL1	1					1	1			2		3	12		14				26	30
41	FOXA1-SKA1	1					1	10				3	13	9		4				13	27
42	MLPH-TBC1D9	2					2	2			1		3	12		10				22	27
43	DEPDC1B-FOXA1	4				4	8	4				4	8	7		3				10	26
44	FOXA1-GALNT6	1	1				2	1				4	5	5		14				19	26
45	FOXA1-TPI1					1	1	16				5	21	1		2				3	25

46	FOXC1-MLPH	11		1			12	1					1	6		1			5	12	25
47	FOXC1-SKA1	4		2			6	2					2	9					7	16	24
48	CEP55-FOXC1	3		1			4	1				1	2	16					1	17	23
49	CCNE1-MLPH	3			2		5	11				2	13	2		1			1	4	22
50	DEPDC1B-MLPH	7			6		13	2				3	5	4						4	22
51	ANP32E-SIDT1	1					1		1				1	11		7	1			19	21
52	AR-TBC1D9	1					1	2			1		3	12	2	2	1			17	21
53	CENPN-MLPH	1			1		2	1				3	4	9		6				15	21
54	FOXA1-KCMF1	2					2	8				8	16	3						3	21
55	MLPH-RRM2	2			1		3	9				4	13	5						5	21
56	FOXC1-NDC80	1		1			2	2					2	14					2	16	20
57	ABCC11-FOXC1	12					12	1					1	3					3	6	19
58	CCNE1-FOXC1	3	1	2			6	1					1	8					4	12	19
59	CDKN3-MLPH	1					1	4				7	11	2		4				6	18
60	FOXC1-SIDT1	1					1	2					2	11	1				3	15	18
61	MLPH-MUCL1	1			1		2	1					1	11		4				15	18
62	MLPH-PGR	4					4	6					6	8						8	18
63	AGR3-FOXC1	2					2	1					1	8					6	14	17
64	MIA-MLPH	6			2		8					6	6	2					1	3	17
65	MLPH-NCAPG	1					1	9				2	11	4		1				5	17
66	CCNE1-FOXA1	6				3	9	4				1	5	1		1				2	16
67	DSCC1-FOXC1	1					1	1					1	10					4	14	16
68	FOXC1-MYBL2	1		1			2	6					6	6		1			1	8	16
69	GATA3-MLPH	8					8	1			1		2	6						6	16
70	MLPH-TRIM29	5			2		7	4				3	7	1		1				2	16
71	CCNE2-FOXA1	1					1	11					11	3						3	15

72	CDKN3-FOXC1	4		1			5	4				1	5	4		1				5	15
73	FOXC1-GMPS	2		1		1	4	3					3	7						7	14
74	AR-XBP1	3					3	2			2	2	6	3		1				4	13
75	ESR1-MLPH	9					9	1					1	1	1					2	12
76	FOXA1-RERGL	1				1	2	1					1	7		2				9	12
77	FOXA1-RRM2	1				1	2	7				2	9	1						1	12
78	SIDT1-XBP1	1					1	1				1	2	3	3	3				9	12
79	ABCG2-FOXA1	2					2	1				1	2	4		3				7	11
80	AR-TFF1	1					1	5			1	1	7	1	1		1			3	11
81	AR-THSD4	1					1	3			3		6	4						4	11
82	CA12-MLPH	3					3	1			1	1	3	3		2				5	11
83	FAT2-MLPH	4			2		6	1				1	2	1		2				3	11
84	MLPH-THSD4	5					5				1		1	4		1				5	11
85	PTX3-TBC1D9	3	1				4	2					2	2	3					5	11
86	ASNS-FOXA1	1					1	1					1	8						8	10
87	BCAN-MLPH	2			1		3	1					1	2		4				6	10
88	SIDT1-THSD4	1					1	1					1	6		1	1			8	10
89	ANP32E-TBC1D9	2					2	3					3	4						4	9
90	CA12-FOXA1	4				3	7	1					1	1						1	9
91	CEP55-FOXC1-GATA3	1		1		1	3	3					3	2					1	3	9
92	CMTM7-TTK	1					1	5					5	3						3	9
93	FOXA1-TFF1	1					1	7					7	1						1	9
94	FOXC1-TBX19	1					1	1	1				2	5					1	6	9
95	AR-SFRP1-THSD4	1	1				2	1					1	1		4				5	8
96	BCL2-MLPH	1					1	4				2	6	1						1	8
97	CEP55-ESR1-FOXC1		1			2	3	2	1				3	2						2	8

