Classifying Breast Cancer Molecular Subtypes using Deep Clustering Approach

Narjes Rohani
Shahid Beheshti University
Changiz Eslahchi (✉ Ch-eslahchi@sbu.ac.ir)
Shahid Beheshti University https://orcid.org/0000-0002-8913-3904

Research article

Keywords: Machine learning, Cancer, Molecular subtypes, Breast cancer, Tumor classication, Cancer heterogeneity

Posted Date: December 23rd, 2019

DOI: https://doi.org/10.21203/rs.2.19530/v1

Version of Record: A version of this preprint was published at Frontiers in Genetics on November 25th, 2020. See the published version at https://doi.org/10.3389/fgene.2020.553587.

RESEARCH

Classifying Breast Cancer Molecular Subtypes using Deep Clustering Approach

Narjes Rohani[1†] and Changiz Eslahchi[1,2*]

[*]Correspondence:
ch-eslahchi@sbu.ac.ir
[1]Department of Mathematics, Shahid-Beheshti University, GC, Tehran, Iran
[2]Department of Biological Sciences, Institute for Research in Fundamental Sciences, GC, Tehran, Iran
Full list of author information is available at the end of the article
[†]Contact mail:
n.rohani@mail.sbu.ac.ir

Abstract

Background: Cancer is a complex disease with a high rate of mortality. The characteristics of tumor masses are very heterogeneous; thus, the appropriate classification of tumors is a critical point in the correct treatment. A high level of heterogeneity has also been observed in breast cancer. Therefore, detecting the molecular subtypes of this disease is a worthwhile issue for medicine that could be facilitated using bioinformatics.

Method: Numerous methods have already classified breast cancer based on gene expression data; however, they are not reliable due to the dynamic nature of these data. In contrast, gene mutation data are relatively stable and may lead to better classification. The aim of this study is to introduce a novel method for detecting the molecular subtypes of breast cancer. In this study, somatic mutation profiles of tumors are used; nonetheless, the somatic mutation profiles are very sparse. To address this issue, we made use of the network propagation method on gene interaction network and made the mutation profiles dense. Afterward, we used deep embedded clustering (DEC) method to classify breast tumors into four subtypes. In the next step, gene signatures of each subtype obtained by Fisher exact test and Benjamini-Hochberg procedure.

Results: Clinical and molecular analyses are executed, besides enrichment of results in numerous databases have shown that the proposed method, using mutation profiles can efficiently detect the molecular subtypes of breast cancer. Finally, a supervised classifier is proposed based on discovered subtypes to predict the molecular subtype of a new patient.

Keywords: Machine learning; Cancer; Molecular subtypes; Breast cancer; Tumor classification; Cancer heterogeneity

1 Introduction

Breast cancer is a heterogeneous disease at the molecular and clinical level; thus, the effectiveness of a treatment is hugely different based on the patients. This heterogeneity is a challenge for tumor classification to reach an appropriate clinical outcome. To solve this problem, researcher developed numerous methods to classify tumor masses, such as histopathological classification based on the morphological characteristics or immunohistochemical (IHC) markers like estrogen receptor (ER), progesterone receptor (PR), and HER2 [1, 2, 3, 4, 5, 6, 7]. Moreover, Sørlie *et al.* used hierarchical clustering on gene expression data that led to the identification of significant breast cancer subtypes namely $luminal\ A$, $luminal\ B$, $HER2$ (human epidermal growth factor receptor 2) and $basal-like$ [2]. The high cost of gene expression analysis for a large number of genes was a significant obstacle in applying this method. To overcome this issue, researchers reduced the gene list to relevant gene signatures for the breast cancer subtypes. Parker *et al.* [8] have presented biomarker genes that can efficiently separate molecular subtypes, which could be an excellent alternative to whole transcriptome micro-array analysis.

Diversity of gene expression data in the subtypes is an indicator for clinical prognosis of the patients, such as survival outcome [9]. Especially, *luminal A* subtype patients are found to have better prognosis while basal-like subtype patients have the poorest prognosis. Importantly, this molecular classification has successfully discovered subtypes of $ER+$ and/or $PR-$ breast cancer as *luminal A* and *luminal B*. [3, 4, 5, 6, 7, 9].

In some studies, the microarray-based breast cancer classification has been considered as the gold standard [10]. However, the critical limitation of the microarray-based method is its failure to classify tumors consistently to specific molecular subtypes [11, 12, 13]. The main reason for this failure is that gene expression is dynamic within a patient, and this may yield misleading results for classification. In contrast, somatic mutations can be used for stable subtypes detection. As all cancers lead somatic mutations and mutational heterogeneity broadly exists in tumor masses, the classification of cancers based on the mutation profile can be helpful for cancer diagnosis and treatment. On the other side, with the development of new sequencing technologies, genome sequencing has become an applicable tool for diagnostic purposes. Therefore, cancer classification based on gene mutation profiles and association of the classification into the clinical decisions can be a key point in the personalized medicine of cancer patients.

Some studies have merged different kinds of molecular data for breast cancer classification. Curtis *et al.* [14] developed a method to classify breast cancer via integrating genome and transcriptome data of 2000 breast cancer patients. Based on the impact of somatic copy-number alterations on the transcriptome, they introduced new subtypes of breast cancer. Furthermore, Ali *et al.* [5] classified breast cancer into ten subtypes based on the combination of Copy Number Alterations (CNA) and gene expression data. In another study, List *et al.*[6] proposed a computational method that merges gene expression and DNA methylation data to execute machine learning-based classification of breast cancer patients. In a novel recent study, Hofree *et al.* [7] proposed a network-based stratification algorithm to classify tumors via fusing somatic mutation profiles with gene interaction network and identified four subtypes of breast cancer. As somatic mutations are often sparse, it is sometimes challenging to predict cancer subtypes from somatic mutations. Therefore, previous studies used somatic mutation data along with other molecular information to classify cancers [7].

In most of the previous works, conventional clustering methods are used for clustering tumors, and novel, innovative clustering methods are not used.

Moreover, the number of clusters typically has been determined using the silhouette criterion, which sometimes leads to biologically meaningless clusters. In addition to the mentioned issues, the discovered clusters in previous works are not analyzed extensively. In this study, we proposed a novel method to classify breast cancer based on the integration of somatic mutation profiles and gene interaction network. We analyzed the somatic mutations and CNAs data from 861 breast tumors in The Cancer Genome Atlas (TCGA) database [15]. We used the network propagation method for smoothing somatic mutation profiles besides the gene interaction network and used deep embedded clustering [16] to find breast cancer subtypes. Moreover, for finding the best number of clusters, we used novel metrics such as AUMF [17] and MMR [18] and examined the biological associations of the subtypes that are discovered. Finally, we developed a supervised model to predict the subtypes of a new breast cancer patient. Also, the Random Forest (RF) used to find the most important genes for classification.

2 Material and Methods

2.1 Data Extraction and Smoothing

We used somatic mutation profiles collected by Zhang *et al.* [19]. They obtained somatic mutation data of 861 breast tumors from the TCGA. A gene recognized altered if at least one of the following conditions satisfies:

- It has a non-silent somatic mutation.
- It is a well-defined oncogene or tumor suppressor.
- It happens within a CNA.

The somatic mutation profiles are sparse; i.e., in each tumor, the number of genes that are mutated is relatively small compared to the total number of genes. In most machine learning techniques, sparse data cannot train the model well, so data needs to be smoothed. One of the most effective solutions for smoothing data is network propagation. By combining somatic mutation profiles and gene interaction network, we can obtain profiles that are not sparse. This study used protein-protein interaction (PPI) information in the STRING database [20] to create a gene interaction network. For this purpose, the $Homosapiense$ PPI network obtained from the STRING database. Then, the gene interaction network created from the PPI network. For each tumor, its mutation profile integrated with the gene interactions network. In fact, for each tumor, the entire vertex of the network is labeled, such that if a gene in the tumor has a mutation, the corresponding vertex for that gene is labeled one and zero otherwise.

