Mathematical Framework for Breast Cancer Stratification Using Somatic Mutation-Based Network Propagation

Abstract

Summary: According to the World Health Organization, breast cancer is responsible for 685,000 deaths worldwide in 2020. Because cancer is heterogeneous, treatment selection is highly customized and based on stratification. Thus, breast cancer stratification is of critical clinical utility. As pointed out in the previous research, certain inherited “high penetrance” gene mutations greatly increase breast cancer risk. In this passage, we set a mathematical framework for somatic mutation-based network propagation breast cancer stratification. In *Section 1*, we give an overview of the previous research around the data extraction of breast cancer, i.e., assay methods, and former mathematical approaches to data clustering. In *Section 2*, we describe the network propagation model from a random walk process. Finally, we discuss the current prognosis analysis framework in *Section 3*.

Keywords: Breast Cancer, Somatic Mutation, Gene-Protein Network, Network Propagation, Cox Regression, Random Walk

Contents

1. Breast Cancer Stratification

1.1 Overview

Breast cancer stratification has been developed for a prolonged period and is still a topic under debate due to the heterogeneous nature of cancer.[1,2] The stratification of breast cancer is vital for treatment method selection as it provides a reference for treatment selection. For instance, adjuvant chemotherapy is only effective for roughly 15% of the patient population having early-stage breast cancer, this ratio could be alleviated by introducing a more specified stratification.[3] The hormonal treatment for metastasis risk reduction is dependent on the expression level.[4] Moreover, a more specified stratification method is able to contribute to a better understanding of cancer diagnosis and treatment.[5,6]

Breast cancer stratification is based on the morphology in the first place, including patient age, axillary lymph node status, tumor size, histological features, hormone receptor status, and HER-2 status.[7] As the detection methods advance, more information can be derived on a molecular level. Based on the molecule subtypes, breast cancer can be further classified by the expression of three key receptors, i.e., estrogen receptor, progesterone receptors, and HER-2.[8] However, patients with similar clinical and pathological features vary greatly in therapy outcomes.[9] In the first decade of the 21st century, a lot of attention had gathered around microarray profiling because of the emergence of DNA microarray technology.[10,11] The two main types are cDNA microarrays and oligonucleotide microarrays.[12,13] From then on, much progrhavehasn made. Currently, there are multiple commercial microarray tests for different clinical decision-making such as Oncotype DX, MammaPrint, PAM50-risk of recurrence score, Breast Cancer Index, and EndoPredict.[14–18] A thorough review of the statistical methods and current developments of the microarray can be found in the citation here.[19] However, the drawback of gene profiling based on microarrays is the instability of the stratification result.[1,20,21] Relatively speaking, the stratification based on somatic mutation is more stable and is becoming widely recognized nowadays.[22]

Currently, a huge number of clinical guidelines are based on single gene mutation.[23] As reported by a lot of research, somatic mutations in the PIK3CA gene are strongly correlated with the expression of the estrogen receptor.[24–26] By integrating the breast cancer somatic mutation data with the gene-protein network, a more detailed stratification of breast cancer can be derived. We give a glimpse of the current methods applied in breast cancer stratification in the next section, i.e., clustering methods, and provide a mathematical framework specifically for breast cancer stratification using somatic mutation-based network propagation.

1.2 Stratification Approaches

An enormous amount of effort has been made searching for a one-size-fits-all stratification method for breast cancer. However, there is no such thing as a perfect stratification method, as is pointed out by Jane et al.[27] By categorizing gene-related data into different strata, this problem can be treated as a data clustering problem. We introduce the core ideas of the commonly used clustering methods along with the introduction of some state-of-the-art methods to provide the readers with a broader view of gene profiling.

Multiple summary articles have given detailed categorization of clustering methods, however, none of them are specifically designed for gene profiling.[28–33] In gene profiling, there are approximately 5 major classical clustering methods, and we discuss them respectively, including (a). Hierarchical clustering/clustering based on connection, (b). Partitional clustering, (c). Clustering based on density, (d). Clustering based on finite elements, (e). Clustering based on distribution.

Hierarchical clustering is the earliest applied method in gene profiling. In 2000, Alzadeh et al. discovered three subtypes of diffuse large B-cell lymphoma using hierarchical clustering.[28] Therese et al. classified breast carcinomas based on the variations in gene expressions using hierarchical clustering in 2001.[34] It can be done either through an agglomerative process or a divisive process.[30] The drawback of hierarchical clustering includes the rejection of missing data, difficulty in finding the cluster quantity, and the inability to handle a massive amount of data.

Partitional clustering can be divided into two steps: selecting the initial clusters and using an iterative algorithm to update the clusters. Of which k-means and k-medoid are applied the most. Inderjit et al. derived the weighted kernel k-means and spectral clustering to overcome the subjective shortcoming of the partitional clustering methods.[35] In 2021, Sehhati et al. applied the combination method of k-means and random forest to predict recurrence in biological networks made from gene expression data.[36]

Clustering based on density is able to perform clustering on massive datasets through a two-step process. By constructing a map of the relevant portion of the data space, the calculation for density function is promoted.[28] Of all the density-based clustering methods, DBSCAN, DENCLUE, and OPTIC are favored for their low algorithm complexity.[37–39] Joshua et al. designed the PSCAN system, which is a parallel implementation of DBSCAN, for robust protein clustering.[40] Edla et al. combined k-means with DBSCAN to cluster the gene expression data.[41]

Clustering based on finite elements owns the advantage of fast calculation. In regards to computational complexity, STING and OptiGrid are the optimal choices[42,43]. However, adopting these methods requires the initial setting of the threshold value of the size of the grid. To overcome this shortcoming, adaptive grids are recommended for optimization.

Clustering based on distribution requires prerequisites, i.e., the distribution of each variable. In practice, the distributions are usually yet to be elucidated. Thus, this clustering method has a relatively narrow application in gene profiling.

In recent years, more novel clustering methods are coming into sight due to the upgrade in computing ability. For instance, the minimum spanning tree, subspace clustering, mutual information clustering, and multi-clustering. Liu et al. proposed a multiple graph spectral clustering method integrated with graph association to contribute to the discovery of functional molecules, which in turn determine their association among cancer diseases.[43] Yongfu et al. proposed a deep embedded refined clustering method for breast cancer differentiation based on DNA methylation to study the differentiation of breast cancer.[44] This method is based on the dimensionality reduction of the methylation data and the clustering algorithm based on the soft assignment of the latent space provided by the autoencoder. Zhaopeng et al. proposed a single-cell self-paced clustering based on nonnegative matrix factorization using the F norm.[45] An illustration of the relationship among the clustering methods we discussed is

shown in **Figure 1**.

