Deep Learning Approach to Identifying Breast Cancer Subtypes Using High-Dimensional Genomic Data

Runpu Chen[§], Le Yang[§], Steve Goodison[‡], Yijun Sun[§][¶][*]

[§]Department of Computer Science and Engineering
[¶]Department of Microbiology and Immunology
The State University of New York at Buffalo, Buffalo, NY 14203
[‡]Department of Health Sciences Research
Mayo Clinic, Jacksonville, FL 32224

Abstract

Motivation: Cancer subtype classification has the potential to significantly improve disease prognosis and develop individualized patient management. Existing methods are limited by their ability to handle extremely high-dimensional data and by the influence of misleading, irrelevant factors, resulting in ambiguous and overlapping subtypes.
Results: To address the above issues, we proposed a novel approach to disentangling and eliminating irrelevant factors by leveraging the power of deep learning. Specifically, we designed a deep-learning framework, referred to as DeepType, that performs joint supervised classification, unsupervised clustering and dimensionality reduction to learn cancer-relevant data representation with cluster structure. We applied DeepType to the METABRIC breast cancer dataset and compared its performance to alternative state-of-the-art methods. DeepType significantly outperformed the existing methods, identifying more robust subtypes while using fewer genes. The new approach provides a framework for the derivation of more accurate and robust molecular cancer subtypes by using increasingly complex, multi-source data.
Availability and implementation: An open-source software package for the proposed method is freely available at www.acsu.buffalo.edu/~yijunsun/lab/DeepType.html.

Bioinformatics
submitted in January 2019, revised in May 2019

[*] Please address all correspondence to Dr. Yijun Sun (yijunsun@buffalo.edu).

1 Introduction

Human cancer is a heterogeneous disease initiated by random somatic mutations and driven by multiple genomic alterations (Hanahan and Weinberg, 2011; Sun *et al.*, 2017). In order to move towards personalized patient treatment regimes, cancers of specific tissues have been divided into molecular subtypes based on the gene expression profiles of primary tumors (Sørlie *et al.*, 2001, 2003; Curtis *et al.*, 2012; Parker *et al.*, 2009). The premise is that patients who share molecular subtypes are likely to have similar disease etiology, response to therapy, and clinical outcomes. Thus, the analysis of molecular subtyping can reveal information valuable for a range of cancer studies from etiology and tumor biology to prognosis and personalized medicine.

Most early work on molecular subtyping for oncology has been performed on data obtained from breast cancer tissues (Sørlie *et al.*, 2001, 2003). Typically, breast cancer is not lethal immediately and thus, there is an opportunity to assist with prognostication and patient management using molecular information. Molecular subtyping of breast cancer initially focused on mRNA data obtained from microarray platforms and parsed molecular profiles to stratify patients according to clinical outcomes (Sørlie *et al.*, 2001). Refinement of the subtype categories through validation in independent datasets identified five broad subtypes, including normal-like, luminal A, luminal B, basal, and HER2+, each with distinct clinical outcomes (Sørlie *et al.*, 2003; Parker *et al.*, 2009). These early studies altered our understanding of breast cancer and offered a foundation for the development of therapies tailored to specific subtypes. However, possibly due to the small number of tumor samples used in initial analyses and the technical limitations of the methods used for gene selection and clustering analysis, several large-scale benchmark studies have demonstrated that the current stratification of breast cancer is only approximate, and that the high degree of ambiguity in existing classification systems can induce uncertainty in the classification of new samples (Weigelt *et al.*, 2010; Mackay *et al.*, 2011).

The desire for levels of accuracy that can ultimately lead to clinical utility continues to drive the field to refine breast cancer subtypes (Shen *et al.*, 2013; Parker *et al.*, 2009; Sun *et al.*, 2017; Haibe-Kains *et al.*, 2012; Sun *et al.*, 2014) and to identify molecular subtypes in other cancers (Abeshouse *et al.*, 2015; Cancer Genome Atlas Network, 2014). The recent establishment of international cancer genome consortia (Cancer Genome Atlas Network, 2012; Abeshouse *et al.*, 2015; Cancer Genome Atlas Network, 2014; Curtis *et al.*, 2012) has generally overcome the sample size issue. In this paper, we focus mainly on developing methods to address the computational challenges associated with detecting cancer related genes and biologically meaningful subtypes using high-dimensional genomics data. Molecular subtyping can be formulated as a supervised-learning problem, that is, to use established tumor subtypes as class labels to perform feature selection and construct a model for the classification of new patients. The strategy, though computationally simple, may not enable us to identify novel subtypes. Consequently, most existing methods were developed within the unsupervised learning framework. Representative work includes SparseK (Witten and Tibshirani, 2010), iCluster (Shen *et al.*, 2009, 2013) and non-negative matrix factorization (Kormaksson *et al.*, 2012). The basic idea is to perform gene selection and clustering analysis simultaneously to detect compact tumor groups by optimizing a certain cost function. A major issue with the existing methods is that there is no guarantee that subtypes identified through *de novo* clustering analysis are biologically relevant. Presumably, genomics data records all ongoing biological processes in a cell or tissue, where multiple factors interact with each other in a complex and entangled manner. Tumor samples can be grouped based on factors that are not related to the actual disease (e.g., race and eye color). Another major limitation is that for computational considerations most existing methods perform data dimensionality reduction through linear transformation (e.g., feature weighting used in SparseK (Witten and Tibshirani, 2010)). Thus, they cannot adequately deal with complex non-linear data and extract pertinent information to detect subtypes residing in non-linear manifolds in a high-dimensional space. Note that the molecular subtyping analysis was initially performed on mRNA data (Sørlie *et al.*, 2001, 2003). Conceivably, the integration of the information from multi-source genomics data could refine established cancer subtypes. However, some existing methods do not scale well to handle high-dimensional data. For example, iCluster (Shen *et al.*, 2009, 2013) involves matrix inversion and thus can only process a few thousands of genes. A commonly used practice is to perform preprocessing and retain only the most variant genes (Curtis *et al.*, 2012). However, there is no guarantee that low-variant genes contain no information and the cut-offs used to select variant genes were usually set somehow arbitrarily.