Now the network propagation process applies a random walk with the following function over the networks:

$$D_{i+1} = \alpha D_i A + (1-\alpha) D_0 \qquad \text{i = 0, 1, 2, ...} \tag{1}$$

The adjustment parameter α sets the amount of distance that a mutation can be propagate on the network. The optimal value of α varies for each network (in this study it is subjectively set to 0.4). The network propagation operation iterates until D_{i+1} is converged (i.e. $||D_{i+1} - D_i|| < 1 \times 10^{-6}$), where D_0 is the original profile of tumor mutations, which is a $k * n$ matrix where k is number of tumors and n is number of genes, D_i is the modified profile of mutations in the ith iteration. Matrix A is a $n * n$ matrix that is computed by $A = H * D$, which $H = [h_{ij}]$ is the adjacent matrix of the network and $D = [d_{ij}]$ is a diagonal matrix such that:

$$d_{ij} = \begin{cases} \frac{1}{\sum_j h_{ij}} & \text{If } i = j \\ 0 & \text{Otherwise} \end{cases} \tag{2}$$

2.2 Clustering Method

After the propagating step, the mutation profile is a matrix with values between zero and one. In order to cluster this data, we used Deep Embedded Clustering (DEC) method. Suppose we have n tumors with the feature vectors $X_i = (x_{i1}, ..., x_{im})$ in space X with m dimension that should be clustered to k classes. Each cluster center is represented with $\mu_j, j = 1, \ldots, k$. Instead of clustering the data in the initial space X, the data are mapped to the latent feature space Z. The mapping is done by a nonlinear function $f_\theta : X \to Z$ which θ is a set of trainable parameters. Usually, in order to avoid the curse of dimensionality, the dimension of Z is less than m. The deep neural network can be used to implement f_θ, because of its theoretical function approximation characteristics [21], and the capabilities of learning features [22].

DEC is an iterative method, which learns feature maps and clusters using deep neural network simultaneously. In each iteration, the clustering representative $\{\mu_j \in Z\}_{j=1}^{k}$ as well as *parameters* θ of deep neural network are learned. This algorithm consists of two parts:

1 Initializing parameters using the stacked auto-encoder and centroids by k-means algorithm.

2 Parameter optimization that eventually leads to clustering. This section contains the continuous iteration of two-step: calculation of the auxiliary target distribution function, and minimization of the Kullback-Leibler divergence metric.

We tuned hyper-parameters of the model, and the best number of neurons in the stacked auto-encoder layers are 514, 500, 200, 500, and 514, respectively. There is also a layer with four neurons for clustering. See [16] for more details of DEC. The schema of the method is presented in Figure 1.

2.3 Finding the Best Number of Clusters

The clustering method needs the number of clusters (k) as input. Here, to select the best number of clusters, the clustering algorithm is implemented with different ks. There are some appropriate criteria to compare results and choose the best number of clusters. An approach to finding the number of clusters is to evaluate the classification based on microarray-based classes (PAM50) as the gold standard. For this purpose, a weighted bipartite graph G is formed, where the nodes of one part are the clusters of PAM50, represented by p_i symbols, and the nodes of another part are the clusters of the method, represented by c_j symbols. We weighted the edge (p_i, c_j) by the number of tumors shared between the cluster p_i and c_j. In Figure 2 you can see the general scheme of such a graph. After creating the graph, the following metrics are calculated in order to find best number of clusters:

$$PPV = \frac{\sum_{j=1}^{K} \max_i v_{ij}}{\sum_{i=1}^{L} \sum_{j=1}^{K} v_{ij}} \tag{3}$$

$$SN = \frac{\sum_{i=1}^{L} \max_j v_{ij}}{\sum_{i=1}^{L} l_i} \tag{4}$$

$$ACC = \sqrt{SN \times PPV} \tag{5}$$

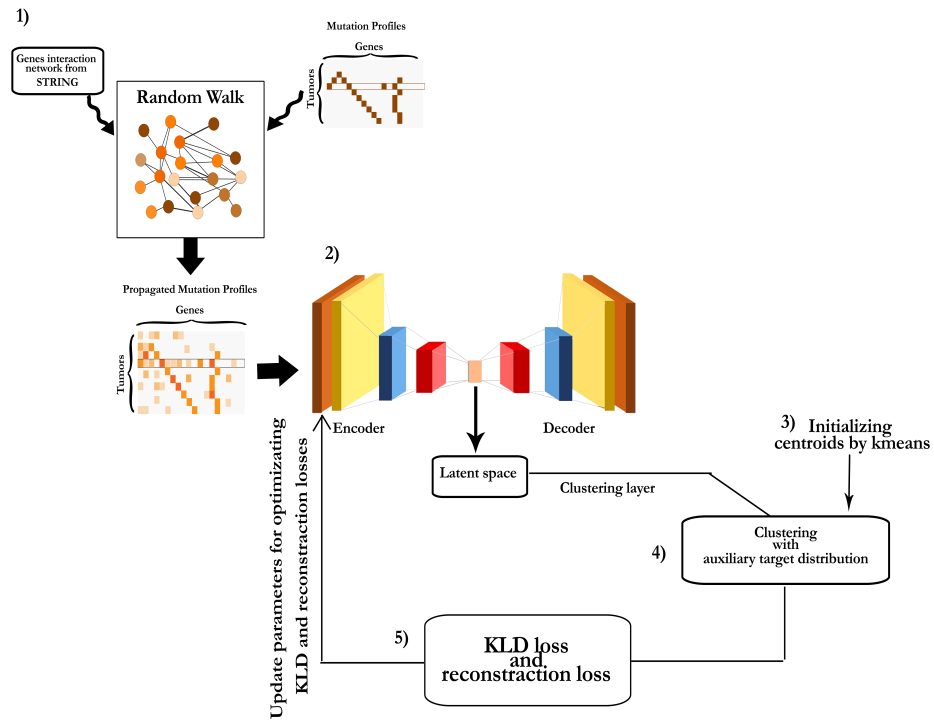

1)
Genes interaction network from STRING
Random Walk
Mutation Profiles
Genes
Tumors
Propagated Mutation Profiles
Genes
Tumors
2)
Encoder
Decoder
Latent space
Clustering layer
3) Initializing centroids by kmeans
4)
Clustering with auxiliary target distribution
5)
KLD loss and reconstruction loss
Update parameters for optimizating KLD and reconstruction losses

Figure 1 The schema of the proposed method

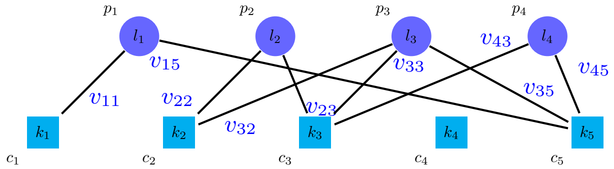


Figure 2 Bipartite graph between method to be evaluated and PAM50

These criteria have been introduced by Brohee and Holden [18]. In fact, ACC is the geometric mean of the two criteria PPV and SN. So ACC is a more comprehensive measure than PPV and SN.

Another important criterion is MMR [18]. To calculate this criterion, graph G is made and the weight on its edges is calculated based the threshold θ and the affinity score $NA(p_i, c_j)$, which represents the similarity of p_i and c_j.

$$v_{ij} = \begin{cases} NA(p_i, c_j) & NA(p_i, c_j) \geq \theta \\ 0 & (p_i, c_j) < \theta \end{cases} \quad (6)$$

$$NA(p_i, c_j) = \frac{|p_i \cap c_j|^2}{|p_i||c_j|} \quad (7)$$

MMR is calculated as follows:

$$MMR = \frac{\sum_{v_{ij} \in Match_w(\mathcal{P},\mathcal{C},\theta)} v_{ij}}{|\mathcal{P}|} \tag{8}$$

where $Match_w(\mathcal{P},\mathcal{C},\theta)$ is the maximum weighted matching of G.

The criteria discussed are qualitative criteria for comparison. Another approach of comparisons is the quantitative one. Suppose we have a graph similar to that made for computing MMR, and we have now ignored the weight of the edges. Let $Match(\mathcal{P},\mathcal{C},\theta)$ to be the maximum non-weighted matching of this graph:

$$N_p^+ = |\{p_i \mid \ \exists c_j, \ NA(p_i, c_j) \geq \theta, \ (p_i, c_j) \in Match(\mathcal{P},\mathcal{C},\theta)\}| \tag{9}$$

$$N_c^+ = |\{c_j \mid \ \exists p_i, \ NA(p_i, c_j) \geq \theta, \ (p_i, c_j) \in Match(\mathcal{P},\mathcal{C},\theta)\}| \tag{10}$$

$$Precision^+ = \frac{N_p^+}{|\mathcal{P}|} \tag{11}$$

$$Recall^+ = \frac{N_c^+}{|\mathcal{C}|} \tag{12}$$

$$F - measure^+ = \frac{2 \times Precision^+ \times Recall^+}{Precision^+ + Recall^+} \tag{13}$$

This set of criteria introduced by Maddi *et al.* [17], in which $F - measure^+$ is the harmonic mean of the two criteria $Precision^+$ and $Recall^+$. So $F - measure^+$ is a more comprehensive and meaningful measure than the $Precision^+$ and $Recall^+$ criteria. All criteria examined are in the $[0, 1]$ range.

One of the most comprehensive criteria in this area is the AUMF [17], which combines qualitative and quantitative attitudes. In fact, in this criterion the area under the curve $(MMR + Fmeasure^+, \theta)$ considered as a clustering measure called AUMF, which is in the $[0, 2]$ range.