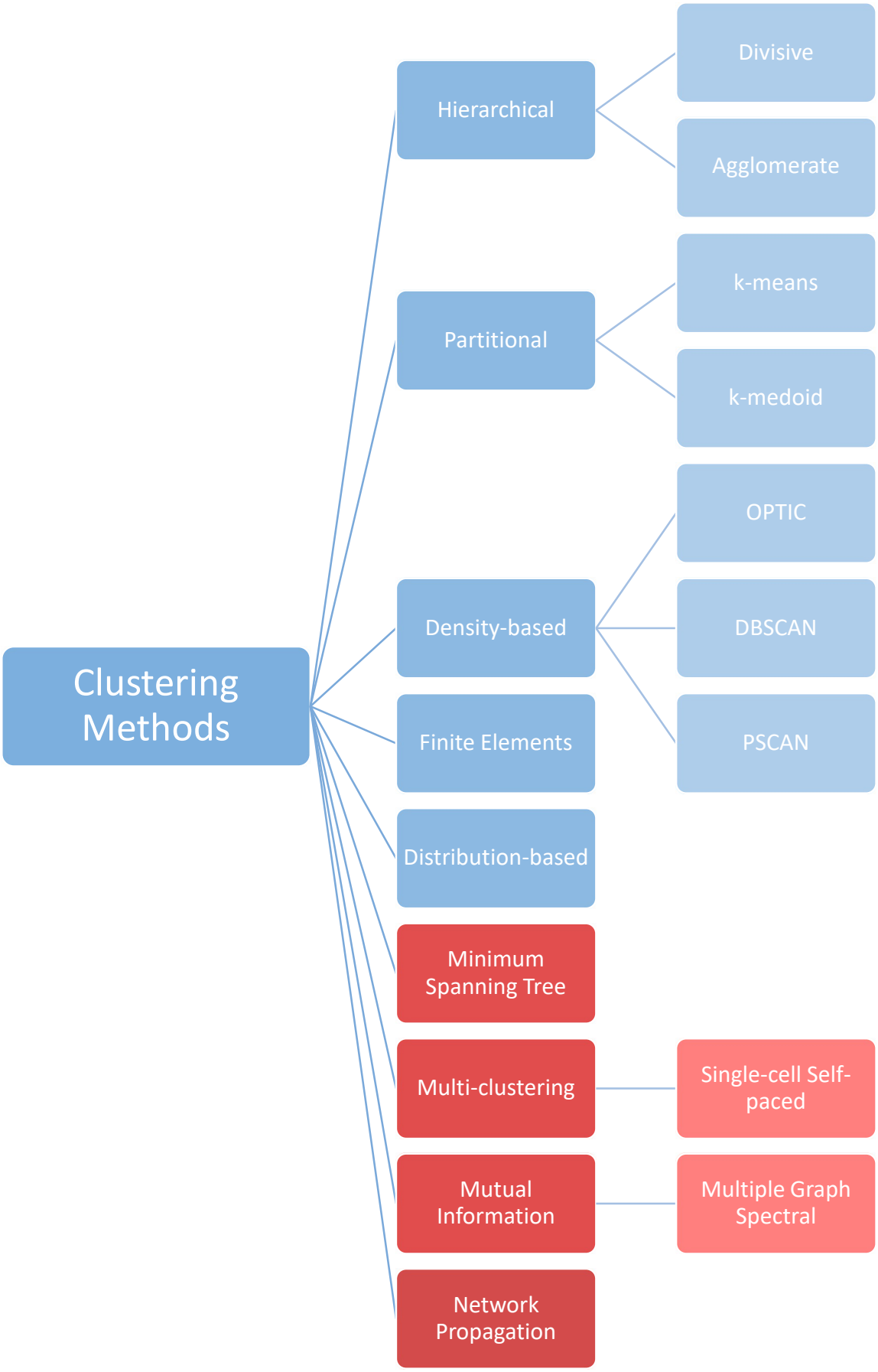

Figure 1

Figure 1. The columns in blue indicate the traditional clustering methods and the columns in red indicate the novel clustering methods.

2. Stratification Based on Somatic Mutation

In the following subsections, we first give some available data sources in *Section 2.1*. In *Section 2.2*, we give the mathematical layout of the network-based stratification in a logical order. In *Section 2.3*, we discuss the applicability of machine learning techniques in network propagation.

2.1 Data Sources

There are multiple relevant research focusing on breast cancer stratification based on somatic mutation profiles. The major data sources are The Cancer Genome Atlas (TCGA) project, the COSMIC database, the STRING database, KEGG database, the Genecards, and the GEO database.[46–51] The Cancer Genome Atlas was brought up by the National Human Genome Research Institute in 2006, more than 20000 types of primary cancers are molecularly characterized with corresponding normal samples.[46] The Catalogue Of Somatic Mutations In Cancer provides the largest and widest database for somatic mutation influence on human cancer.[48] STRING is a database providing the interactions among proteins.[49] Gene Expression Omnibus is a public functional genomics database that provides users with curated gene expression profiles and corresponding experiments.[47] KEGG is a knowledge database for systematic analysis of gene functions, linking genomic information with higher-order functional information.[51] GeneCards is a database of human genomes, transcriptomes, and proteomes.[50]

Of note, one should consider eliminating the effect of other irrelevant factors such as cigarette smoking, drinking, race, and ethnicity based on breast cancer epidemiology surveys.[52–54]

2.2 Network-Based Stratification

2.2.1 Gene-Protein Network

The gene-protein network is composed of the protein-protein network and the gene-gene network. Their relationship is hierarchical since the translation of protein is dependent on the genes' expression status. The interactions among proteins constitute a vital part of cell physiological activities. Proteins are the main structural elements of biological tissues as well as the carriers for most of the cellular biochemical reactions, for instance, large protein complexes contribute to the fundamental processes in cell mechanisms including DNA transcription, DNA replication, and so on.[55] The introduction of the protein-protein network is to illustrate the interactions among key molecules and provide guidelines for the design of the targeted therapeutic agents accordingly.[56–58] By complementing the gene-gene network, more valuable biomarkers could be identified.[59]

Currently, one of the most difficult challenges in the past genome era is the identification of the protein function, and a widely recognized solution is through the network.[60] By recognizing the protein types as nodes and relationships as edges, one can use the graph theory for further analysis, which is indicated by $G =< V, E >$ in the next section. Many biological features can be represented intuitively by the graphs. Jeong et al. found that the network topological structure is merely affected by random mutations in their research on Saccharomyces cerevisiae.[61] Han et al. proved the distorting effect of sampling on the apparent topology of scale-free networks by yeast two-hybrid technique.[61]

There are multiple ways to visualize the network, one way is the integrated circuit with signaling molecules affecting cancer cells as nodes and metabolic pathways as edges, and according to Douglas et al., the circuit can be segmented into subcircuits responsible for diverse functions of the cell.[62] After a thorough review of breast cancer metabolic pathways, we give a visualized demonstration of triple-negative breast cancer metabolic pathways shown in **Figure 2**. Besides, we retrieved a figure of the gene-gene interaction network from the Genecards, which is shown in **Figure 3**.

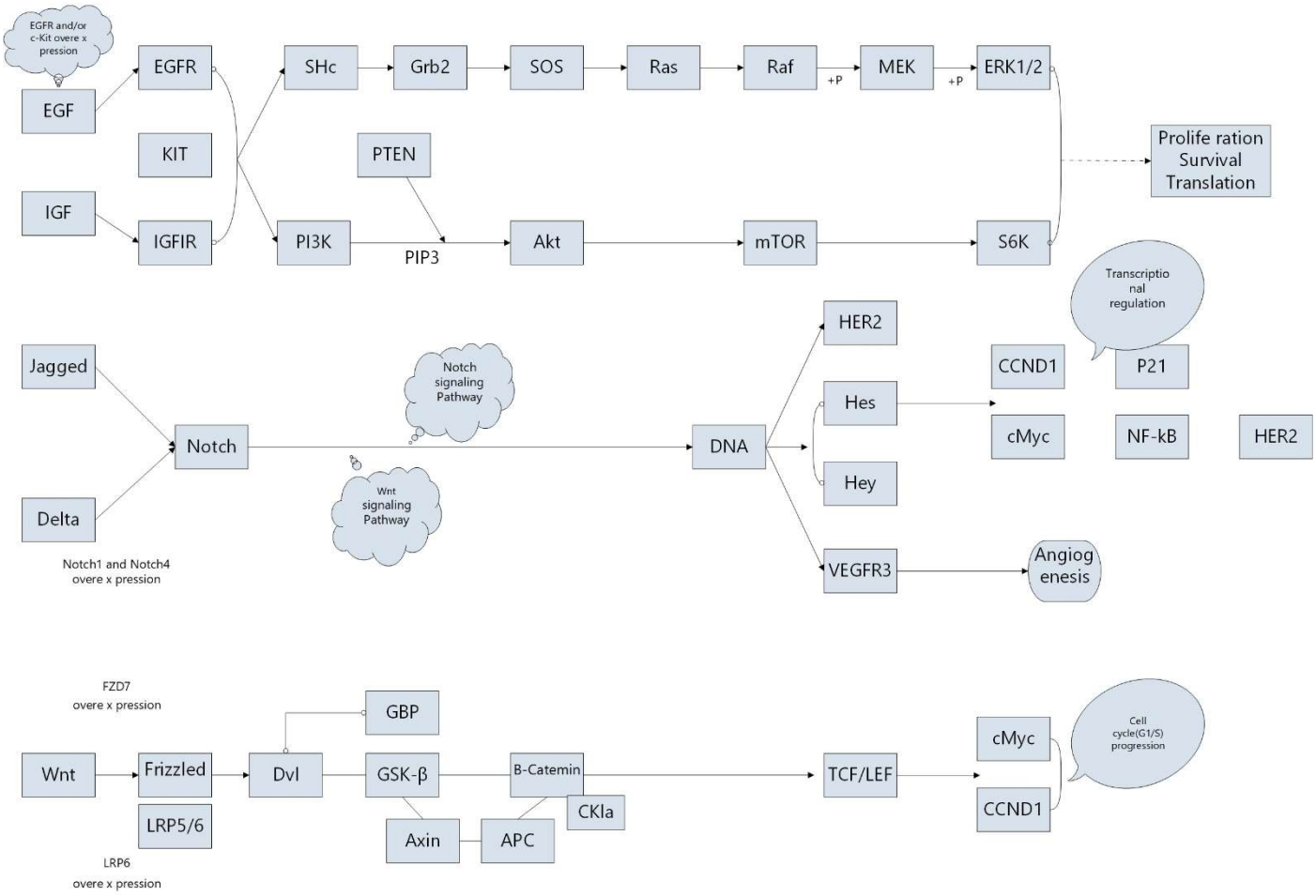

Figure 2

Figure 2. This figure demonstrates the major metabolic pathways of triple-negative breast cancer.