In this paper, we propose a novel deep-learning based method, referred to as DeepType, that addresses all the aforementioned technical limitations. Due to the ability to learn good representation in problem solving, deep learning has recently achieved state-of-the-art performance in computer vision (LeCun *et al.*, 2015), pattern recognition (Parkhi *et al.*, 2015) and bioinformatics (Zheng *et al.*, in press). For our purpose, by leveraging the power of a multi-layer neural network for representation learning, we map raw genomics data into a space where clusters can be easily detected. To ensure the biological relevance of detected clusters, we incorporate prior biological knowledge to guide representation learning. We train the neural network by minimizing a unified objective function consisting of a classification loss, a clustering loss and a sparsity penalty. The training process can be easily performed by using a mini-batch gradient descent method. Thus, our method can handle large datasets with extremely high dimensionality. Although the idea of using deep learning for clustering is not new (see, e.g., Xie *et al.* (2016)), to the best of our knowledge, this work represents the *first* attempt to use deep learning to perform joint supervised and unsupervised learning for cancer subtype classification. A large-scale experiment was performed that demonstrated that DeepType significantly outperformed the existing approaches. The new approach provides a framework for the derivation of more accurate and robust molecular cancer subtypes by using increasingly complex, multi-source data.

2 Methods

In this section, we give a detailed description of the proposed method for cancer subtype identification. We also propose novel procedures for optimizing the associated objective function and estimating the hyper-parameters.

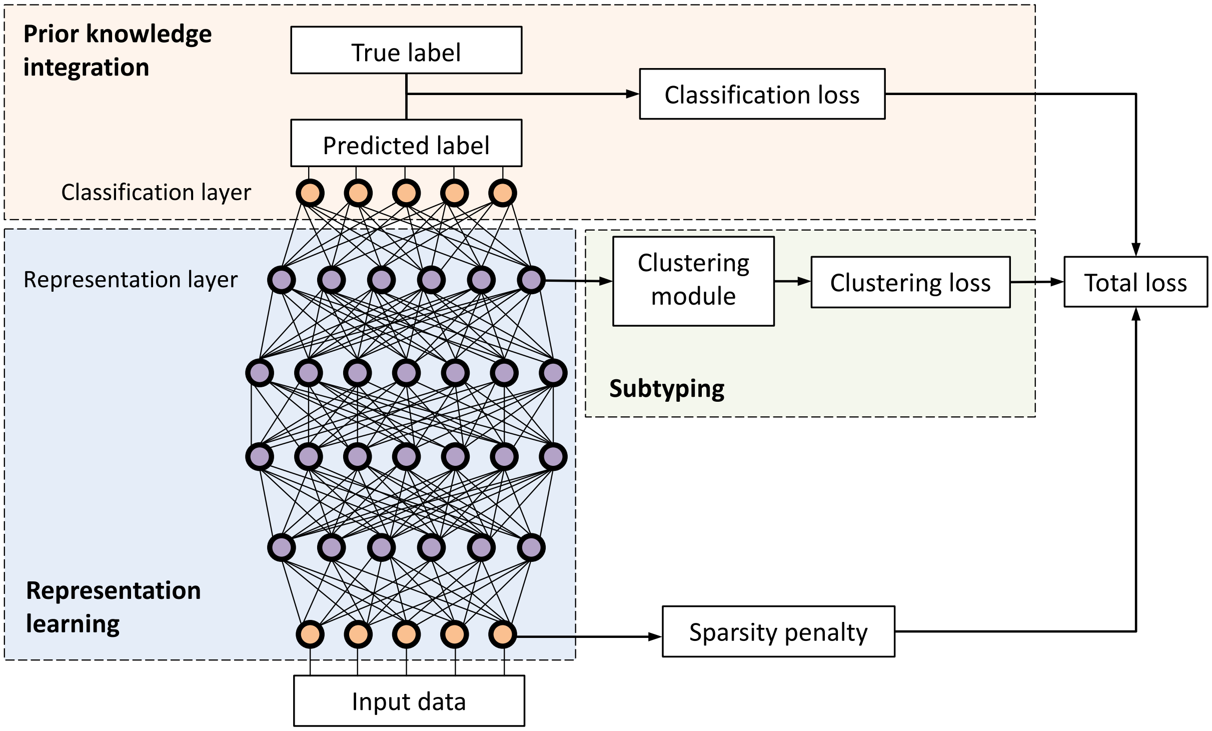


Figure 1: Overview of the proposed deep-learning based method for cancer molecular subtyping. It consists of three major components: representation learning, biological knowledge integration, and subtyping. The first part maps raw genomics data onto a representation space, the second part introduces prior biological knowledge to guide representation learning, and the third part generates subtyping results. The network parameters are learned by minimizing a unified objective function consisting of a classification loss, a clustering loss and a sparsity penalty.