We executed DEC on different number of clusters and results showed that best number of clusters is four (see additional file #1). Also, to evaluate the performance of the proposed clustering method, this method is compared with other popular and common clustering methods such as Hierarchical Clustering (HC), $k - means$ clustering, and Spectral Clustering (SPC). DEC can achieved better performance on comparison with other clustering methods.

2.4 Supervised Classification for new tumors

Using the discovered breast cancer subtypes, we labeled each tumor with its discovered label and proposed a supervised classifier to understand how accurate new breast tumors can be predicted based on their somatic mutations. With this model, we can predict the subtype of a new patient, using mutation profile as input.

We labeled each tumor with its assigned subtype and run five common machine learning methods: random forest, Support Vector Machine (SVM), Multi-layer Perceptron (MLP), Naïve Bayes (NB), and k-Nearest Neighbors(KNN) to classify the tumors into k subtypes $\{C_i\}_{i=1}^k$.

The 10-fold cross-validation is used for evaluation of different classifier performances. In 10-fold cross-validation, the whole set of tumors is randomly divided into ten subset with almost the same size. Then, one of the subsets is eliminated and the model is trained with the remaining nine subsets and evaluated with removed subset. This process is repeated such that each of ten subsets is considered as test data once. In this study, the 10-fold cross-validation is repeated 100 times and average performance of model is reported. The performance of the model is measured by standard evaluation criteria such as accuracy, sensitivity, precision, F-measure and AUC.

$$Accuracy = \frac{\sum_{i=1}^{k} \frac{TP_i+TN_i}{TP_i+TN_i+FP_i+FN_i}}{k} \tag{14}$$

$$Precision = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k}(TP_i + FP_i)} \tag{15}$$

$$Recall = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k}(TP_i + FN_i)} \tag{16}$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{17}$$

Where TP_i, TN_i, FP_i, and FN_i stand for the number of True Positives, True Negatives, False Positives, and False Negatives of class $\{C_i\}_{i=1}^{k}$. Since the values of accuracy, precision, recall, and F-measure is dependent to the value of the threshold, we also evaluate methods via AUC, which is the area under the receiver operating characteristic (ROC) curve. These criteria indicate the efficiency of methods independent of the threshold value.

3 Results

After Clustering tumors using DEC method, four clusters are obtained.

- Subtype 1: 182 tumors
- Subtype 2: 82 tumors
- Subtype 3: 499 tumors
- Subtype 4: 98 tumors

We did the following evaluations for investigation of the results and their correlations with PAM50 clusters.

3.1 Finding the Gene Signatures for Each Subtype

One of the efficient evaluations is finding influential genes in each subtype. This evaluation is important in two ways. First, it is possible to examine the biological significance of clustering method; second, these genes can be considered as candidates for the therapeutic purposes in each subtype patients. For this purpose, the Fisher exact test with Benjamini-Hochberg correction was used to find the gene signatures of each subtype. In the gene signature list, the top 50 genes with the p-value lower than 0.05 are calculated and shown in Additional Files (See Additional File #2). The gene interaction subnetwork of each subtype is obtained by enriching the gene signatures into STRING. Figure 3 illustrates the subnetworks of each subtype.

Many important genes are found in the gene signatures of the first subtype. For example, the $PIK3CA$ mutations occur in $20-30\%$ of breast cancer patients [23], which is a good diagnostic factor for hormone receptor-positive breast cancer patients [24]. Various treatments have been suggested for people with the $PI3KCA$

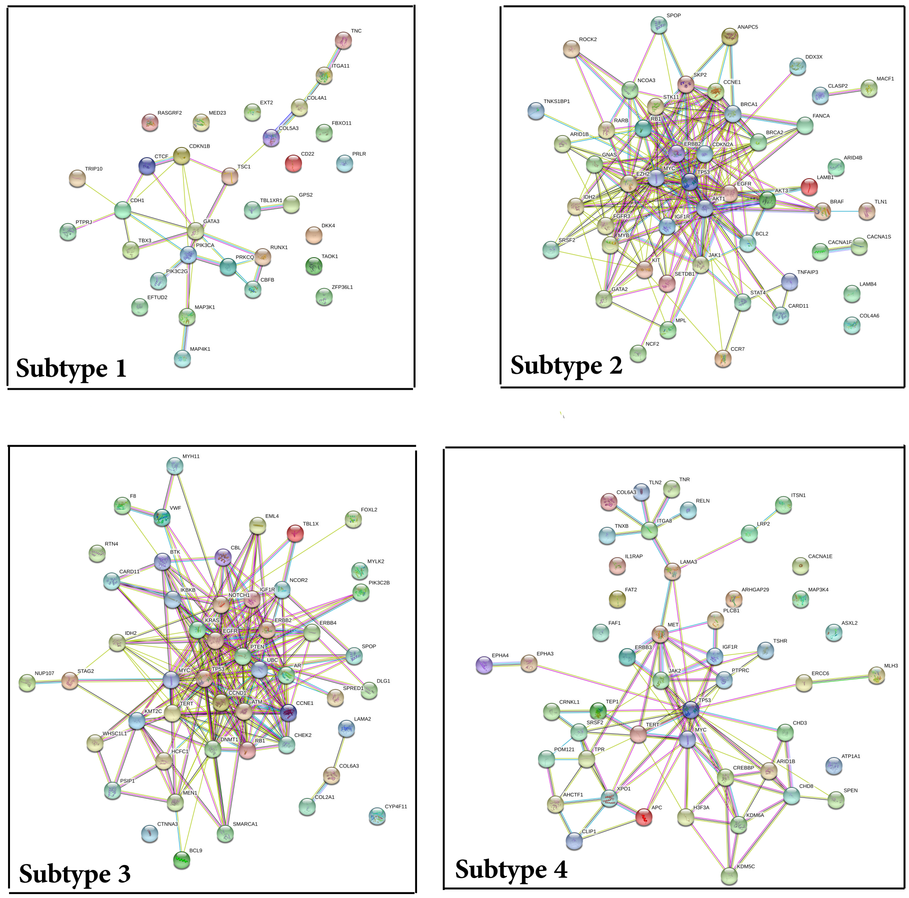


Figure 3 Subnetworks of all subtypes

mutation [25]. Another important gene in subtype1 is $CDH1$. Its mutation rate varies significantly across subtypes. This gene is highly expressed in the *luminal A* and *luminal B* subtypes, while it has low activity in the other two subtypes [26]. One of the genes that mutate with $PIK3CA$ is the $MAP3K1$ gene. In 11% breast tumors, both of these genes are mutate [27]. Extensive studies on genetic sequencing data suggest that the $MAP3K1$ mutations often occur in tumors of the *luminal A* subtype [23]. Moreover, this gene is one of the driver genes that is important in the diagnosis of breast cancer [23]. $CDKN1B$ is another gene involved in many cancers such as prostate cancer, small intestine cancer, and breast cancer. The $CDKN1B$ is one of the driver genes in the mentioned cancers [28].

Many important genes like $ERBB2$, $TP53$, MYC and $BRCA1$ are presented in the gene signatures of the subtype2. One of the driver genes in breast cancer is $ERBB2$, which is an indicator for tumor invasion [29]. Mutations and overexpression of this oncogene shows the tendency of a tumor mass to become an invasive subtype of breast cancer and is one of the predictors of poor prognosis. One of the critical regulators of cell growth, proliferation, metabolism, differentiation, and apoptosis is MYC. Mutations of this gene have many roles in the development and progression of breast cancer, activation of oncogenes, and inactivation of tumor suppressors. The MYC gene is highly expressed in the basal-like subtype of breast cancer, which is being targeted for treatment in these patients. The $BRCA1$ repressor gene inhibits the expression and activity of MYC. Mutations of the $BRCA1$ and MYC genes exacerbate breast cancer, especially basal-like subtype [30]. The $TP53$ gene also is mutated in about 20-40 % of breast cancer patients. It is useful to note that the mutation frequency of this gene is higher in patients with recurrent breast cancer [31]. Therefore, the second subtype is more invasive because its significant genes are mostly mutated in invasive cancers. The probability of poor prognosis and metastasis may be high in this subtype.

The third subtype contains many important genes, such as $Notch$, $CCND1$, and $IGF1R$. The $Notch$ family genes, including $Notch1, Notch2, Notch3$ and $Notch4$, are highly expressed in breast cancer patients. These genes play an important role in the differentiation, proliferation, and cell cycle [32]. The $Notch1$ gene also indicates aggressive breast cancer. About 80% of cancers have estrogen receptors and are treated with anti-estrogen drugs. One of the leading causes of death in such patients is their resistance to anti-estrogen drugs. Estrogen pathways have a positive association with anti-estrogen drug resistance in ER-positive breast cancers via suppressing $Notch1$ [33]. In ER positive breast cancers, the amount of $cyclinD1$ is high due to overexpression of the $CCND1$ gene and overexpression of the insulin-like growth factor receptor (IGF1R) [34].