98	ERBB4-MLPH	2			1		3	1	1		1	1	4	1						1	8
99	FOXA1-GATA3-GMPS	1					1	2					2	4		1				5	8
100	FOXC1-PSMB2					1	1	2					2	2		2			1	5	8
101	AFF3-MLPH	1					1	3			2		5	1						1	7
102	AR-GALNT6	1					1	1					1	4			1			5	7
103	BCL2-FOXC1	1					1	1				1	2	4						4	7
104	FOXA1-SCCPDH	1					1	5					5			1				1	7
105	FOXC1-GATA3-MYBL2	1		1			2	1	1	1			3	1					1	2	7
106	GREB1-MLPH	1					1	2				1	3	3						3	7
107	MAPT-MLPH	2			2		4				2		2	1						1	7
108	ADCY4-MLPH-THSD4	1		1			2	1					1	2		1				3	6
109	AR-FOXA1	2					2	1			1		2	2						2	6
110	ASPM-CMTM7	2					2	3					3	1						1	6
111	ASPM-FOXC1-TBX19	1					1		1				1	1		1			2	4	6
112	CA12-ESPL1-FOXC1	1				1	2	1	1			1	3					1		1	6
113	CDKN3-FOXA1-SKA1		1			1	2	1				1	2	1		1				2	6
114	ESR1-FOXA1	3					3	1					1	1	1					2	6
115	FOXA1-SFRP1	1					1	1				1	2	1		2				3	6
116	FOXA1-SKA1-TBC1D9	1					1					1	1			3			1	4	6
117	PTX3-THSD4	2					2	1					1	3						3	6
11	ABCC11-MLPH	3					3	1					1	1						1	5

8																					
119	ANP32E-ESR1				1		1	1					1	3						3	5
120	ANP32E-NAT1	1					1	1					1	3						3	5
121	AURKA-ESR1-SIDT1			1			1	1					1	1		1			1	3	5
122	CENPL-FOXA1-THSD4	1				1	2	1					1	1					1	2	5
123	CEP55-FOXC1-FSIP1	1					1	2					2	2						2	5
124	ESR1-FOXC1-NCAPG	1	1				2	1	1				2	1						1	5
125	FOXA1-GREB1	1					1	2				1	3	1						1	5
126	FOXC1-NCAPG	2					2	2					2	1						1	5
127	FOXC1-RRM2-THSD4			1		1	2					1	1			1			1	2	5
128	GMPS-MLPH-SFRP1	1					1	1				1	2	1		1				2	5
129	ABCC11-ESR1-PTX3	1	1				2					1	1	1						1	4
130	ADCY4-ASPM-TBC1D9			1			1		1				1	1	1					2	4
131	ASPM-CMTM7-ESR1	1			1		2	1					1	1						1	4
132	AURKA-FOXA1-GATA3	1					1	1					1	1		1				2	4
133	AURKA-PTX3-TBC1D9	2					2					1	1	1						1	4
134	CEP55-FOXC1-SFRP1			1			1	1			1		2	1						1	4
135	CEP55-SFRP1-THSD4					1	1					1	1				1		1	2	4
136	CHI3L1-TBC1D9	1	1				2	1					1				1			1	4
137	ESR1-EXO1-FOXA1	1		1			2	1					1		1					1	4
138	ESR1-FOXC1-			1			1					1	1	1	1					2	4

8	RERGL																				
139	ESR1-MLPH-NCAPG	1					1	1			1		2	1						1	4
140	ESR1-SIDT1	2					2	1					1	1						1	4
141	MLPH-POU4F1	1					1					1	1	1		1				2	4
142	ABCC11-EXO1-GATA3				1		1					1	1			1				1	3
143	AFF3-FOXC1-NCAPG			1			1	1					1	1						1	3
144	AURKA-ESR1-TMEM45B	1					1	1					1	1						1	3
145	AURKA-GATA3-GRB7		1				1	1					1						1	1	3
146	AURKA-GATA3-MLPH	1					1	1					1	1						1	3
147	BCL2-SFRP1	1					1	1					1	1						1	3
148	CEP55-GATA3					1	1				1		1				1			1	3
149	CSRNP1-FOXA1	1					1					1	1			1				1	3
150	ERBB4-SIDT1	1					1	1					1	1						1	3
151	ESPL1-FGD3			1			1			1			1				1			1	3
152	ESR1-FOXC1-SFRP1	1					1	1					1	1						1	3
153	ESR1-UBE2T			1			1			1			1				1			1	3
154	FOXC1-MAPT-TTK			1			1	1					1	1						1	3
155	KCMF1-XBP1					1	1	1					1	1						1	3
156	NDC80-PTX3	1					1	1					1	1						1	3