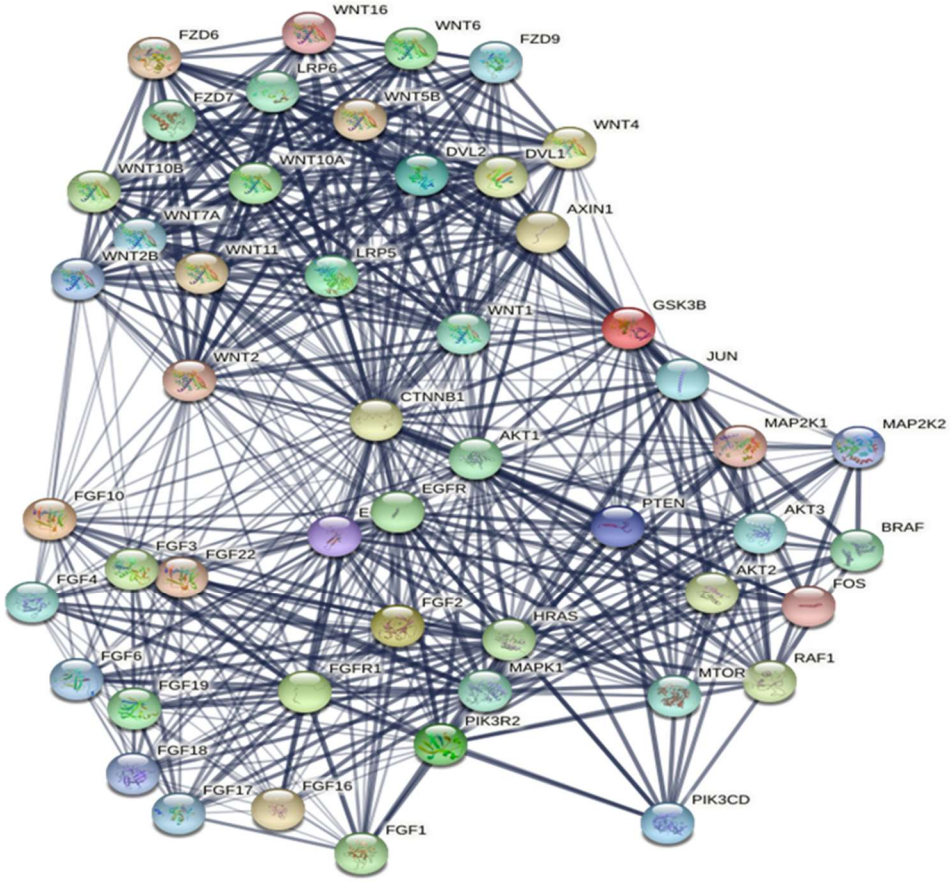

Figure 3

Figure 3. This figure is retrieved from the Genecards.[50] This figure provides a network view of gene-gene interaction in breast cancer.

2.2.2 Network Propagation

Since each gene only corresponds to a few biomolecules, the data we obtained are sparse, which is not applicable for machine learning techniques.[48] Here we introduce the supervised random walk method to amplify the gene mutation effect on breast cancer. To avoid the erosion of pathway signals and the complication of cancer subtype identification, we adapt the mathematical framework established by Zhang et al., who augmented a cost function term in the original supervised random walk model applied in social network analyzing.[63,64]

Table 1

Notation	Interpretation
V	The set containing genes as nodes
E	The set containing the molecular interactions as edges
G	The graph composed of nodes and edges
P^b	The tumor-by-gene binary matrix ($b = 0$ for wild-type and 1 for mutated type)
Q	The degree normalized adjacency matrix of a graph
α	The probability for a mutated gene to mutate back to the wild-type gene
$P^{(t)}$	The random walk score matrix at time t
P	The stable random walk score matrix
a_{ij}	The activation score for a molecular interaction
$\vec{x}_{ij}$	The feature vector corresponding to the vertices (i, j)
$\vec{w}$	The feature weight vector containing weights for each feature
M	The total number of the tumors

$\vec{\boldsymbol{p}}_{\boldsymbol{u}}$	The u-th row of the matrix P, indicates the mutation profile of tumor u
$\vec{\boldsymbol{c}}_{\boldsymbol{A}}$	The centroid vector of the true subtype A of u
M_A	The number of tumors in subtype A
λ	The hyperparameter used for parameters' sparsity control
β	The hyperparameter used for model nonlinearity control
S	The set of all subtypes
$\mathrm{F}(\cdot)$	The edge strength function takes feature vectors as input and outputs the corresponding strength
$f(\cdot)$	The encoder function takes feature vectors as input for future processing
$g(\cdot)$	The decoder function takes the processed feature vectors as input and outputs centroids and latent features
ω_i	The ith layer weight for stacked auto encoder
b_i	The ith layer bias for stacked auto encoder
$\tilde{P}$	The auxiliary distribution matrix
$\tilde{p}_{ij}$	The (i,j) component of the matrix $\tilde{P}$
$\tilde{Q}$	The transformation matrix
$\tilde{q}_{ij}$	The (i,j) component of the matrix $\tilde{Q}$
z_i	The ith component of latent feature
f_j	The frequency of the jth soft cluster
μ_j	The centroid of the jth column
$\vec{\boldsymbol{u}}$	The output vector of the decoder

The interaction among genes can be viewed as a network with each gene as a node and their relationships as edges. Now that the layout of the graph $G =< V, E >$ is given. Since both the network of the graph and the relationship among nodes contain partial information, we take them into account. As the network propagates with time, the random walk process can be represented as

$$P^{(t)} = (1-\alpha)\cdot P^{(t-1)}\cdot Q + \alpha\cdot P^{0} \tag{1}$$

The P^{b} term represents the mutation profile of each tumor. As the convergence criterion is met, i.e.,

$$P^{(t)} \approx P^{(t-1)} \tag{2}$$

The obtained stationary random walk score matrix $P^{(t)}$ becomes a smoothed tumor-by-gene matrix, which is referred to as the propagated mutation profiles.[63] The matrix Q is updated in each iteration by

$$q_{ij} = \frac{I\big((i,j)\in E\big)\cdot a_{ij}}{\sum_k a_{ik}} \tag{3}$$

Where

$$a_{ij} = \frac{1}{1+e^{-\vec{\boldsymbol{w}}\cdot\vec{\boldsymbol{x}}_{ij}}} \tag{4}$$

The feature weight vector $\vec{\boldsymbol{w}}$ is obtained as the process iterates. The centroid vector of the breast cancer subtype a is defined as the average of all profiles in subtype A besides u, in order to prevent the leak of information during the iteration.

$$\vec{\boldsymbol{c}}_{\boldsymbol{A}} \coloneqq \frac{1}{M_A-1}\sum_{v\in A, v\notin u}\vec{\boldsymbol{p}}_{\boldsymbol{v}} \tag{5}$$

The optimization problem can be stated as below.

$$\min_{\vec{\boldsymbol{w}}} F(\vec{\boldsymbol{w}}) = \lambda\cdot\|\vec{\boldsymbol{w}}\|_1 + \sum_{m=1}^{M}\frac{1}{1+e^{-\beta\cdot D_u}} = \lambda\cdot\|\vec{\boldsymbol{w}}\|_1 + f(\vec{\boldsymbol{w}}) \tag{6}$$

$$D_u = \|\vec{\boldsymbol{p}}_{\boldsymbol{u}} - \vec{\boldsymbol{c}}_{\boldsymbol{A}}\|_2^2 - \min_{B\neq A}\|\vec{\boldsymbol{p}}_{\boldsymbol{u}} - \vec{\boldsymbol{c}}_{\boldsymbol{B}}\|_2^2 \tag{7}$$

The first equation indicates that the edge strength is optimized when for any node, the random walk process is more likely to establish a connection with those who are known connected to the node. The second term introduces the effect of classification error, which has a great sensitivity due to the property of the sigmoid function.[65] The second equation exhibits the quantity difference of the distance between tumor u and its corresponding centroid and the distance to the nearest centroid.