Central genes of subtype4 have a high intersection with the central genes of the subtype2. Also, many essential genes are found among the gene signatures of this subtype. The MET gene is the tyrosine kinase receptor, which initiates the activity of its downstream pathways by binding to its ligand, the hepatocyte growth factor (HGF). It has different cellular activities in cell growth and cancer progression. Mutation of this gene often occurs in the basal-like subtype, where there is no estrogen receptor and $HER2$ [35].

3.2 Survival Analysis

We used cox hazard regression [36] for survival analysis in each subtypes. In order to investigate the significance of subtypes in the prediction of patient survival, chi2-test was used, which showed that subtypes are important features in cox hazard regression (p-value=0.00475). Cox hazard regression analysis showed that molecular subtypes have a significant correlation with the hazard rate. Figure 4 shows a diagram of cox hazard regression. It was mentioned in subsection 3.1 that $subtype2$ is invasive, due to the set of significant genes in this subtype. This issue is in consistent with survival Analysis. It can be seen that the second subtype has lower survival.

3.3 KEGG Enrichment

To find important pathways in each subtype, we enriched gene signatures of each subtype into KEGG database [37]. The results are described in the Additional Files (see Additional file #3).

One of the most important pathways in subtype1 is $PI3K/AKT/mTOR$, which promotes cell growth and tumor proliferation in breast cancer. This pathway has a significant role in resistance to Endocrine and Trastuzumab drugs in breast cancer

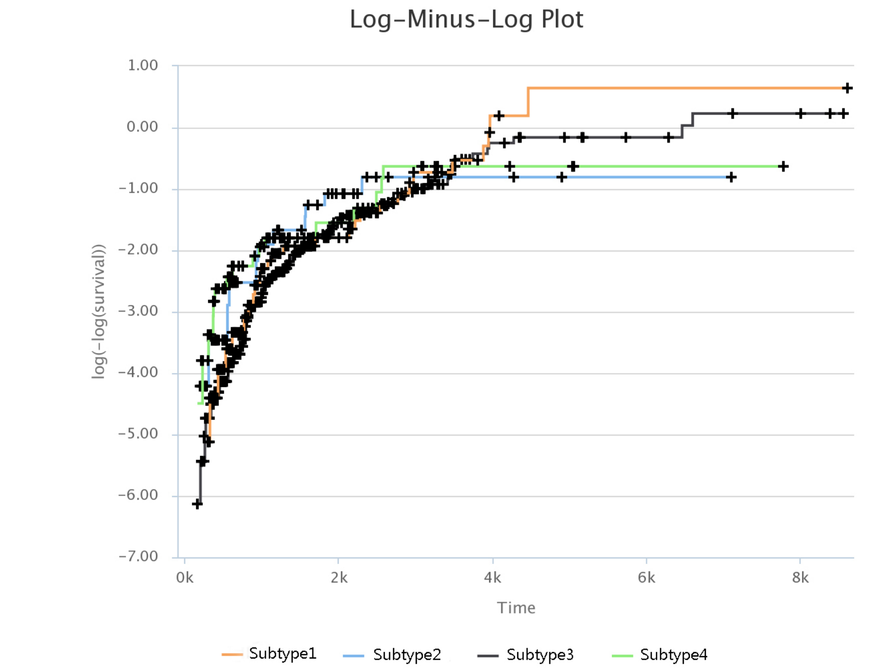


Figure 4 Cox Hazard Regression Survival Diagram

patients [38, 39]. Many studies have examined the relationship between the $MAPK$ pathway and the $PI3K/AKT$ pathway [40]. In fact, $PI3K$ suppresses the path of $MAPK$. This may be consistent with the $TCGA$ study that nonsense mutations and truncating $MAPK$ mutations are present simultaneously in breast cancer tumors [23]. Although $PIK3CA$ mutations often occur in $luminal\ A$ tumors [23]; the $PI3K/AKT$ messenger pathway is usually active in basal-like tumors [39]. The $TP53$ mutations and the $PI3K/AKT/mTOR$ pathway activity are found in breast tumors, especially the basal-like subtype [23].

3.4 Investigation of Protein Complexes of each Subtype

Since most of the cell activity is carried out by protein complexes, we have also investigated protein complexes in each subtype. The gene signatures of each subtype are entered to the iRefWeb website; then, the ordered list of complexes of each subtype is obtained [41]. More information on these complexes is available in the CORUM database [42]. The results are described in Additional Files (see Additional file #4). However, we discussed some results below.

One of the notable complexes in the first subtype is the $p27 - cyclinE - CDK2$ complex, which contains two $CDK2$ and $CDKN1B$ genes. This complex is involved in cell cycle regulation, cell cycle control, and DNA processing. One of the crucial regulators of the cell cycle is $CDKN1B$, which inhibits $G1/S$ via clinging to $CDK2$ and suppressing it. Over-expression of $CDKN1B$ gene in specific cancer cells in mice prevents DNA replication and tumorigenesis, while its deficiency plays an inhibitory role in human cancers and decreases the chance for developing breast, prostate, colon, lung, and esophagus [43].

BRCC complex includes the genes BRCA1, BRCA2, BRCC3, RAD51, and BRE, which is among the influential complexes in the second subtype. The function of the BRCA1 gene in DNA repair and cell cycle control in response to DNA damage is regulated by other complexes. Interaction of BRCA1 with RAD51 has a direct impact on the double-strand breaks of DNA [44]. Not only does ERCC complex has a direct interaction with TP53 in the destruction of DNA, but also it causes the displacement of DNA. Recently, the expression of two new members of the complex, namely BRCC36 and BRCC45 has been discovered in breast cancer cells [45].

The set of $TBL1X$, $HDAC3$, and $NCOR2$ genes together make the $SMRT$ complex, which plays a vital role in subtype3 tumors. The $SMRT$ complex is both an activator and a suppressor of the estrogen receptor-$\aleph$ ($ER-\aleph$), which its overexpression in breast cancer can make therapeutic outcomes more complicated. The activity of this complex inhibits the regulated cell death via the genes involved in apoptosis. This complex activates the anti-apoptotic genes and suppresses the pro-apoptotic genes. Thus, by activating multiple pathways, this complex leads to the progression and proliferation of breast cancer and declining the apoptosis [46].

$ESR1 - MDM4$ complex that is consisted of two genes $ESR1$ and $MDM4$ proteins is essential in the fourth subtype. The estrogen hormone receptor $ESR1$ is a nuclear hormone receptor that is expressed in approximately 70% of patients with breast cancer [47]. Expression of $MDM4$ gene is positively correlated with expression of $ER\alpha$ in primary breast tumors. Also, $ER\alpha$ enhances the expression of $MDM2$ [48].

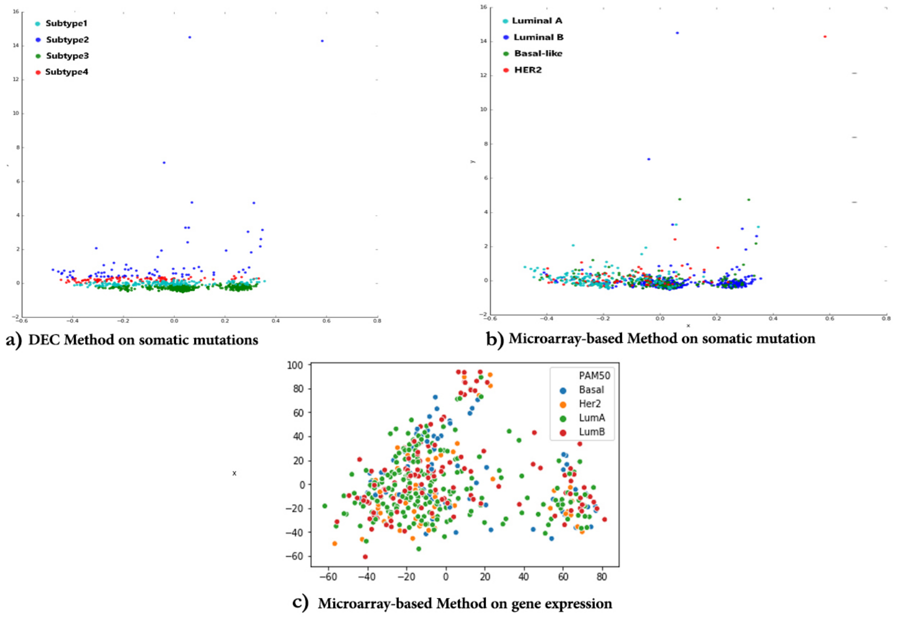
3.5 Clinical Examination

We investigated the relationship between each subtype and its clinical features such as ER status, PR status, $HER2$ status, TP53 status, PAM50 subtypes, and histopathological subtypes with chi2-test. The results are shown in Additional Files (see additional file #5). The discovered subtypes have a significant relation with the mentioned clinical features. The discovered $subtypes1$ and $subtype3$ mostly contain $luminal\ A$ and $luminal\ B$ tumors, while the majority of tumors in discovered $subtypes2$ and $subtypes4$ are $Her2-positive$ and $basal-like$.