2.1 Deep Learning for Cancer Subtype Identification

Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ be a cohort of tumor samples and $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]$ be a rough stratification of the samples (e.g., subtyping results from previous studies), where $\mathbf{x}_n \in \mathbb{R}^D$ is the n-th sample and $\mathbf{y}_n \in \mathbb{R}^J$ is the corresponding class label vector with $y_{jn} = 1$ if $\mathbf{x}_n$ belongs to the j-th group and 0 otherwise. Our goal is to identify a small set of cancer related genes and perform clustering analysis on the detected genes to refine existing classification systems and detect novel subtypes. To this end, we utilize the representation power of a multi-layer neural network to project raw data onto a representation space where clusters can be easily detected. As discussed above, clusters identified through unsupervised learning may not be biologically relevant. To address the issue, we impose an additional constraint that the detected clusters are concordance with previous results. Specifically, we cast it as a supervised-learning problem, that is, to find a representation space where the class labels can be accurately predicted.

Figure 1 depicts the network structure of the proposed method. It consists of an input layer, M hidden layers, a classification layer and a clustering module. The M-th hidden layer is designated as the representation layer, the output of which is fed into the classification layer and the clustering module. Mathematically, the neural network can be described as follows:

$$\begin{aligned}
\mathbf{o}_1 &= \mathrm{ReLU}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)\ , \\
\mathbf{o}_m &= \mathrm{ReLU}(\mathbf{W}_m\mathbf{o}_{m-1} + \mathbf{b}_m), 2 \le m \le M\ , \\
\bar{\mathbf{y}} &= \mathrm{softmax}(\mathbf{W}_{m+1}\mathbf{o}_M + \mathbf{b}_{m+1})\ ,
\end{aligned} \tag{1}$$

where $\mathbf{W}_m$, $\mathbf{b}_m$, and $\mathbf{o}_m$ are the weight matrix, bias term and output of the m-th layer, respectively, and $\bar{\mathbf{y}}$ is the output of the classification layer. For the purpose of this study, we use the rectified linear unit (ReLU) (Nair and Hinton, 2010) and softmax as the activation functions for the hidden and classification layers, respectively. For notational convenience, let $\boldsymbol{\Theta} = \{(\mathbf{W}_m, \mathbf{b}_m)\}_{m=1}^{M}$ and denote $f(\mathbf{x}|\boldsymbol{\Theta}) : \mathbb{R}^D \to \mathbb{R}^{D_M}$ as the mapping function that projects raw data onto a representation space, where D_M is the number of the nodes in the representation layer and $D_M << D$.

We optimize network parameters $\boldsymbol{\Theta}$ through joint supervised and unsupervised learning by minimizing an objective function that consists of a classification loss, a clustering loss and a regularization term. The classification loss measures the discrepancy between the predicted and given class labels. By construction, the j-th element of $\bar{\mathbf{y}}_n$ can be interpreted as the probability of $\mathbf{x}_n$ belonging to the j-th group. Thus, we use the cross entropy to quantify the classification loss:

$$L_{\text{classification}} = -\sum_{n=1}^{N}\sum_{j=1}^{J} y_{jn}\log \bar{y}_{jn}\ . \tag{2}$$

We use the K-means method (Lloyd, 1982) to detect clusters in the representation space. The clustering loss optimized by K-means is given by

$$L_{\text{clustering}} = \sum_{n=1}^{N} \|f(\mathbf{x}_n|\boldsymbol{\Theta}) - \mathbf{C}\mathbf{s}_n\|_2^2\ , \tag{3}$$

subject to $\sum_{k=1}^{K} s_{kn} = 1$, $s_{kn} \in \{0, 1\}$, $\forall k, \forall n$, where K is the number of clusters, $\mathbf{C}$ is a center matrix with each column representing a cluster center, and $\mathbf{s}_n$ is a binary vector where $s_{kn} = 1$ if $\mathbf{x}_n$ is assigned to cluster k and 0 otherwise.

Finally, we impose an $\ell_{2,1}$-norm regularization (Nie *et al.*, 2010) on the weight matrix of the first layer to control the model complexity and to select cancer related genes:

$$L_{\text{sparsity}} = \|\mathbf{W}_1^T\|_{2,1} = \sum_{j=1}^{D}\sqrt{\sum_{i=1}^{D_2} W_{1ij}^2}\ , \tag{4}$$

where W_{1ij} is the ij-th element of $\mathbf{W}_1$ and D_2 is the number of the nodes in the second layer. The $\ell_{2,1}$-norm regularization has an effect of automatically determining the number of nodes activated in the input layer, and thus the number of genes used in subtyping analysis.