Appendix 7

Table 8-7. Correlation of Block 1 Genes

#	Gene 1	Gene 2	20685		21653		TCGA	
			Correlation	p-value	Correlation	p-value	Correlation	p-value
1	MLPH	FOXA1	0.8596154	0.0000	0.8922986	0.0000	0.6714588	0.00E+00
2	MLPH	SIDT1	0.7225649	0.0000	0.7296534	0.0000	0.2967918	7.16E-13
3	FOXA1	SIDT1	0.7407922	0.0000	0.7794672	0.0000	0.4228451	0.00E+00

Table 8-8. Correlation of Block 2 Genes

#	Gene 1	Gene 2	20685		21653		TCGA	
			Correlation	p-value	Correlation	p-value	Correlation	p-value
1	CEP55	ASPM	0.8575111	0.0000	0.8745455	0.0000	0.6356562	0.0000
2	CEP55	CENPL	0.6982602	0.0000	0.6321183	0.0000	0.5479947	0.0000
3	ASPM	CENPL	0.7683048	0.0000	0.7216883	0.0000	0.7473495	0.0000
4	CEP55	AURKA	0.8265305	0.0000	0.7785445	0.0000	0.4995046	0.0000
5	ASPM	AURKA	0.8047447	0.0000	0.8187281	0.0000	0.532519	0.0000
6	CENPL	AURKA	0.6607924	0.0000	0.6851631	0.0000	0.4424389	0.0000
7	CEP55	ESPL1	0.8096087	0.0000	0.7428849	0.0000	0.514083	0.0000
8	ASPM	ESPL1	0.8142705	0.0000	0.8092953	0.0000	0.7076613	0.0000
9	CENPL	ESPL1	0.6276137	0.0000	0.6605905	0.0000	0.5752906	0.0000
10	AURKA	ESPL1	0.8078814	0.0000	0.8147408	0.0000	0.5888395	0.0000
11	CEP55	TTK	0.8486258	0.0000	0.8703485	0.0000	0.6533323	0.0000
12	ASPM	TTK	0.8492475	0.0000	0.8657433	0.0000	0.7475246	0.0000
13	CENPL	TTK	0.712594	0.0000	0.706519	0.0000	0.7095809	0.0000
14	AURKA	TTK	0.7966269	0.0000	0.7902278	0.0000	0.5739677	0.0000
15	ESPL1	TTK	0.7351018	0.0000	0.7516126	0.0000	0.5882558	0.0000
16	CEP55	UBE2T	0.7479331	0.0000	0.8359129	0.0000	0.5720968	0.0000
17	ASPM	UBE2T	0.8200321	0.0000	0.8326855	0.0000	0.6456034	0.0000
18	CENPL	UBE2T	0.7121803	0.0000	0.6981502	0.0000	0.657473	0.0000
19	AURKA	UBE2T	0.7496623	0.0000	0.7493184	0.0000	0.5711274	0.0000
20	ESPL1	UBE2T	0.7610574	0.0000	0.7303998	0.0000	0.6151175	0.0000
21	TTK	UBE2T	0.6966261	0.0000	0.7893964	0.0000	0.6657555	0.0000
22	CEP55	NCAPG	0.8545552	0.0000	0.8082357	0.0000	0.6624174	0.0000