In particular, $subtypes1$ and $subtype3$ are consisted of tumors that are $ER+$, $PR+$, have wild type $TP53$, and their most significant genes are $PI3KCA$, $CDH1$, and MYC. Moreover, $subtypes2$ and $subtypes4$ mostly contain tumors that are $PR-$, have mutant $P53$, and $TP53$, $ERBB2$, MYC are their significant genes.

It is noteworthy that although the majority of tumors in $subtypes1$ and $subtype3$ are $luminal\ A$ and $luminal\ B$, numerous $Her2-positive$ and $basal-like$ tumors are included in these two subtypes. A similar issue is shown in the case of $subtypes2$ and $subtypes4$. Thus, the discovered subtypes are not fully matched with PAM50 subtypes.

To compare the illustration of discovered and microarray-based subtypes in two-dimensional space, we used Principal Component Analysis (PCA) and reduced somatic mutation profiles of tumors to two-dimension. Figure 5.a shows that the discovered subtypes are linearly separable and reveals more discriminant features than PAM50 subtypes. Furthermore, figure 5.b shows the illustration of $PAM50$ clusters based on somatic mutation and figure 5.c shows the illustration of PAM50 clusters based on gene expression data does not demonstrate a high separability.



a) DEC Method on somatic mutations

b) Microarray-based Method on somatic mutation

c) Microarray-based Method on gene expression

Figure 5 Visualization of each Classification

3.6 Supervised Methods Evaluation

After clustering tumors, they are labeled according to the discovered subtypes. Then, five classifiers, namely, random forest, support vector machine, multi-layer neural network, k nearest neighbors, and naive Bayes, have been implemented to classify tumors based on the discovered subtypes. Ten-fold cross-validation was used to evaluate the performance of the classifiers. The evaluation criteria Accuracy, AUC, F-measure, Precision, and Recall are calculated for each method. According to additional file #7, naive Bayes has the worst performance. Support vector machines, k nearest neighbor, and multilayer perceptron have average performance. The best method in all criteria is the random forest with AUC 99%, the accuracy of 86%, the precision of 90%, recall of 85%, and F-measure of 87%, which has achieved great results. It can be concluded that the discovered subtypes by DEC method are separable; also, these subtypes can be predicted only by receiving mutations of important genes for new tumors. According to additional file #6, we enriched these important genes in GSEA, and surprisingly, many of them are the most important genes in cancer. Results are available in Additional File. Figure 6 shows the Roc curves of random forest classifier for each subtype. The value of AUC is excellent for each subtype and very close to one. However, the value of AUC for the third subtype is equal to one, which indicates that the model fits well on the tumors of the third subtype.

3.7 GSEA Enrichment

To find a family of genes that are related to cancer, we enriched gene signatures by Gene Set Enrichment Analysis (GSEA) tool [49]. We recognized many essential genes in transcription factor and protein kinase gene families that are well known to

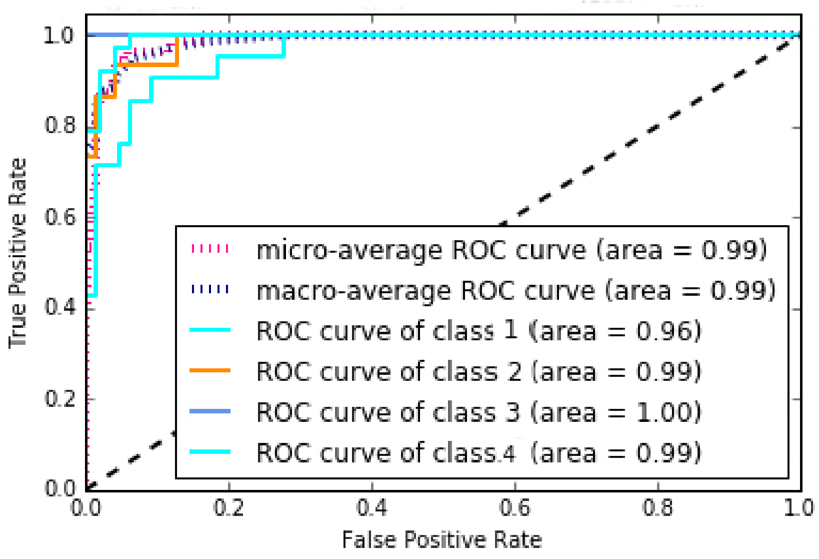


Figure 6 Area under curve plot of random forest

be associated with the progression of breast cancer. We found out that many gene signatures are related to cancer, and they are the essential genes in each subtype. The results are described in Additional Files (see additional file #7).

4 Conclusion

Cancer is a very heterogeneous disease; so, accurate classification of cancer is an important step to find the appropriate treatment. Recent advances in molecular biology have provided high quality and diverse data for the researchers. These data are included sequencing, transcriptomics, copy number variation, and methylation profiling. In this study, a novel cancer classification method was developed that identifies breast cancer subtypes using the profile of somatic mutations. The proposed method uses network propagation with deep embedded clustering and classified breast tumors into four subtypes. This method utilizes somatic mutation profile data of breast tumors in TCGA. DNA mutation data are more appropriate for identifying molecular cancer subtypes; however, gene expression is dynamic and variable in a time-dependent manner. Therefore, in this study, a deep clustering method based on somatic mutation data was used to classify breast tumors. Finding the gene signatures of each subtype can help better detection of the subtypes. These genes may also be targeted in the future to treat patients. In this study, the gene signatures of each subtype were detected via Fisher's exact test; then, the families of these genes were identified via $GSEA$ tool. Significant complexes, biological processes, molecular functions, cellular components, and pathways regulated via these genes have also been identified. Gene signatures are enriched in $KEGG$ to check the critical pathways of each subtype. Besides, the association of different clinical features with each subtype has been investigated. Finally, the random forest classification algorithm was used for supervised classification to provide predictions for new breast cancer patients, which could provide insights about the disease and its highly effective genes. The results of this study indicate that the subtypes of breast cancer can be clinically diagnosed using somatic mutation profiles. Besides, the proposed method can be used to predict the subtypes of new tumors. This study is not cancer-specific, and it can be used to classify any other cancers as well. For future research, we intend to address the following:

- To use the proposed method to detect subtypes of other cancers, such as brain cancer.
- To use other data such as gene expression, methylation, etc. Furthermore, we aim to examine the importance of each data in detecting cancer subtypes.
- To use cell-line samples along with tumors to study a larger dataset. The larger the dataset is, the more the accuracy can be improved.
- To use semi-supervised methods to classify cancer subtypes.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent to publish
Not applicable.

Funding
Not applicable.

Author's contributions
NR and CHE conceived the analysis. NR implemented the method and calculated the results. NR wrote the manuscript, and CHE helped to improve the paper. All authors have read and approved the final manuscript.

Acknowledgments
All authors thank Farzaneh Rami and Fatemeh Ahmadi Moughari for their helpful comments.

Availability of Data and Material
The data and material of the method are available at https://github.com/nrohani/MolecularSubtypes

Abbreviations
DEC: Deep Embedded Clustering; IHC: Immunohistochemical; ER: Estrogen Receptor; PR: Progesterone Receptor; CNA: Copy Number Alteration; TCGA: The Cancer Genome Atlas; RF: Random Forest; AUMF: Area Under MMR and F-measure; MMR: Maximal Marginal Relevance; PPI: PPI: Protein-Protein Interaction; ACC: Accuracy; PPV: Positive Predictive Value; SN: Sensitivity; HC: Hierarchical Clustering; KM: k-means Clustering; SPC: Spectral Clustering; SVM: Support Vector Machine; MLP: Multi-Layer Perceptron; NB: Naive Bayes; KNN: k-Nearest Neighbors; ROC: Receiver Operating Characteristic; PCA: Principal Component Analysis; AUC: Area Under Receiver Operating Characteristic Curve; AUPR: Area Under Precision-Recall Curve; FN: False-negative; FP: False positive; TN: true negative; TP: True Positive; GSEA: Gene Set Enrichment Analysis.

Author details
[1]Department of Mathematics, Shahid-Beheshti University, GC, Tehran, Iran. [2]Department of Biological Sciences, Institute for Research in Fundamental Sciences, GC, Tehran, Iran.