Combining the above three losses, we obtain the following novel formulation for cancer subtype identification:

$$\begin{aligned}&\min_{\{\mathbf{\Theta},\mathbf{S},\mathbf{C}\}} \sum_{n=1}^{N} \|f(\mathbf{x}_n|\mathbf{\Theta}) - \mathbf{C}\mathbf{s}_n\|_2^2 + \lambda\|\mathbf{W}_1^T\|_{2,1} \\ &\text{subject to } -\sum_{n=1}^{N}\sum_{j=1}^{J} y_{jn}\log\bar{y}_{jn} \leq \zeta, \sum_{k=1}^{K} s_{kn} = 1,\ s_{kn} \in \{0,1\},\ \forall k, \forall n\ ,\end{aligned} \tag{5}$$

where $\mathbf{S} = [\mathbf{s}_1, \cdots, \mathbf{s}_N]$ and λ is a regularization parameter that controls the sparseness of a solution. The above formulation can be interpreted as finding a representation space to minimize the clustering loss while maintaining the classification loss smaller than a user defined upper bound ζ. For ease of optimization, we move the classification-loss constraint to the objective function and write the problem in the following equivalent form:

$$\begin{aligned}&\min_{\{\mathbf{\Theta},\mathbf{S},\mathbf{C}\}} -\sum_{n=1}^{N}\sum_{j=1}^{J} y_{jn}\log\bar{y}_{jn} + \alpha\sum_{n=1}^{N} \|f(\mathbf{x}_n|\mathbf{\Theta}) - \mathbf{C}\mathbf{s}_n\|_2^2 + \lambda\|\mathbf{W}_1^T\|_{2,1} \\ &\text{subject to } \sum_{k=1}^{K} s_{kn} = 1,\ s_{kn} \in \{0,1\},\ \forall k, \forall n\ ,\end{aligned} \tag{6}$$

where α is a tradeoff parameter that controls the balance between the classification and clustering performance. In the following sections, we describe how to solve the above optimization problem and estimate the hyper-parameters.

2.2 Optimization

The above optimization problem contains three sets of variables, namely, network parameters $\mathbf{\Theta}$, assignment matrix $\mathbf{S}$, and cluster centers $\mathbf{C}$. It is difficult to solve the problem directly since the parameters are coupled and $\mathbf{S}$ is a binary matrix. To address the issue, we partition the variables into two groups, i.e., $\mathbf{\Theta}$ and $(\mathbf{S}, \mathbf{C})$, and employ an alternating optimization strategy to solve the problem. Specifically, we first perform pre-training to initialize the network by ignoring the clustering module (i.e., setting $\alpha = 0$). Then, we fix $\mathbf{\Theta}$ and transform the problem into

$$\min_{\{\mathbf{C},\mathbf{S}\}} \sum_{n=1}^{N} \|f(\mathbf{x}_n|\mathbf{\Theta}) - \mathbf{C}\mathbf{s}_n\|_2^2,\ \text{subject to } \sum_{k=1}^{K} s_{kn} = 1,\ s_{kn} \in \{0,1\},\ \forall k, \forall n, \tag{7}$$

which can be readily solved by using the standard K-means method (Lloyd, 1982). Then, we fix $(\mathbf{S}, \mathbf{C})$ and write the problem as

$$\min_{\mathbf{\Theta}} -\sum_{n=1}^{N}\sum_{j=1}^{J} y_{jn}\log\bar{y}_{jn} + \alpha\sum_{n=1}^{N} \|f(\mathbf{x}_n|\mathbf{\Theta}) + \mathbf{C}\mathbf{s}_n\|_2^2 + \lambda\|\mathbf{W}_1^T\|_{2,1}\ , \tag{8}$$

which can be optimized through back-propagation by using the mini-batch based stochastic gradient descent method (Kingma and Ba, 2014). The above procedures iterate until convergence.

```
Algorithm 1: Hyper-parameter estimation ($\mathbf{X}$, $\mathbf{Y}$, $\mathcal{A}$, $\mathcal{L}$, $T$)
  Input: training data $\mathbf{X}$, class labels $\mathbf{Y}$, $T$, $\mathcal{A} = \{a_1, \cdots, a_J\}$, $\mathcal{L} = \{\lambda_1, \cdots, \lambda_L\}$
  Output: estimated parameters $\alpha^*$, $\lambda^*$, $K^*$
Randomly partition $(\mathbf{X}, \mathbf{Y})$ into ten folds $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{10}$;
Estimate $\lambda^*$ through ten-fold cross validation;
Compute average classification error $e_0$ and one standard error $\sigma_0$;
Estimate $\tilde{K}$ by maximizing average silhouette width;
for $i = 0$ to $T$ do
    $K_i = \tilde{K} + i$;
    for $j = 1$ to $J$ do
        $\alpha = a_j$;
        Solve Problem (6);
        Compute average classification error $e_j$;
        Compute average silhouette width $\tilde{s}_j$;
    end
    $j^* = \arg\max_{1 \le j \le J} j$, subject to $e_j \le e_0 + \sigma_0$;
    /* one-standard-error rule                                    */
    $\alpha_i = a_{j^*}$;
    $s_i = \tilde{s}_{j^*}$;
end
$i^* = \arg\max_{0 \le i \le T} s_i$;
$K^* = K_{i^*}$;
$\alpha^* = \alpha_{i^*}$
```