23	ASPM	NCAPG	0.8506741	0.0000	0.8444412	0.0000	0.7237227	0.0000
24	CENPL	NCAPG	0.7031494	0.0000	0.7685625	0.0000	0.6086388	0.0000
25	AURKA	NCAPG	0.813024	0.0000	0.8418129	0.0000	0.6452357	0.0000
26	ESPL1	NCAPG	0.8088861	0.0000	0.8168438	0.0000	0.6253858	0.0000
27	TTK	NCAPG	0.8510467	0.0000	0.8214419	0.0000	0.7594619	0.0000
28	UBE2T	NCAPG	0.7231205	0.0000	0.7657888	0.0000	0.6975421	0.0000
29	CEP55	GMPS	0.6959857	0.0000	0.7231992	0.0000	0.5854985	0.0000
30	ASPM	GMPS	0.6905887	0.0000	0.6831183	0.0000	0.5959207	0.0000
31	CENPL	GMPS	0.6342785	0.0000	0.6261296	0.0000	0.592883	0.0000
32	AURKA	GMPS	0.6387671	0.0000	0.6606353	0.0000	0.4864475	0.0000
33	ESPL1	GMPS	0.5950142	0.0000	0.5523741	0.0000	0.441224	0.0000
34	TTK	GMPS	0.7049798	0.0000	0.7148428	0.0000	0.7341886	0.0000
35	UBE2T	GMPS	0.6036569	0.0000	0.7320452	0.0000	0.582425	0.0000
36	NCAPG	GMPS	0.6605923	0.0000	0.686573	0.0000	0.659225	0.0000
37	CEP55	NDC80	0.8094364	0.0000	0.8024632	0.0000	0.6846841	0.0000
38	ASPM	NDC80	0.8418813	0.0000	0.841132	0.0000	0.6642475	0.0000
39	CENPL	NDC80	0.6675979	0.0000	0.7153549	0.0000	0.6255195	0.0000
40	AURKA	NDC80	0.7658231	0.0000	0.7851039	0.0000	0.5257431	0.0000
41	ESPL1	NDC80	0.7385113	0.0000	0.7906541	0.0000	0.5359277	0.0000
42	TTK	NDC80	0.8321689	0.0000	0.8425137	0.0000	0.7311886	0.0000
43	UBE2T	NDC80	0.6674397	0.0000	0.7502601	0.0000	0.6256781	0.0000
44	NCAPG	NDC80	0.8547555	0.0000	0.8685311	0.0000	0.754773	0.0000
45	GMPS	NDC80	0.6063186	0.0000	0.6932169	0.0000	0.5695946	0.0000
46	CEP55	MYBL2	0.8261436	0.0000	0.7577678	0.0000	0.5584955	0.0000
47	ASPM	MYBL2	0.7794561	0.0000	0.7827708	0.0000	0.5170572	0.0000
48	CENPL	MYBL2	0.6229741	0.0000	0.603065	0.0000	0.4357157	0.0000
49	AURKA	MYBL2	0.7458181	0.0000	0.8118801	0.0000	0.6502274	0.0000
50	ESPL1	MYBL2	0.7726727	0.0000	0.8070341	0.0000	0.5678579	0.0000
51	TTK	MYBL2	0.7763063	0.0000	0.7586118	0.0000	0.5944671	0.0000
52	UBE2T	MYBL2	0.6954727	0.0000	0.6995077	0.0000	0.4932848	0.0000
53	NCAPG	MYBL2	0.7965997	0.0000	0.7724018	0.0000	0.591174	0.0000
54	GMPS	MYBL2	0.5962744	0.0000	0.639658	0.0000	0.5497663	0.0000
55	NDC80	MYBL2	0.749793	0.0000	0.7561672	0.0000	0.5645452	0.0000
56	CEP55	KIF18B	0.8600357	0.0000	0.7641194	0.0000	0.5998931	0.0000
57	ASPM	KIF18B	0.8671564	0.0000	0.8248792	0.0000	0.6493994	0.0000
58	CENPL	KIF18B	0.6773531	0.0000	0.6068884	0.0000	0.592929	0.0000
59	AURKA	KIF18B	0.8433043	0.0000	0.76702	0.0000	0.5921999	0.0000
60	ESPL1	KIF18B	0.8657786	0.0000	0.854904	0.0000	0.686405	0.0000
61	TTK	KIF18B	0.8192439	0.0000	0.7793224	0.0000	0.6836513	0.0000

62	UBE2T	KIF18B	0.760567	0.0000	0.6877987	0.0000	0.6709445	0.0000
63	NCAPG	KIF18B	0.8480407	0.0000	0.7728841	0.0000	0.7418874	0.0000
64	GMPS	KIF18B	0.6555279	0.0000	0.5706215	0.0000	0.5711871	0.0000
65	NDC80	KIF18B	0.8257414	0.0000	0.7893076	0.0000	0.7234713	0.0000
66	MYBL2	KIF18B	0.7871949	0.0000	0.759387	0.0000	0.6425187	0.0000
67	CEP55	EXO1	0.7966347	0.0000	0.841249	0.0000	0.6175829	0.0000
68	ASPM	EXO1	0.8517266	0.0000	0.8625618	0.0000	0.7424684	0.0000
69	CENPL	EXO1	0.673193	0.0000	0.7154106	0.0000	0.6975411	0.0000
70	AURKA	EXO1	0.7403416	0.0000	0.7647459	0.0000	0.5670555	0.0000
71	ESPL1	EXO1	0.7428094	0.0000	0.7614284	0.0000	0.6116734	0.0000
72	TTK	EXO1	0.7634121	0.0000	0.8409662	0.0000	0.7612638	0.0000
73	UBE2T	EXO1	0.77767	0.0000	0.8537791	0.0000	0.7205	0.0000
74	NCAPG	EXO1	0.7501238	0.0000	0.8053746	0.0000	0.721039	0.0000
75	GMPS	EXO1	0.6523897	0.0000	0.7578817	0.0000	0.6914734	0.0000
76	NDC80	EXO1	0.7286817	0.0000	0.7795468	0.0000	0.6910663	0.0000
77	MYBL2	EXO1	0.721223	0.0000	0.778706	0.0000	0.5901322	0.0000
78	KIF18B	EXO1	0.7974898	0.0000	0.7596262	0.0000	0.7302604	0.0000