References

1. Elston, C.W.: Pathological prognostic factors in breast cancer. Crit Rev Oncol Hematol **31**, 209–223 (1999)
2. Perou, C.M., Sørlie, T., Eisen, M.B., Van De Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., *et al.*: Molecular portraits of human breast tumours. nature **406**(6797), 747 (2000)
3. Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., Van De Rijn, M., Jeffrey, S.S., *et al.*: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences **98**(19), 10869–10874 (2001)
4. Hu, Z., Fan, C., Oh, D.S., Marron, J., He, X., Qaqish, B.F., Livasy, C., Carey, L.A., Reynolds, E., Dressler, L., *et al.*: The molecular portraits of breast tumors are conserved across microarray platforms. BMC genomics **7**(1), 96 (2006)
5. Ali, H.R., Rueda, O.M., Chin, S.-F., Curtis, C., Dunning, M.J., Aparicio, S.A., Caldas, C.: Genome-driven integrated classification of breast cancer validated in over 7,500 samples. Genome biology **15**(8), 431 (2014)
6. List, M., Hauschild, A.-C., Tan, Q., Kruse, T.A., Baumbach, J., Batra, R.: Classification of breast cancer subtypes by combining gene expression and dna methylation data. Journal of integrative bioinformatics **11**(2), 1–14 (2014)

7. Hofree, M., Shen, J.P., Carter, H., Gross, A., Ideker, T.: Network-based stratification of tumor mutations. Nature methods **10**(11), 1108 (2013)
8. Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., *et al.*: Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology **27**(8), 1160 (2009)
9. Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., *et al.*: Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the national academy of sciences **100**(14), 8418–8423 (2003)
10. Peppercorn, J., Perou, C.M., Carey, L.A.: Molecular subtypes in breast cancer evaluation and management: divide and conquer, 125–142 (2007)
11. Gusterson, B.: Do'basal-like'breast cancers really exist? Nature Reviews Cancer **9**(2), 128 (2009)
12. Pusztai, L., Mazouni, C., Anderson, K., Wu, Y., Symmans, W.F.: Molecular classification of breast cancer: limitations and potential. The Oncologist **11**(8), 868–877 (2006)
13. Weigelt, B., Baehner, F.L., Reis-Filho, J.S.: The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland **220**(2), 263–280 (2010)
14. Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., *et al.*: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature **486**(7403), 346 (2012)
15. Consortium, I.C.G., *et al.*: International network of cancer genome projects. Nature **464**(7291), 993 (2010)
16. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis, 478–487 (2016)
17. Maddi, A.M., Moughari, F.A., Balouchi, M.M., Eslahchi, C.: Cdap: An online package for evaluation of complex detection methods. Scientific Reports **9**(1), 1–13 (2019)
18. Brohee, S., Van Helden, J.: Evaluation of clustering algorithms for protein-protein interaction networks. BMC bioinformatics **7**(1), 488 (2006)
19. Zhang, W., Ma, J., Ideker, T.: Classifying tumors by supervised network propagation. Bioinformatics **34**(13), 484–493 (2018)
20. Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al.: The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research, 937 (2016)
21. Hornik, K.: Approximation capabilities of multilayer feedforward networks. Neural networks **4**(2), 251–257 (1991)
22. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
23. Network, C.G.A., *et al.*: Comprehensive molecular portraits of human breast tumours. Nature **490**(7418), 61 (2012)
24. Ellis, M.J., Lin, L., Crowder, R., Tao, Y., Hoog, J., Snider, J., Davies, S., DeSchryver, K., Evans, D.B., Steinseifer, J., *et al.*: Phosphatidyl-inositol-3-kinase alpha catalytic subunit mutation and response to neoadjuvant endocrine therapy for estrogen receptor positive breast cancer. Breast cancer research and treatment **119**(2), 379 (2010)
25. Ramirez-Ardila, D.E., Helmijr, J.C., Look, M.P., Lurkin, I., Ruigrok-Ritstier, K., Van Laere, S., Dirix, L., Sweep, F.C., Span, P.N., Linn, S.C., *et al.*: Hotspot mutations in pik3ca associate with first-line treatment outcome for aromatase inhibitors but not for tamoxifen. Breast cancer research and treatment **139**(1), 39–49 (2013)
26. Zaha, D.C., Jurca, C.M., Bungau, S., Cioca, G., Popa, A., Sava, C., Endres, L., Vesa, C.M.: Luminal versus non-luminal breast cancer cdh1 immunohistochemical expression. REVISTA DE CHIMIE **70**(2), 465–469 (2019)
27. Avivar-Valderas, A., McEwen, R., Taheri-Ghahfarokhi, A., Carnevalli, L.S., Hardaker, E.L., Maresca, M., Hudson, K., Harrington, E.A., Cruzalegui, F.: Functional significance of co-occurring mutations in pik3ca and map3k1 in breast cancer. Oncotarget **9**(30), 21444 (2018)
28. Cusan, M., Mungo, G., De Marco Zompit, M., Segatto, I., Belletti, B., Baldassarre, G.: Landscape of cdkn1b mutations in luminal breast cancer and other hormone-driven human tumors. Frontiers in endocrinology **9**, 393 (2018)
29. Revillion, F., Bonneterre, J., Peyrat, J.: Erbb2 oncogene in human breast cancer and its clinical significance. European Journal of Cancer **34**(6), 791–808 (1998)
30. Xu, J., Chen, Y., Olopade, O.I.: Myc and breast cancer. Genes & cancer **1**(6), 629–640 (2010)
31. Norberg, T., Klaar, S., Lindqvist, L., Lindahl, T., Ahlgren, J., Bergh, J.: Enzymatic mutation detection method evaluated for detection of p53 mutations in cdna from breast cancers. Clinical chemistry **47**(5), 821–828 (2001)
32. Wang, J., Fu, L., Gu, F., Ma, Y.: Notch1 is involved in migration and invasion of human breast cancer cells. Oncology reports **26**(5), 1295–1303 (2011)
33. Hao, L., Rizzo, P., Osipo, C., Pannuti, A., Wyatt, D., Cheung, L.W., Sonenshein, G., Osborne, B., Miele, L.: Notch-1 activates estrogen receptor-α-dependent transcription via ikkα in breast cancer cells. Oncogene **29**(2), 201 (2010)
34. Tian, X., Aruva, M.R., Qin, W., Zhu, W., Duffy, K.T., Sauter, E.R., Thakur, M.L., Wickstrom, E.: External imaging of ccnd1 cancer gene activity in experimental human breast cancer xenografts with 99mtc-peptide-peptide nucleic acid-peptide chimeras. Journal of Nuclear Medicine **45**(12), 2070–2082 (2004)
35. Ho-Yen, C.M., Jones, J.L., Kermorgant, S.: The clinical and functional significance of c-met in breast cancer: a review. Breast Cancer Research **17**(1), 52 (2015)
36. Fox, J.: Cox proportional-hazards regression for survival data. An R and S-PLUS companion to applied regression **2002** (2002)
37. Aoki, K.F., Kanehisa, M.: Using the kegg database resource. Current protocols in bioinformatics **11**(1), 1–12 (2005)
38. Nahta, R.: Pharmacological strategies to overcome her2 cross-talk and trastuzumab resistance. Current

medicinal chemistry **19**(7), 1065–1075 (2012)

39. Ramirez-Ardila, D., Timmermans, A.M., Helmijr, J.A., Martens, J.W., Berns, E.M., Jansen, M.P.: Increased mapk1/3 phosphorylation in luminal breast cancer related with pik3ca hotspot mutations and prognosis. Translational oncology **10**(5), 854–866 (2017)
40. Laprise, P., Langlois, M.-J., Boucher, M.-J., Jobin, C., Rivard, N.: Down-regulation of mek/erk signaling by e-cadherin-dependent pi3k/akt pathway in differentiating intestinal epithelial cells. Journal of cellular physiology **199**(1), 32–39 (2004)
41. Turner, B., Razick, S., Turinsky, A.L., Vlasblom, J., Crowdy, E.K., Cho, E., Morrison, K., Donaldson, I.M., Wodak, S.J.: irefweb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database **2010** (2010)
42. Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., Mewes, H.-W.: Corum: the comprehensive resource of mammalian protein complexes—2009. Nucleic acids research **38**(suppl_1), 497–501 (2009)
43. Xu, S., Abbasian, M., Patel, P., Jensen-Pergakes, K., Lombardo, C.R., Cathers, B.E., Xie, W., Mercurio, F., Pagano, M., Giegel, D., *et al.*: Substrate recognition and ubiquitination of scfskp2/cks1 ubiquitin-protein isopeptide ligase. Journal of Biological Chemistry **282**(21), 15462–15470 (2007)
44. Christou, C., Kyriacou, K.: Brca1 and its network of interacting partners. Biology **2**(1), 40–63 (2013)
45. Dong, Y., Hakimi, M.-A., Chen, X., Kumaraswamy, E., Cooch, N.S., Godwin, A.K., Shiekhattar, R.: Regulation of brcc, a holoenzyme complex containing brca1 and brca2, by a signalosome-like subunit and its role in dna repair. Molecular cell **12**(5), 1087–1099 (2003)
46. Blackmore, J.K., Karmakar, S., Gu, G., Chaubal, V., Wang, L., Li, W., Smith, C.L.: The smrt coregulator enhances growth of estrogen receptor-α-positive breast cancer cells by promotion of cell cycle progression and inhibition of apoptosis. Endocrinology **155**(9), 3251–3261 (2014)
47. Stanford, J.L., Szklo, M., Brinton, L.A.: Estrogen receptors and breast cancer. Epidemiologic reviews **8**, 42–59 (1986)
48. Baunoch, D., Watkins, L., Tewari, A., Reece, M., Adams, L., Stack, R., Brown, A., Jones, L., Christian, D., Latif, N., *et al.*: Mdm2 overexpression in benign and malignant lesions of the human breast. International journal of oncology **8**(5), 895–899 (1996)
49. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences **102**(43), 15545–15550 (2005)