The edge strength function can be obtained using the gradient descent method.[66] By taking the partial derivative of the above two equations with respect to the weight vector, we have

$$\frac{\partial F(\vec{\boldsymbol{w}})}{\partial \vec{\boldsymbol{w}}} = \mathrm{sgn}(\vec{\boldsymbol{w}}) \cdot \lambda + \sum_{m=1}^{M} \beta \cdot f(\vec{\boldsymbol{w}}) \cdot (1 - f(\vec{\boldsymbol{w}})) \cdot \frac{\partial D_m}{\partial \vec{\boldsymbol{w}}} \tag{8}$$

$$\frac{\partial D_m}{\partial \vec{\boldsymbol{w}}} = 2(\vec{\boldsymbol{p}}_{\boldsymbol{u}} - \vec{\boldsymbol{c}}_{\boldsymbol{A}}) \cdot \left(\frac{\partial \vec{\boldsymbol{p}}_{\boldsymbol{u}}}{\partial \vec{\boldsymbol{w}}} - \frac{\partial \vec{\boldsymbol{c}}_{\boldsymbol{A}}}{\partial \vec{\boldsymbol{w}}}\right) - 2(\vec{\boldsymbol{p}}_{\boldsymbol{u}} - \vec{\boldsymbol{c}}_{\boldsymbol{B}}) \cdot \left(\frac{\partial \vec{\boldsymbol{p}}_{\boldsymbol{u}}}{\partial \vec{\boldsymbol{w}}} - \frac{\partial \vec{\boldsymbol{c}}_{\boldsymbol{B}}}{\partial \vec{\boldsymbol{w}}}\right) \tag{9}$$

Calculating the above two formulas requires the partial derivative of $\vec{\boldsymbol{p}}_{\boldsymbol{u}}$ with respect to the weight vector, then we take the partial derivatives of the matrix P, Q.

$$\frac{\partial P}{\partial \vec{\boldsymbol{w}}} = (1 - \alpha) \cdot \left(\frac{\partial P}{\partial \vec{\boldsymbol{w}}} \cdot Q + \frac{\partial Q}{\partial \vec{\boldsymbol{w}}} \cdot P\right) \tag{10}$$

$$\frac{\partial q_{ij}}{\partial \vec{\boldsymbol{w}}} = \frac{I\big((i,j) \in E\big) \cdot \left(\frac{\partial a_{ij}}{\partial \vec{\boldsymbol{w}}} \cdot \sum_k a_{ik} + a_{ij} \cdot \sum_k \frac{\partial a_{ik}}{\partial \vec{\boldsymbol{w}}}\right)}{(\sum_k a_{ik})^2} \tag{11}$$

Where

$$\frac{\partial a_{ij}}{\partial \vec{\boldsymbol{w}}} = x_{ij} \cdot a_{ij} \cdot \big(1 - a_{ij}\big) \tag{12}$$

Thus, we have obtained all the mathematical framework for the random walk-based network propagation.

Based on the network established, the categorization of a tumor x by its profile P_x^0 can be expressed as

$$X = arg\min_{X \in S} \|\vec{\boldsymbol{p}}_{\boldsymbol{x}} - \vec{\boldsymbol{c}}_{\boldsymbol{X}}\|_2^2 \tag{13}$$

X is the type that tumor x belongs to. The propagated mutation profile is obtained by iteration.

2.3 Machine Learning Techniques

Several studies have validated machine learning as an adjuvant strategy for cancer stratification. Machine learning can aid in the construction of gene-protein networks, data clustering, and prognosis under the network propagation paradigm outlined. We discuss them separately.

Machine learning aided prognosis emerged in recent years. Mark et al. suggested a unique method for identifying somatic mutation-based prognostic signatures using a neural network. They divided the data at random into two sets, one for training and one for validation. They identified the mutant genes related to patient survival by utilizing single variate Cox regression. The multivariate Cox regression is then subjected to bidirectional stepwise model selection to yield the final predictive somatic mutation gene signature. The final signature is used to compute risk scores for each patient in the dataset.[67] The flow chart is shown in **Figure 4**.

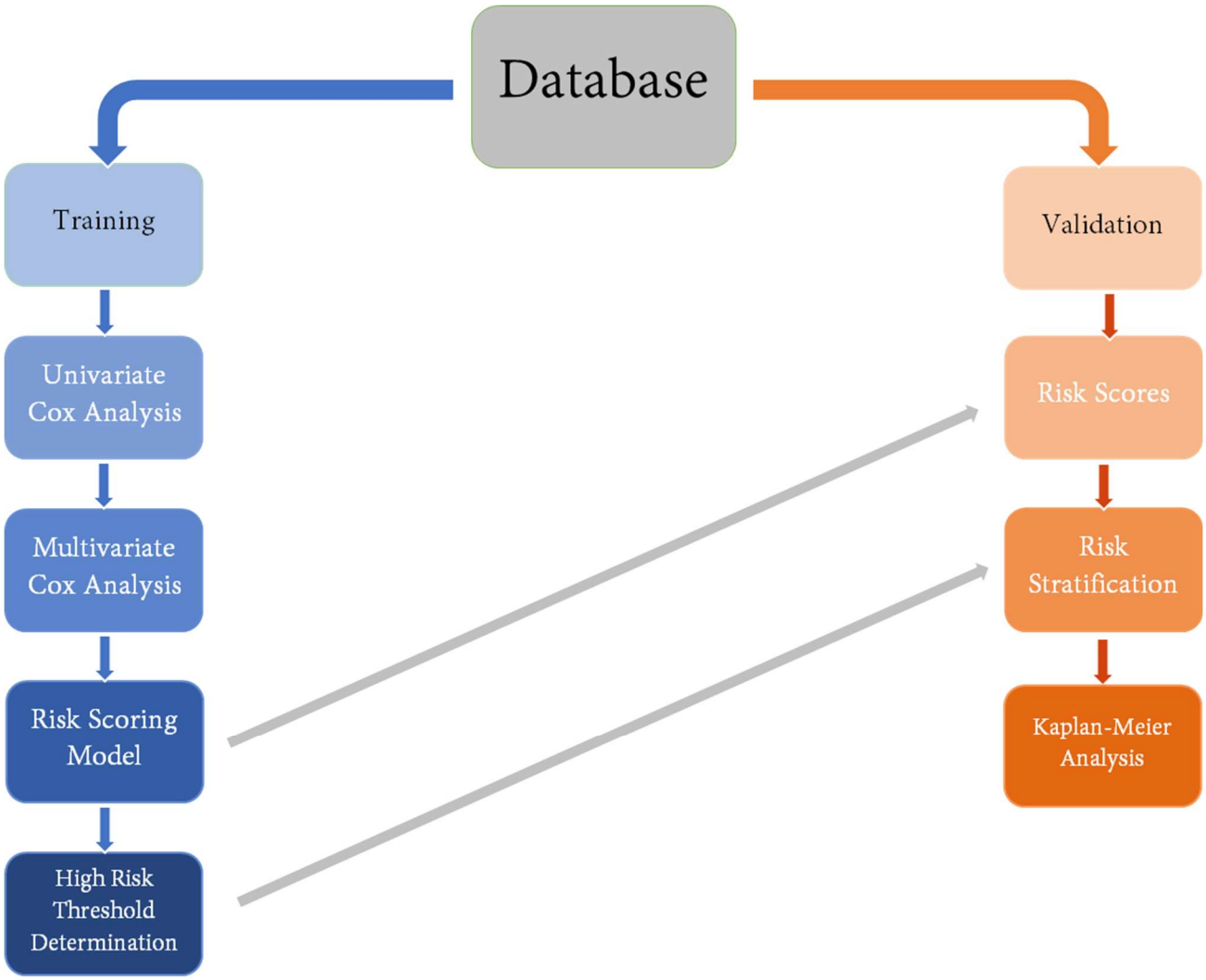

Figure 4

Machine learning techniques are used in cancer subtype clustering to improve model accuracy and clustering convergence speed. Suleyman et al. investigated the robustness and accuracy of five major techniques: random forest, support vector machine, C4.5, naïve Bayesian, and k-nearest neighbor. They discovered that when dealing with sparse data, such as somatic mutation, the random forest works well. [68] Narjes et al. used deep embedded clustering to classify breast cancer into four categories.[69] We include DEC in our discussion since it shows potential for future research utilizing the methodology proposed by Narjes et al.