2.3 Parameter Estimation

We describe how to estimate the three hyper-parameters of the proposed method, namely regularization parameter λ, tradeoff parameter α, and number of clusters K. In order to avoid a computationally expensive three-dimensional grid search, we first ignore the clustering module by setting $\alpha = 0$ and perform supervised learning to estimate λ. The rationale is that previous subtyping results could provide us with sufficient information to determine the value of λ. Specifically, we randomly partition training data into ten equally-sized sub-datasets, perform ten-fold cross-validation and estimate λ by using the one-standard-error rule (Hastie *et al.*, 2009). Once we determine the value of λ, we perform K-means analysis on the outputs of the representation layer and pre-estimate the number of clusters, denoted as $\tilde{K}$, as the one that maximizes the average silhouette width (Wiwie *et al.*, 2015). Since the data representation is obtained through supervised learning, which tends to group samples with the same labels together, $\tilde{K}$ is likely to be the lower bound of the true value. Let $K_i = \tilde{K} + i, 0 \le i \le T$. For each K_i, we train a deep-learning model by using different α values and record the corresponding ten-fold cross-validation classification errors. By design, α controls the tradeoff between the classification and clustering performance, and the classification error increases with the increase of α. Again, by using the one-standard-error rule, for each K_i, we find the largest α, denoted as α_i, that results in a classification error that is within one standard deviation of the one obtained by setting $\alpha = 0$ (i.e., we require that the obtained classifier does not perform significantly worse than the existing subtyping system), and record the corresponding average silhouette width s_i. Once we run over all possible K_i, we obtain $T+1$ triplets $\{K_i, \alpha_i, s_i\}_{i=0}^{T}$. Finally, we determine the number of clusters K and the tradeoff parameter α as the pair that yields the largest average silhouette width. The pseudo-code of parameter estimation is given in Algorithm 1, and the proposed procedure performed quite well in our numerical experiment (See Figure 2).

3 Experiments

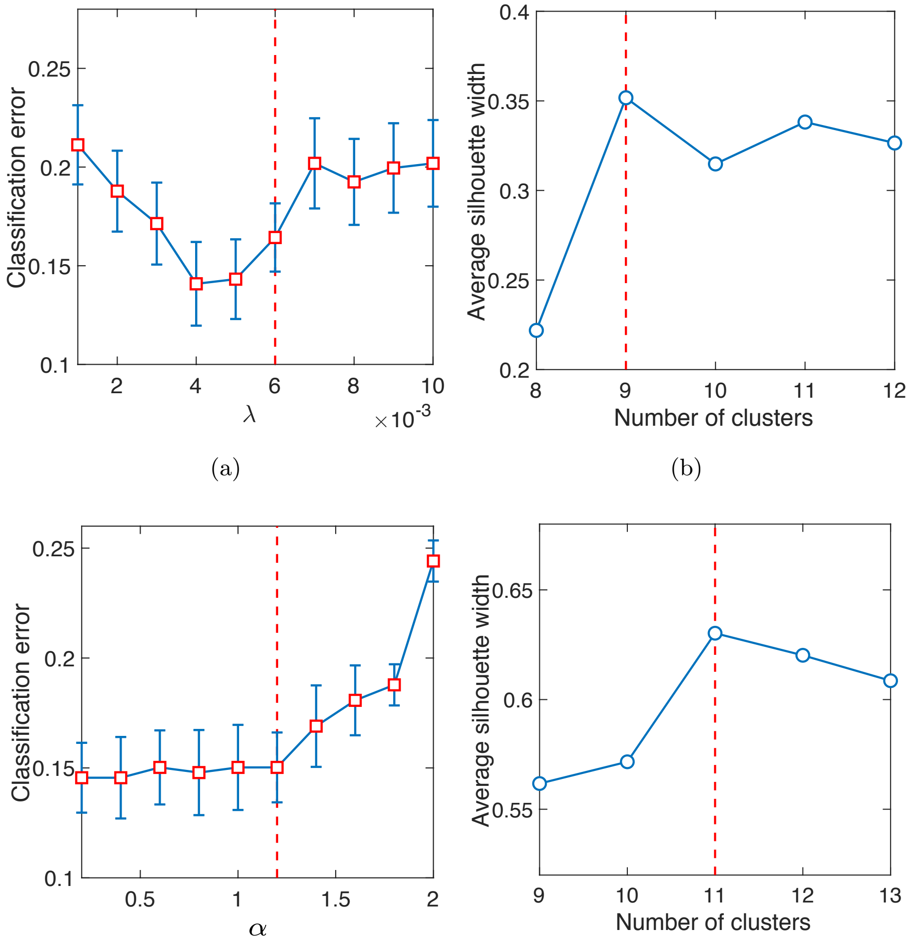
We conducted a large-scale experiment that demonstrated the effectiveness of the proposed method.

3.1 Experiment Setting

The experiment was performed on the data from the METABRIC project (Curtis *et al.*, 2012), which contains the expression profiles of 25,160 genes from 1,989 surgically excised primary breast tumor samples and 144 normal breast tissue samples. It is probably the largest single breast cancer dataset assayed to date. For computational convenience, we retained only the top 20,000 most variant genes for the downstream analysis. For model construction and performance evaluation, we randomly partitioned the data into a training and test datasets, each containing 80% and 20% of the samples, respectively. In this study, we used the PAM50 subtypes (Parker *et al.*, 2009) as class labels in the training process. We designed a four-layer neural network model for the joint supervised and unsupervised learning. The numbers of the nodes in the input layer, the two hidden layers and the output layer were set to 20,000, 1,024, 512, and 6, respectively. We employed the Adam method (Kingma and Ba, 2014) to tune the parameters of the model. The learning rate was set to 1e-3, the numbers of training epochs for model initialization and the joint supervised and unsupervised training were set to 300 and 1,500, respectively, and the batch size was set to 256. By using the method proposed in Section 2.3, the number of clusters K, the tradeoff parameter α and the regularization parameter λ were estimated to be 11, 1.2 and 0.006, respectively (Figure 2).