Figures

Tables

Additional Files

Additional File 1 — Additional.pdf

Figures

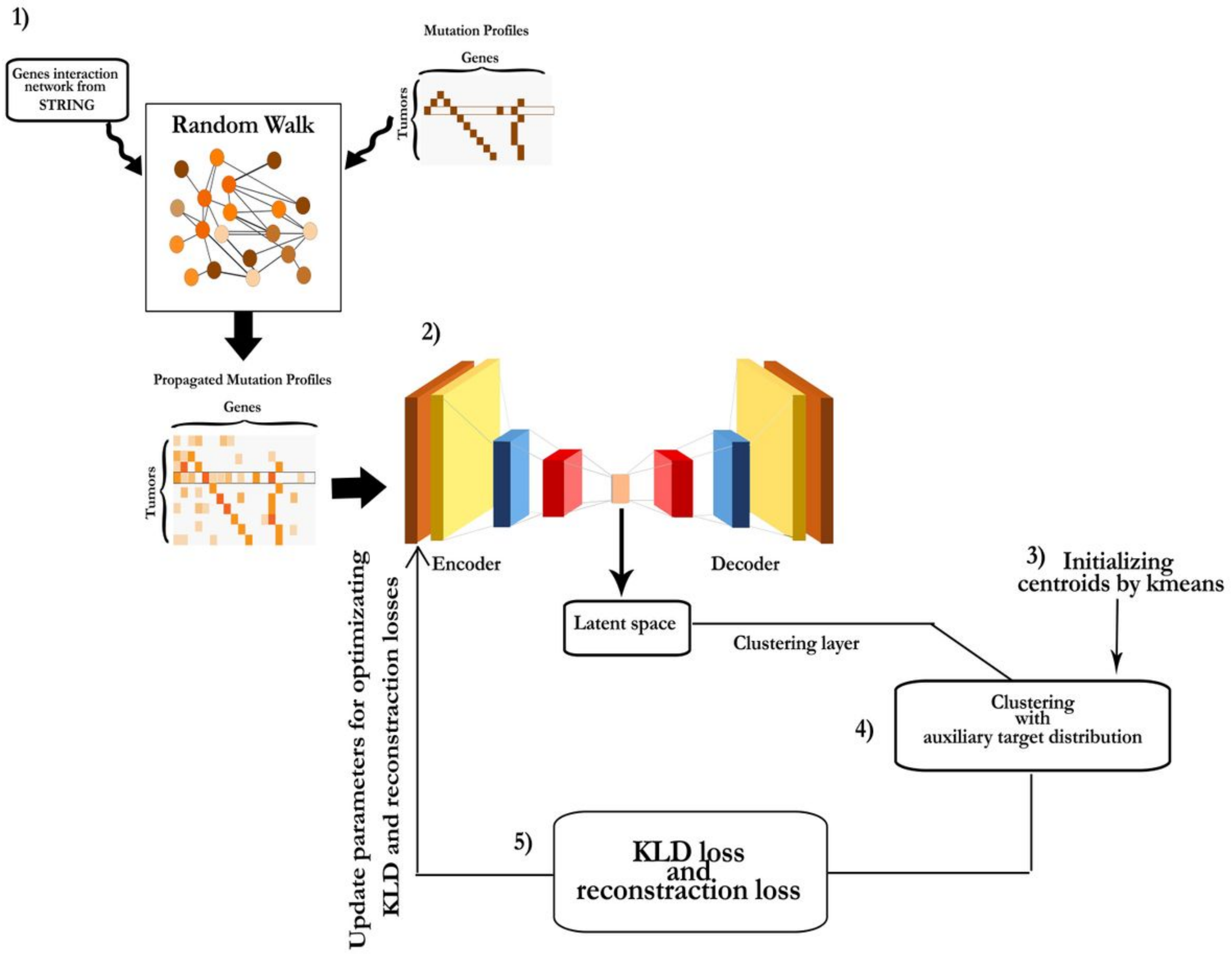

1)
Mutation Profiles
Genes
Genes interaction network from STRING
Tumors
Random Walk
Propagated Mutation Profiles
Genes
Tumors
2)
Encoder
Decoder
3) Initializing centroids by kmeans
Latent space
Clustering layer
4)
Clustering with auxiliary target distribution
5)
KLD loss and reconstraction loss
Update parameters for optimizating KLD and reconstraction losses