DEC is an iterative method combining clustering and feature embedment. There are two major procedures in each cycle. The data in Euclidean space are transformed into a parametric space using a stack automatic encoder, where the similarity of the original sparse data is increased. To cluster the data in the parametric space, k-means is used. The centroids are refreshed in each iteration using Kullback-Leibler divergence minimization and gradually approach the defined underlying real subtype.

The construction of the gene-protein network was supported by machine learning. Zhang et al. used a random walk within the context of a machine learning-based gene-protein network to classify cancers based on somatic mutations.[70] Their model NBS^2 is designed as a supervised model, where a minimization target function is designed towards the Euclidean distance-based tumor classification error.

3. Prognosis

The purpose of this section is to verify the validity of the obtained stratification of breast cancer. For completeness, we give the mathematical framework of the prognosis model in *Section 3.1* and a pilot result interpretation in *Section 3.2*.

3.1 Proportional Hazards Model

To verify the stratification of breast cancer subtypes obtained in *Section 2*, we perform the prognosis analysis in this section. First and foremost, we list the mathematical symbols in Table 2 for later convenience.

Table 2

Notation	Interpretation
t	The interested time ($t > 0$ is assumed)
t_{ji}	The survey time of the j-th individual
d_i	The reported number of death occurred at t_i
n_i	The number of individuals known to have survived at t_i
T	The continuous random variable indicating the time
X_{jk}	The random variable indicating the expression level of the k-th gene on the j-th individual
$\overrightarrow{\boldsymbol{X_j}}$	The gene expression vector of the j-th individual
β_{mk}	The weight for the k-th gene belonging to the m-th cluster
$\overrightarrow{\boldsymbol{\beta_m}}$	The synthesized weight factor for the selected genes
J	The total number of data samples/individuals
M	The total number of clusters obtained
N	The total number of the genes we selected
$S(\cdot)$	The survival function
$\lambda(\cdot)$	The hazard function
$P(\cdot)$	The probability function
c_j	The censoring time of the j-th individual

The survival function $S(t)$ indicates the probability that a patient of interest will survive at time t.[71] Since the survival function cannot be derived directly, we apply the Kaplan-Meier estimator of $S(t)$ for approximation, which is represented with a hat[72]

$$\hat{S}(t) = \prod_{i:t_i \le t} \left(1 - \frac{d_i}{n_i}\right) \tag{14}$$

However, since the Kaplan-Meier estimator doesn't take covariates into account, we introduce the Breslow estimator to derive the estimator for the survival function.[27] Since the survival function can also be expressed as

$$S(t) = P\{T > t\} = 1 - F(t) \tag{15}$$

Where $\mathrm{F}(\cdot)$ indicates the cumulative distribution function. Thus, we can apply the quantities defined in the probability theory to elucidate the effect brought by genes. The probability density function $f(\cdot)$ satisfies

$$f(t) = -\frac{d}{dt}S(t) \tag{16}$$

The hazard function is introduced to measure the event rate at the interested time, by definition,[73]

$$\lambda(\mathrm{t}) = \lim_{\mathrm{dt}\to\infty} \frac{P(t \le T \le t + dt)}{S(t) \cdot dt} \tag{17}$$

Which is tightly linked to the survival function. Combining the above two formulas, their relationship satisfies

$$\lambda(\mathrm{t}) = -\frac{d}{dt} lgS(t) \tag{18}$$

By integrating Formula (18) over time, we get the cumulative hazard function, indicating the hazard accumulated over time.

$$\Lambda(\mathrm{t}) = \int_0^t \lambda(x)dx \tag{19}$$

Considering the censoring time, we make the following definitions for later convenience.

$$\widetilde{t_{jl}} := \min\left(t_{ji}, c_j\right) \tag{20}$$

$$\Delta_j := \mathrm{I}\left(t_{ji} < c_j\right) \tag{21}$$

$$\mathcal{R}_k := \left\{k : \widetilde{t_{kl}} \geq \widetilde{t_{jl}}\right\} \tag{22}$$

$\widetilde{t_{jl}}$ indicates the latest time for valid information. Δ_j indicates the validity of information at the interested time. We now introduce the proportional hazards model to verify the utility of the stratification of breast cancer subtypes we obtained.

Assumption 1. The covariates (X_{jk}) are multiplicatively related to the hazard according to the condition of the proportional hazard.[73]

Assumption 2. The covariates are constant with respect to time.

Following the above assumption, we have

$$\lambda_j\left(t|\overrightarrow{\boldsymbol{X_j}}\right) = \lambda_0(t) \cdot e^{\overrightarrow{X_j} \cdot \overrightarrow{\beta_m}} \tag{23}$$

Where $\lambda_0(t)$ indicates the baseline hazard function obtained using Breslow's estimation, whose core idea is the maximum likelihood.[74–77] The estimator for the accumulative baseline hazard function $\Lambda_0(\cdot)$ and the synthesized weight vector $\overrightarrow{\boldsymbol{\beta_m}}$ can be expressed as[75]

$$\widehat{\Lambda}_0(t) = \sum_{j=1}^{J} \frac{\Delta_j \cdot I\left(\widetilde{t_{jl}} \leq t\right)}{\sum_{k \in \mathcal{R}_k} e^{\widehat{\boldsymbol{\beta_m}} \cdot \overrightarrow{\boldsymbol{X_j}}}} \tag{24}$$

$$\widehat{\boldsymbol{\beta_m}} = \prod_{j=1}^{J} \left(\frac{e^{\overrightarrow{\boldsymbol{\beta_m}} \cdot \overrightarrow{\boldsymbol{X_j}}}}{\sum_{k \in \mathcal{R}_k} e^{\overrightarrow{\boldsymbol{\beta_m}} \cdot \overrightarrow{\boldsymbol{X_j}}}}\right)^{\Delta_j} \tag{25}$$

Thus, we have the estimator for the survival function

$$\hat{S}(t) = e^{-\int_0^t e^{\widehat{\boldsymbol{\beta_m}} \cdot \overrightarrow{\boldsymbol{X_j}}} d\widehat{\lambda}_0(u)} \tag{26}$$

3.2 Result

Using the data obtained from the TCGA project, we select the non-Hispanic, non-Latin Caucasian females to eliminate the effect of irrelevant factors. Of the total 560 cases, the overall survival Kaplan-Meier plot is shown in **Figure 5**.