3.2 Clinically Relevant Subtypes Revealed by DeepType

By applying the proposed DeepType method to the breast cancer dataset, a total of 218 genes were selected and 11 clusters were detected. To visualize the identified clusters, we applied t-SNE (van der Maaten and Hinton, 2008) to the outputs of the representation layer. Figure 3 (a-b) present the sample distributions of the identified clusters and their PAM50 compositions, respectively. We can see that nearly all of the normal tissue samples were grouped into a single cluster (i.e., Cluster 0), and the tumor samples were grouped into ten well-separated clusters, labelled as DeepType 1-10. To demonstrate the clinical relevance of the identified tumor subtypes, a disease-specific survival data analysis was performed. Figure 3(c) shows that the ten subtypes



(a)

(b)

(c)

(d)

Figure 2: Hyper-parameter estimation. (a) The regularization parameter λ was estimated to be 0.006 based on the one-standard-error rule. (b) The number of clusters was pre-estimated to be 9 based on the average silhouette width. (c) We searched a range of values to estimate the number of clusters. For each $K \geq 9$, we trained a deep-learning model by using different α values and estimated the optimal α by using the one-standard-error rule. The figure presents an example showing that the optimal α was estimated to be 1.2 for $K = 11$. (d) The number of clusters was finally determined to be 11 by maximizing the average silhouette width. See Section 2.3 for a detailed description of hyper-parameter estimation.

are associated with distinct prognostic outcomes (logrank test, p-value $<$ 1.22e-19). Further internal and external validation analysis of the detected clusters is presented in Section 3.3.

Figure 3(d) presents the heatmap of the 218 selected genes, and the descriptions of the genes are given in Supplementary Table S1. The discovered subtypes contain distinct transcriptional characteristics associated with several gene co-expression modules and key cancer genes. Most normal-like samples were grouped into DeepType 1, and have an expression pattern similar to