Figure 1

The schema of the proposed method

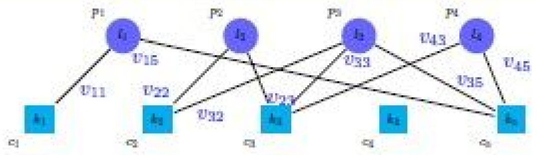


Figure 2 Bipartite graph between method to be evaluated and PAM50

Figure 2

Bipartite graph between method to be evaluated and PAM50

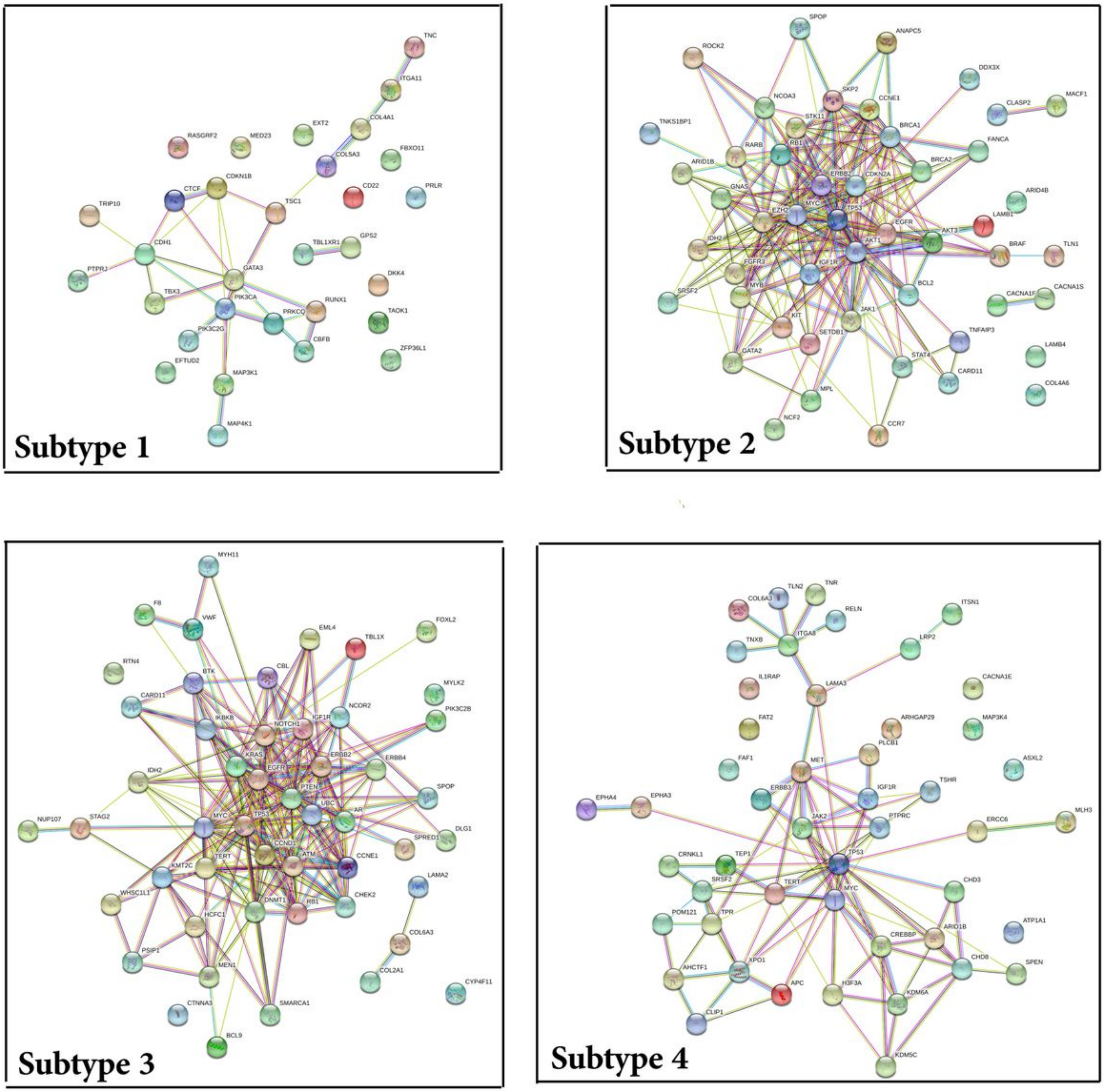


Figure 3

Subnetworks of all subtypes

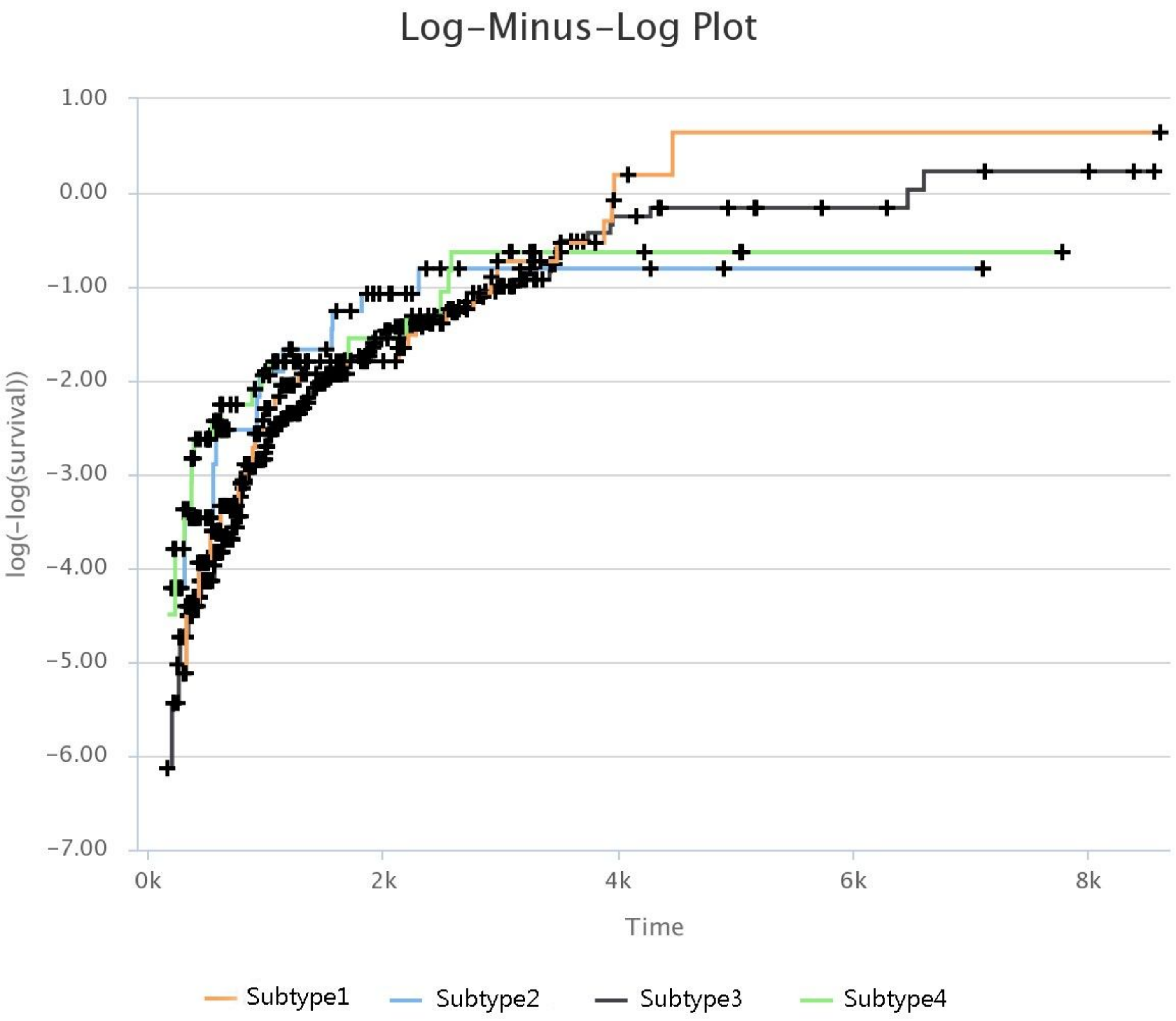


Figure 4

Cox Hazard Regression Survival Diagram

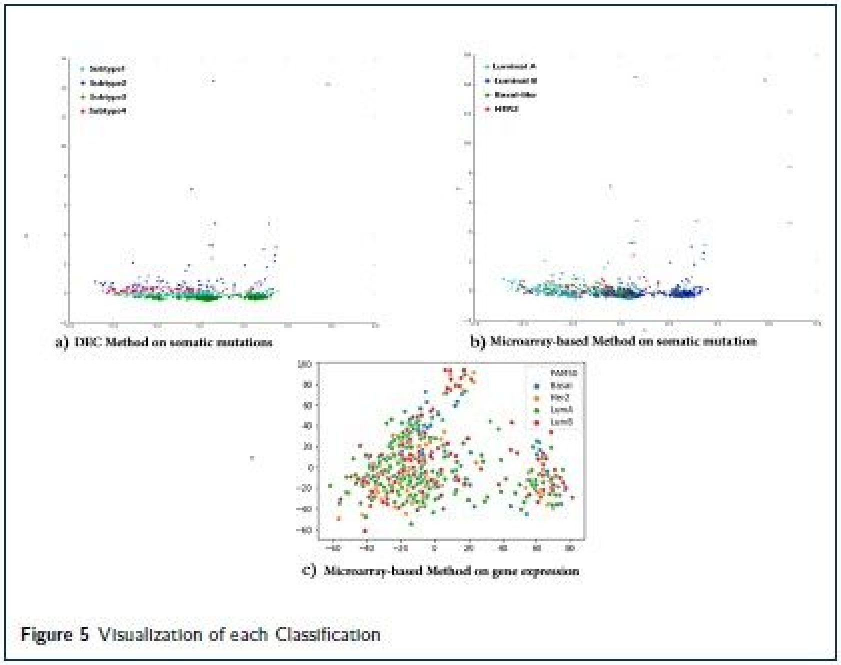


Figure 5

Visualization of each Classication

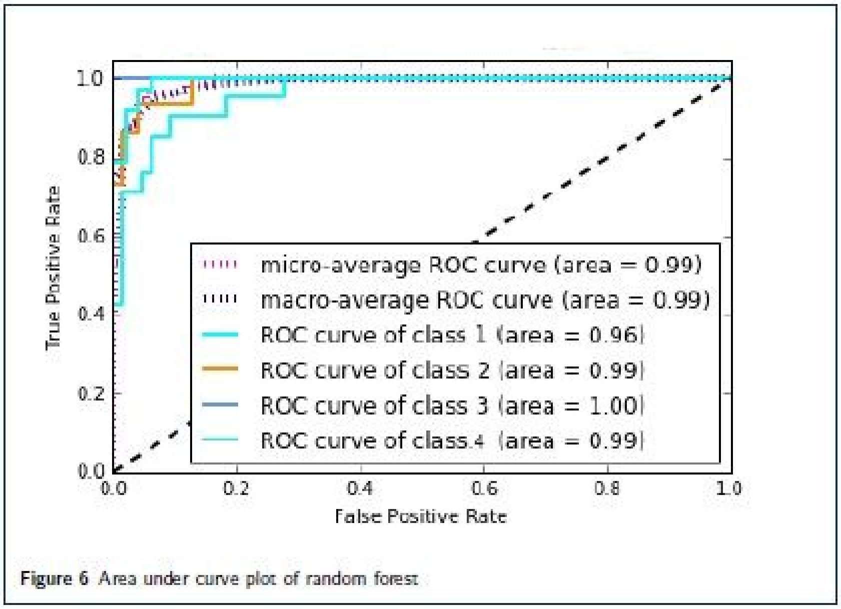


Figure 6

Area under curve plot of random forest

Supplementary Files