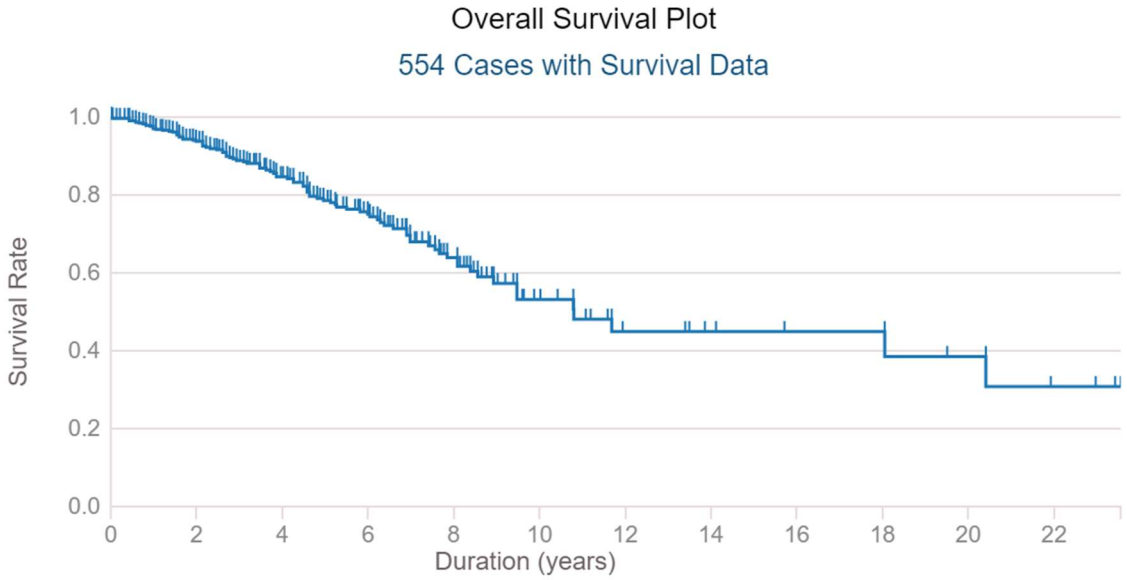

Figure 5

Figure 5. This figure depicts the overall survival rate with respect to time.

By comparing the genes with normal mutations, we obtain the relevant mutated genes. A bar chart is drawn to illustrate the mutation

frequency of genes. We can directly see in Figure 5 that the most frequently mutated gene is PIK3CA, which affects more than 30% of the total cases. PIK3CA genes are known as the regulators of cellular growth and proliferation.[78] Followed by the TP53 gene, whose corresponding protein is a multifunctional transcription factor involved in cell cycle progression, DNA maintenance, and DNA damage repair.[79,80] Other genes affecting more than 10% and less than 20% of the total cases are CDH1, MUC16, and GATA3. CDH1 is presumed as a tumor suppressor controlling the expression of a cell-cell adhesion molecule.[81] MUC16 mutation may be associated with superior response to ICIs in patients with solid tumors, thus could be used to stratify patients into prognostically distinct groups.[82,83] GATA3 is an essential transcription factor relevant to immune responses.[84] However, it seems that only in ER-positive patients does GATA3 mutation status matter.[85] Jiang et al. defined a unique subtype based on the GATA3 mutation status.[86]

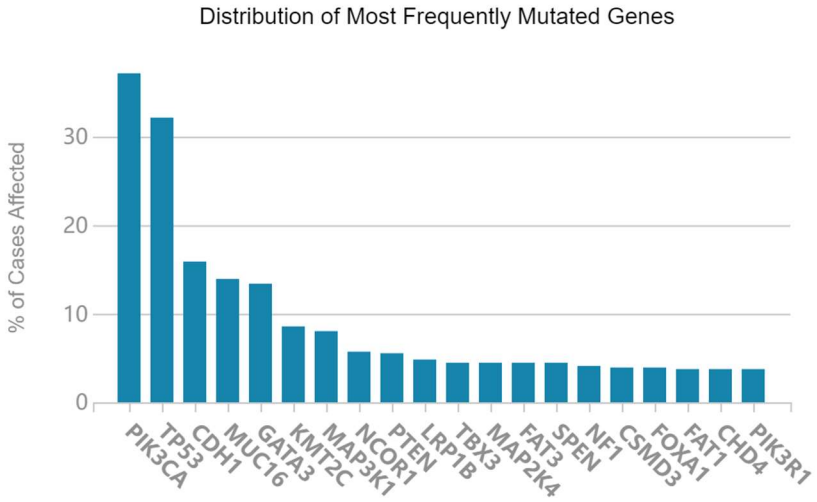

Figure 6

Figure 6. The top 20 most frequently mutated genes in breast cancer. The horizontal axis displays the names of the genes while the vertical axis displays the percentage of the affected cases.

And the effect of each mutated gene can be directly observed in the Kaplan-Meier plot. For instance, the PIK3CA H1047R mutated cases have a similar drop curve in the survival rate in initial 10 years but a relatively higher survival rate after 20 years. As is reported in prior studies, PIK3CA gene mutations has a positive prognostic significance and are vital for PI3K-targeted therapy.[87]

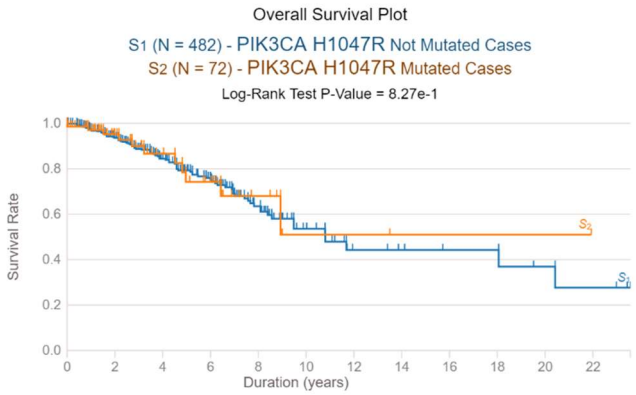

Figure 7

Figure 7. The overall survival plot with respect to the PIK3CA H1047R gene mutation status. The orange line S2 indicates the mutated cases and the blue line S1 indicates the not mutated cases.

However, no valid conclusions can be drawn since breast cancer-induced death is affected by the whole gene-protein network. To give a more detailed illustration, we plot **Figure 8**.

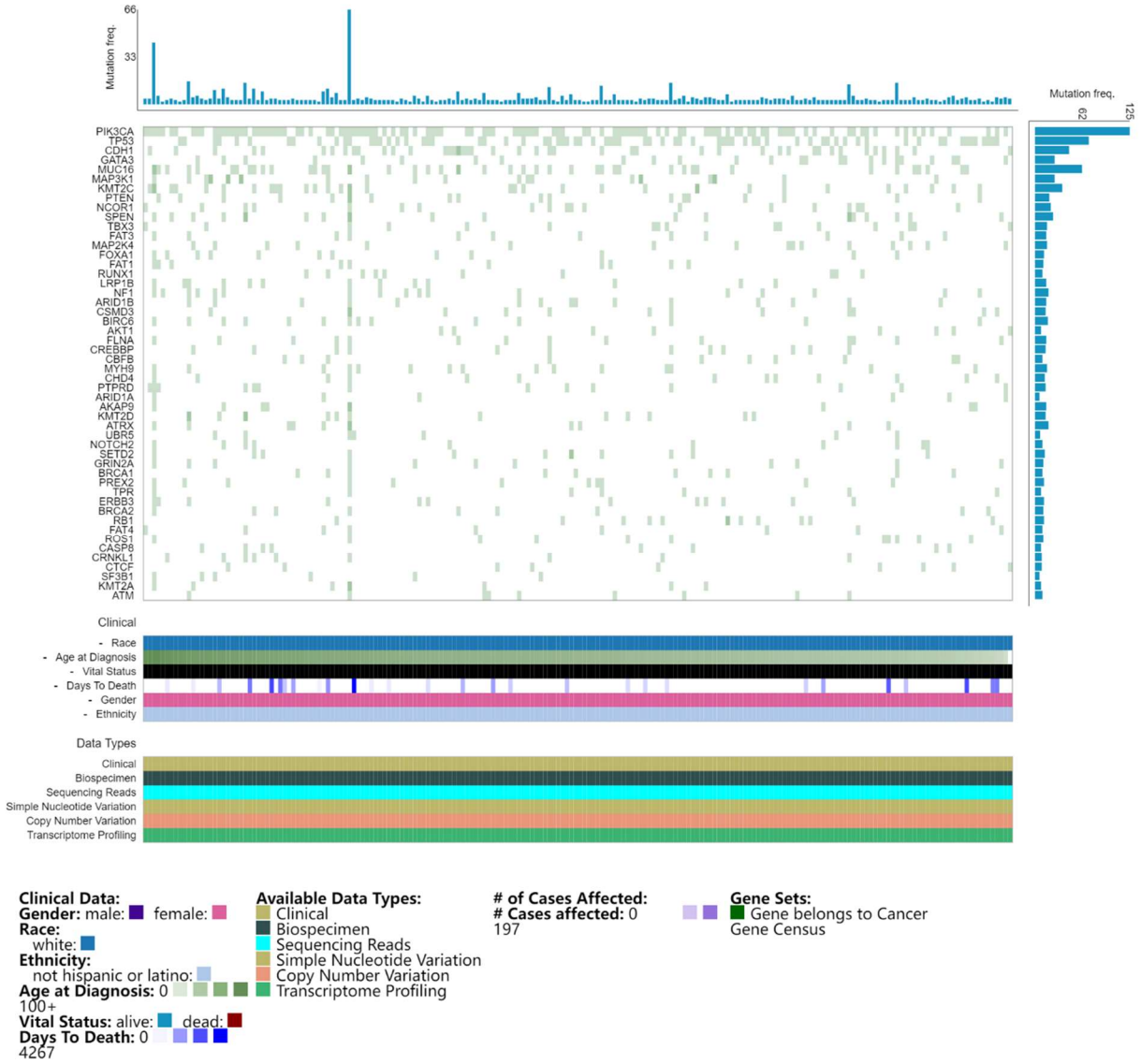

Figure 8

Figure 8. 200 most mutated cases and top 50 most mutated genes by somatic mutation.