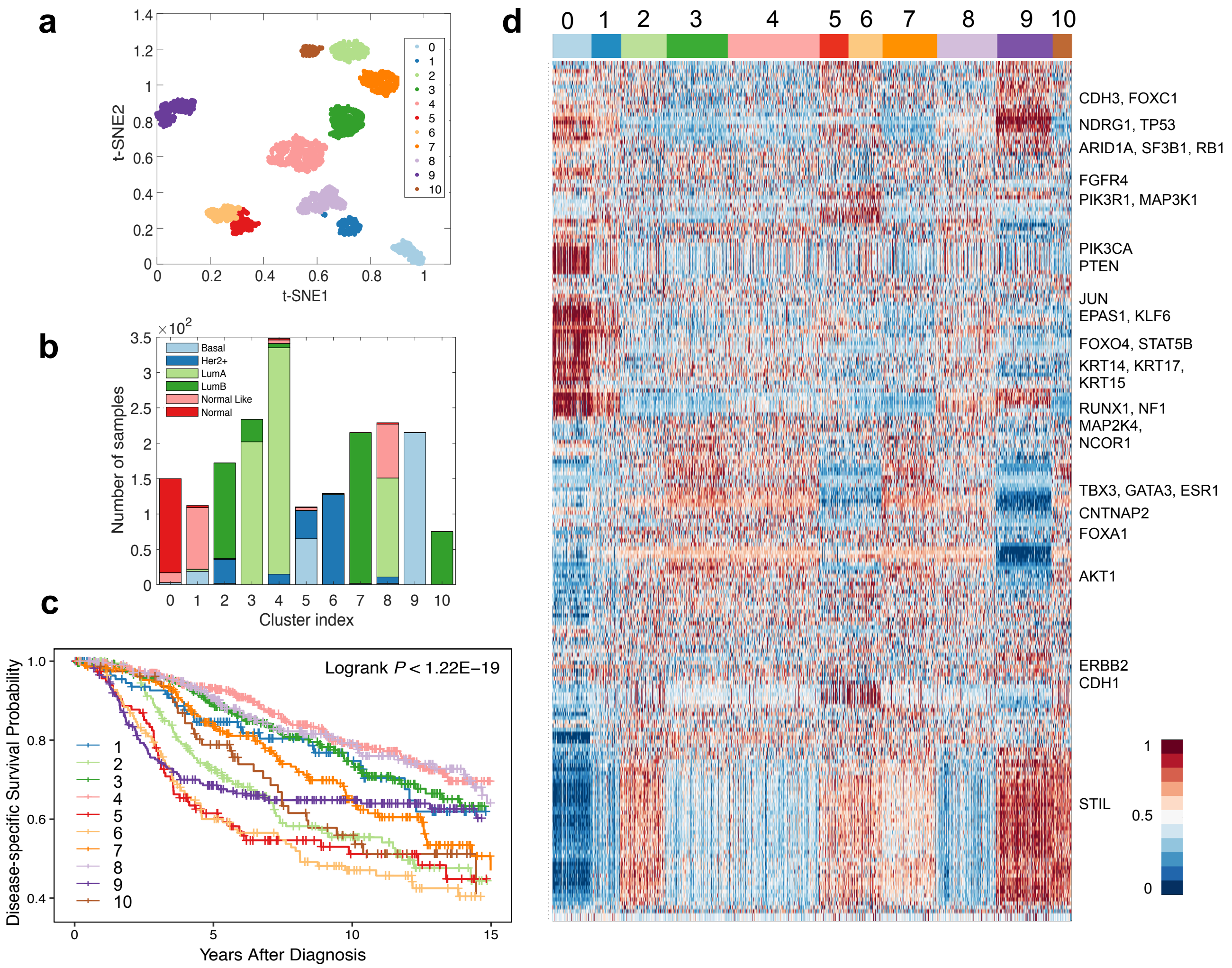


Figure 3: DeepType identified ten clinically relevant breast cancer subtypes. (a) The sample distributions of the identified clusters visualized by t-SNE. Nearly all of the normal tissue samples were grouped into a single cluster (i.e., Cluster 0), and the tumor samples were grouped into ten well-separated clusters, labelled as DeepType 1-10. (b) The PAM50 composition of the identified clusters. (c) The ten identified subtypes were associated with distinct clinical outcomes. (d) The heatmap of the 218 selected genes showed that the identified clusters exhibited distinct transcriptional characteristics on several gene modules. The samples were arranged by the clustering assignments, and the expression levels were linearly scaled into [0, 1] across samples.

the normal samples. The luminal A samples were separated as DeepTypes 3, 4 and 8 with low expression on the *STIL* module (key gene: *STIL*) and intermediate expression on the *GATA3* module (key genes: *TBX3*, *GATA3*, *ESR1*, *CNTNAP2* and *FOXA1*). Among the three subtypes, the expression of the *KRT* family (key genes: *KRT14*, *KRT15* and *KRT17*) were highest in DeepType 8, intermediate in DeepType 4 and lowest in DeepType 3. The luminal B samples

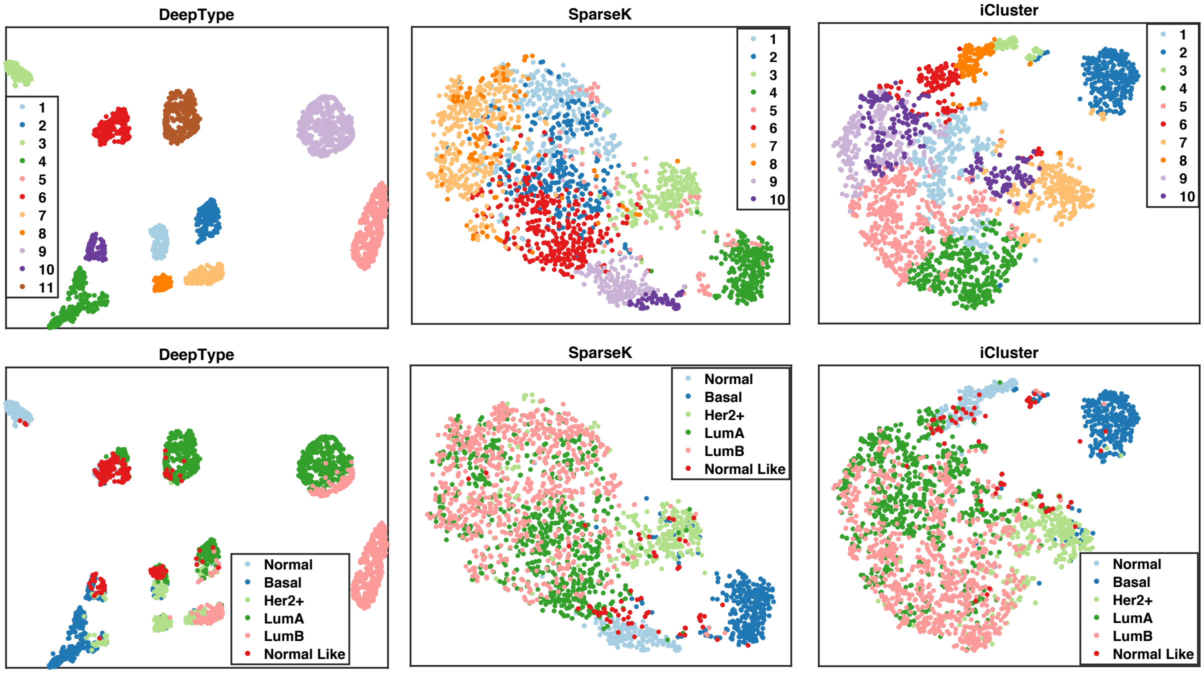


Figure 4: Visualization of the sample distributions of the clusters detected by three methods applied to data containing 10,000 most variant genes. Each sample was color-coded by its clustering assignment (top) and PAM50 label (bottom). DeepType revealed a clear eleven-cluster structure including a cluster comprising primarily normal tissue samples.

were grouped into DeepTypes 2, 7 and 10, with intermediate to high expression of the *GATA3* and *STIL* gene modules, and low expression of *CDH3* and *FOXC1*. Among the three subtypes, the expression of the genes in the *STIL* module were highest in DeepType 10, intermediate in DeepType 2 and lowest in DeepType 7. DeepTypes 5 and 6, which were dominated by HER2+ and mixed HER2+/basal samples, respectively, had very high expression on *ERBB2* and *CDH1* and low expression on *TBX3*, *GATA3* and *ESR1* genes. DeepType 9, composed entirely of basal samples, had low expression in the *GATA3* module and high expression in the *STIL* and *KRT* modules.