The horizontal rows indicate the genes in an order of mutation frequency, the vertical columns indicate patient groups that have similar gene expression statuses. The horizontal rows are arranged by the affected cases and the vertical rows are arranged by the age at diagnosis. In the clinical columns, we can see that the crowd corresponds to the white race and all the cases are female. Moreover, all of our data have the corresponding biospecimen data, sequencing reads, simple nucleotide variation data, copy number data, and transcriptome profiling data. An illustration of the stratified data is shown in **Figure 9**.

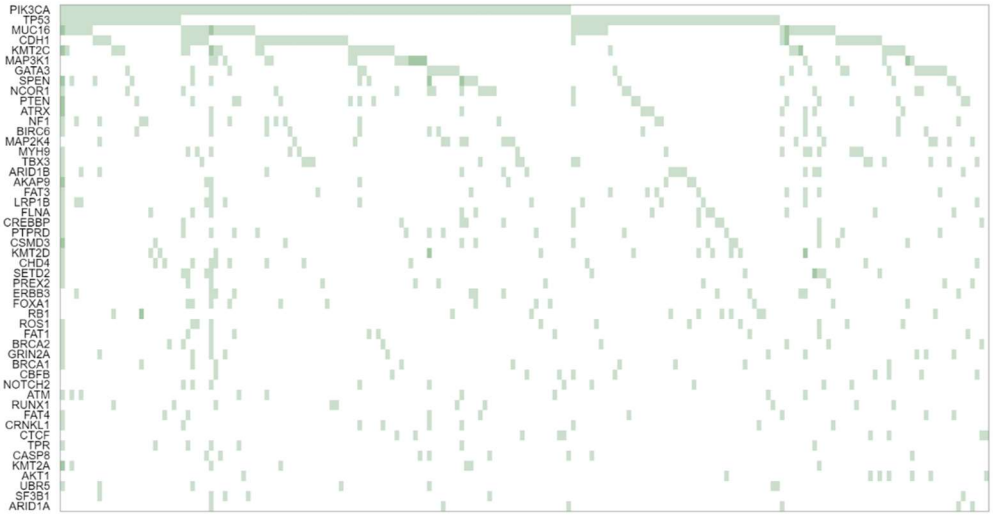

Figure 9

Figure 9. The rearranged heat map.

4. Conclusions

In this summary, we discuss the current mathematical breast cancer stratification methods for its significance in breast cancer treatment method selection. Firstly, we discussed the previous clustering methods and novel algorithms. Due to the nature of somatic mutation and its analogous relationship with network, network propagation methods along with the random walk process are discussed. In the last section, we give a mathematical framework for prognosis. Using the data retrieved from the TCGA project, we illustrate the effect induced by somatic gene mutation.

Author Contributions: Conceptualization, Hongying Zhong, Ruotao Yu, Shengqi Liu; draft preparation, Ruotao Yu, Shengqi Liu, Jiarui Chen, Qianhui Li; writing, review, and editing, Ruotao Yu, Shengqi Liu, Jiarui Chen, Qianhui Li; code and Figure: Shengqi Liu, Ruotao Yu; supervision, Hongying Zhong;

All authors have read and agreed to the published version of the manuscript.

Funding: No funding received.

Conflicts of Interest: The authors declare no conflict of interest.

Citations

1. Pusztai, L., Mazouni, C., Anderson, K., Wu, Y. & Symmans, W. F. Molecular classification of breast cancer: limitations and potential. *Oncologist* **11**, 868–877 (2006).
2. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
3. Group, E. B. C. T. C. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *The Lancet* **365**, 1687–1717 (2005).
4. Güler, E. N. Gene expression profiling in breast cancer and its effect on therapy selection in early-stage breast cancer. *Eur J Breast Health* **13**, 168 (2017).
5. Vandergrift, J. L. *et al.* Time to adjuvant chemotherapy for breast cancer in National Comprehensive Cancer Network institutions. *J Natl Cancer Inst* **105**, 104–112 (2013).
6. Macgregor, P. F. & Squire, J. A. Application of microarrays to the analysis of gene expression in cancer. *Clin Chem* **48**, 1170–1177 (2002).
7. Schnitt, S. J. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. *Modern pathology* **23**, S60–S64 (2010).
8. Bast Jr, R. C. *et al.* 2000 update of recommendations for the use of tumor markers in breast and colorectal cancer: clinical practice guidelines of the American Society of Clinical Oncology. *Journal of clinical oncology* **19**, 1865–1878 (2001).
9. Alizadeh, A. A., Ross, D. T., Perou, C. M. & van de Rijn, M. Towards a novel classification of human malignancies based on gene expression patterns. *J Pathol* **195**, 41–52 (2001).
10. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (1979)* **270**, 467–470 (1995).
11. Shalon, D., Smith, S. J. & Brown, P. O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* **6**, 639–645 (1996).
12. Pease, A. C. *et al.* Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proceedings of the National Academy of Sciences* **91**, 5022–5026 (1994).
13. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression. *Nat. genet* **14**, 457–460 (1996).
14. Dubsky, P. *et al.* The EndoPredict score provides prognostic information on late distant metastases in ER+/HER2− breast cancer patients. *Br J Cancer* **109**, 2959–2964 (2013).
15. Zhang, Y. *et al.* Breast Cancer Index Identifies Early-Stage Estrogen Receptor–Positive Breast Cancer Patients at Risk for Early-and Late-Distant RecurrencePrediction of Early-and Late-Breast Cancer Recurrence. *Clinical Cancer Research* **19**, 4196–4205 (2013).
16. Gnant, M. *et al.* Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone. *Annals of oncology* **25**, 339–345 (2014).
17. Slodkowska, E. A. & Ross, J. S. MammaPrint™ 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Rev Mol Diagn* **9**, 417–422 (2009).
18. McVeigh, T. P. & Kerin, M. J. Clinical use of the Oncotype DX genomic test to guide treatment decisions for patients with invasive breast cancer. *Breast Cancer: Targets and Therapy* **9**, 393 (2017).
19. Cheang, M. C. U., van de Rijn, M. & Nielsen, T. O. Gene expression profiling of breast cancer. *Annu. Rev. Pathol. Mech. Dis.* **3**, 67–97 (2008).
20. Gusterson, B. Do'basal-like'breast cancers really exist? *Nat Rev Cancer* **9**, 128–134 (2009).
21. 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**, 263–280 (2010).
22. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat Commun* **7**, 1–16 (2016).
23. Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W. & Quackenbush, J. Cancer subtype identification using somatic mutation data. *Br J Cancer* **118**, 1492–1501 (2018).
24. Maruyama, N. *et al.* Clinicopathologic analysis of breast cancers with PIK3CA mutations in Japanese women. *Clinical cancer research*

13, 408–414 (2007).

25. Mangone, F. R., Bobrovnitchaia, I. G., Salaorni, S., Manuli, E. & Nagai, M. A. PIK3CA exon 20 mutations are associated with poor prognosis in breast cancer patients. *Clinics* **67**, 1285–1290 (2012).
26. Tavtigian, S. V *et al.* Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet* **43**, 295–305 (2006).
27. Jain, A. K. & Dubes, R. C. Algorithms for Clustering Data. Preprint at (1998).
28. Oyelade, J. *et al.* Clustering algorithms: their application to gene expression data. *Bioinform Biol Insights* **10**, BBI-S38316 (2016).
29. Fa, R., Nandi, A. K. & Gong, L.-Y. Clustering analysis for gene expression data: A methodological review. in *2012 5th International Symposium on Communications, Control and Signal Processing* 1–6 (IEEE, 2012).
30. Belacel, N., Wang, Q. & Cuperlovic-Culf, M. Clustering methods for microarray gene expression data. *OMICS* **10**, 507–531 (2006).
31. D'haeseleer, P. How does gene expression clustering work? *Nat Biotechnol* **23**, 1499–1501 (2005).
32. Kerr, G., Ruskin, H. J., Crane, M. & Doolan, P. Techniques for clustering gene expression data. *Comput Biol Med* **38**, 283–293 (2008).
33. Bailey, J. Alternative clustering analysis: A review. *Data Clustering* 535–550 (2018).
34. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
35. Dhillon, I. S., Guan, Y. & Kulis, B. Kernel k-means: spectral clustering and normalized cuts. in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* 551–556 (2004).
36. Sehhati, M., Tabatabaiefar, M. A., Gholami, A. H. & Sattari, M. Using classification and K-means methods to predict breast cancer recurrence in gene expression data. *J Med Signals Sens* **12**, 122–126 (2022).
37. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. in *kdd* vol. 96 226–231 (1996).
38. Ankerst, M., Breunig, M. M., Kriegel, H.-P. & Sander, J. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record* **28**, 49–60 (1999).
39. Berkhin, P. A survey of clustering data mining techniques. in *Grouping multidimensional data* 25–71 (Springer, 2006).
40. Brulé, J. PSCAN: parallel, density based clustering of protein sequences. *Intell Data Anal* **1**, 48–57 (2015).
41. Edla, D. R., Jana, P. K. & Member, I. S. A prototype-based modified DBSCAN for gene clustering. *Procedia Technology* **6**, 485–492 (2012).
42. Wang, W., Yang, J. & Muntz, R. STING: A statistical information grid approach to spatial data mining. in *Vldb* vol. 97 186–195 (Citeseer, 1997).
43. Hinneburg, A. & Keim, D. A. Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. in (1999).
44. Amor, R. del, Colomer, A., Monteagudo, C. & Naranjo, V. A deep embedded refined clustering approach for breast cancer distinction based on DNA methylation. *Neural Comput Appl* **34**, 10243–10255 (2022).
45. Zhao, P., Xu, Z., Chen, J., Ren, Y. & King, I. Single Cell Self-Paced Clustering with Transcriptome Sequencing Data. *Int J Mol Sci* **23**, 3900 (2022).
46. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* **45**, 1113–1120 (2013).
47. Barrett, T. *et al.* NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res* **33**, D562–D566 (2005).
48. Bamford, S. *et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* **91**, 355–358 (2004).
49. Mering, C. von *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **31**, 258–261 (2003).
50. Safran, M. *et al.* GeneCards Version 3: the human gene integrator. *Database* **2010**, (2010).
51. Kanehisa, M. The KEGG database. in *Novartis foundation symposium* 91–100 (Wiley Online Library, 2002).
52. Mcpherson, K., Steel, C. M. & Dixon, J. M. Breast cancer — epidemiology , risk factors , and genetics Risk factors for breast cancer. *Mortality* **321**, (2000).
53. Kukasiewicz, S. *et al.* Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—An updated review. *Cancers* vol. 13 1–30 Preprint at https://doi.org/10.3390/cancers13174287 (2021).
54. Verma, R., Bowen, R. L., Slater, S. E., Mihaimeed, F. & Jones, J. L. Pathological and epidemiological factors associated with advanced stage at diagnosis of breast cancer. *British Medical Bulletin* vol. 103 129–145 Preprint at https://doi.org/10.1093/bmb/lds018 (2012).
55. Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. Protein function in the post-genomic era. *Nature* **405**, 823–826 (2000).

56. Ratkaj, I. *et al.* Integrated gene networks in breast cancer development. *Funct Integr Genomics* **10**, 11–19 (2010).
57. Raman, K. Construction and analysis of protein–protein interaction networks. *Autom Exp* **2**, 1–11 (2010).
58. Hozhabri, H. *et al.* Comparative analysis of protein-protein interaction networks in metastatic breast cancer. *PLoS One* **17**, e0260584 (2022).
59. van den Akker, E. *et al.* Integrating protein-protein interaction networks with gene-gene co-expression networks improves gene signatures for classifying breast cancer metastasis. *J Integr Bioinform* **8**, 222–238 (2011).
60. Brohee, S. & Van Helden, J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7**, 1–19 (2006).
61. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
62. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).
63. Backstrom, L. & Leskovec, J. Supervised random walks: predicting and recommending links in social networks. in *Proceedings of the fourth ACM international conference on Web search and data mining* 635–644 (2011).
64. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat Methods* **10**, 1108–1115 (2013).
65. Yan, L., Dodier, R. H., Mozer, M. & Wolniewicz, R. H. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. in *Proceedings of the 20th international conference on machine learning (icml-03)* 848–855 (2003).
66. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
67. Menor, M. *et al.* Development of somatic mutation signatures for risk stratification and prognosis in lung and colorectal adenocarcinomas. *BMC Med Genomics* **12**, 63–79 (2019).
68. Vural, S., Wang, X. & Guda, C. Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC Syst Biol* **10**, 263–276 (2016).
69. Rohani, N. & Eslahchi, C. Classifying breast cancer molecular subtypes by using deep clustering approach. *Front Genet* 1108 (2020).
70. Zhang, W., Ma, J. & Ideker, T. Classifying tumors by supervised network propagation. *Bioinformatics* **34**, i484–i493 (2018).
71. Kleinbaum, D. G. & Klein, M. *Survival analysis: a self-learning text*. vol. 3 (Springer, 2012).
72. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* **53**, 457–481 (1958).
73. Finkelstein, M. *Failure rate modelling for reliability and risk*. (Springer Science & Business Media, 2008).
74. Breslow, N. E. Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique* 45–57 (1975).
75. Kumar, D. & Klefsjö, B. Proportional hazards model: a review. *Reliab Eng Syst Saf* **44**, 177–188 (1994).
76. Crowley, J. & Breslow, N. Statistical analysis of survival data. *Annu Rev Public Health* **5**, 385–411 (1984).
77. Lin, D. Y. On the Breslow estimator. *Lifetime Data Anal* **13**, 471–480 (2007).
78. Bachman, K. E. *et al.* The PIK3CA gene is mutated with high frequency in human breast cancers. *Cancer Biol Ther* **3**, 772–775 (2004).
79. Varna, M., Bousquet, G., Plassa, L.-F., Bertheau, P. & Janin, A. TP53 status and response to treatment in breast cancers. *J Biomed Biotechnol* **2011**, (2011).
80. Børresen-Dale, A. TP53 and breast cancer. *Hum Mutat* **21**, 292–300 (2003).
81. Berx, G., Becker, K., Höfler, H. & Van Roy, F. Mutations of the human E-cadherin (CDH1) gene. *Hum Mutat* **12**, 226–237 (1998).
82. Zhang, L., Han, X. & Shi, Y. Association of MUC16 mutation with response to immune checkpoint inhibitors in solid tumors. *JAMA Netw Open* **3**, e2013201–e2013201 (2020).
83. Li, X., Pasche, B., Zhang, W. & Chen, K. Association of MUC16 mutation with tumor mutation load and outcomes in patients with gastric cancer. *JAMA Oncol* **4**, 1691–1698 (2018).
84. Usary, J. *et al.* Mutation of GATA3 in human breast tumors. *Oncogene* **23**, 7669–7678 (2004).
85. Afzaljavan, F., Sadr, A. S., Savas, S. & Pasdar, A. GATA3 somatic mutations are associated with clinicopathological features and expression profile in TCGA breast cancer patients. *Sci Rep* **11**, 1–13 (2021).
86. Jiang, Y., Yu, K., Zuo, W., Peng, W. & Shao, Z. GATA3 mutations define a unique subtype of luminal-like breast cancer with improved survival. *Cancer* **120**, 1329–1337 (2014).
87. Kalinsky, K. *et al.* PIK3CA Mutation Associates with Improved Outcome in Breast CancerPIK3CA Mutation and Improved Outcome in Breast Cancer. *Clinical cancer research* **15**, 5049–5059 (2009).