3.3 Comparison Study

To further demonstrate the effectiveness of the proposed method, we compared it with two state-of-the-art methods, namely SparseK (Witten and Tibshirani, 2010) and iCluster (Shen *et al.*, 2009). Both methods perform feature selection and clustering analysis simultaneously, and iCluster was also used in the METABRIC project (Curtis *et al.*, 2012). The source code of the two methods was downloaded from the CRAN website[1,2]. Following the procedure described in (Shen *et al.*, 2013), we tuned the parameters of iCluster (i.e., the number of clusters K and the sparsity

[1] https://cran.r-project.org/web/packages/iCluster/index.html

[2] https://cran.r-project.org/web/packages/sparcl/index.html

penalty coefficient λ) by maximizing the reproducibility index. SparseK also contains two parameters, the number of clusters K and the ℓ_1 regularization parameter λ. By using the method described in (Witten and Tibshirani, 2010; Tibshirani *et al.*, 2001), we first estimated the optimal λ for each K, and then determined the value of the optimal K based on gap statistic (Tibshirani *et al.*, 2001). To test the ability of the three methods to handle high-dimensional data, we generated four datasets each containing a different number of the most variant genes, ranging from 5,000, 10,000, 15,000 and 20,000. Although in this study, we considered only gene expression data, it is possible to perform cancer subtyping by integrating genomics data from different platforms, including copy number, methylation and mutational data. Therefore, the ability to handle high-dimensional data is an important consideration in algorithm development. For each dataset, we performed a series of quantitative and qualitative analyses to compare the performance of the three methods.

We first visualized the sample distributions of the clusters detected by the three methods (Figure 4). Since iCluster failed on the datasets with 15,000 and 20,000 genes due to the need of performing matrix inversion of high-dimensional data, we considered only the results generated by using the dataset with 10,000 genes. From the figure, we can see that DeepType identified eleven well-defined clusters, nearly all normal tissue samples were grouped into a single cluster, and the clusters that composed of tumor samples were well-separated and highly concordant with the PAM50 labels. In contrast, for SparseK and iCluster, the normal tissue samples were grouped into multiple clusters, which suggests that genes unrelated to cancer were selected. Moreover, the tumor samples with different PAM50 labels overlapped considerably, and did not exhibit a clear clustering structure.

We then performed a series of external and internal evaluations of the clusters detected by the three methods. For external evaluation, we assessed the concordance between the identified cancer subtypes and some widely used clinical and prognostic characteristics of breast cancer, including the PAM50 subtype (Parker *et al.*, 2009), histological grade, Nottingham prognostic index (NPI) (Haybittle *et al.*, 1982), gene expression grade index (GGI) (Sotiriou *et al.*, 2006) and the Oncotype DX prognostic test (Paik *et al.*, 2004) (See Supplementary Table S2 for a detailed description). Specifically, we used average purity and normalized mutual information (NMI) (Cover and Thomas, 2012) to evaluate the extent to which the identified subtypes matched the above described characteristics (Table 1). Our analysis showed that the subtypes identified by DeepType were highly concordant with the clinical variables and prognostic information. In all cases, the results generated by DeepType matched the PAM50 labels to the highest degree. This is expected since the PAM50 labels were used in training DeepType. Our method also produced the highest agreement with the histological grades, NPI and GGI. Notably, when compared with Oncotype DX, the average purities and NMI scores of DeepType were much higher than the other two methods. This is highly significant since while both NPI and GGI provide some values in predicting the clinical outcomes of breast cancer patients, Oncotype DX is the only test supported by level II evidence (Sparano *et al.*, 2018). We performed a Wilcoxon rank-sum test to compare the overall performance of DeepType and the two competing methods. The p-values are 7.7e-14 (DeepType vs. SparseK) and 1.3e-19 (DeepType vs. iCluster).

We next performed internal evaluation of the subtypes identified by the three methods. Inter-

Table 1: External evaluation of subtypes identified by three competing methods applied to datasets with a various number of input genes. iCluster failed on datasets with 15,000 and 20,000 genes. DeepType significantly outperformed SparseK (p-value $\leq$ 7.7e-14) and iCluster (p-value $\leq$ 1.3e-19, Wilcoxon rank-sum test).

		Average Purity				NMI			
		5000	10000	15000	20000	5000	10000	15000	20000
PAM50	DeepType	**0.86**	**0.80**	**0.85**	**0.87**	**0.62**	**0.56**	**0.62**	**0.61**
	SparseK	0.65	0.68	0.64	0.63	0.39	0.40	0.37	0.38
	iCluster	0.43	0.65	/	/	0.08	0.34	/	/
Histological Grade	DeepType	**0.67**	**0.67**	**0.66**	**0.67**	**0.12**	**0.12**	**0.12**	**0.13**
	SparseK	0.63	0.66	0.65	0.63	0.11	0.10	**0.12**	0.09
	iCluster	0.55	0.63	/	/	0.04	0.11	/	/
NPI	DeepType	**0.57**	0.55	**0.60**	**0.58**	**0.08**	**0.09**	**0.10**	**0.07**
	SparseK	0.56	0.55	0.59	0.58	0.07	**0.09**	0.07	**0.07**
	iCluster	0.56	**0.57**	/	/	0.04	**0.09**	/	/
GGI	DeepType	**0.69**	0.68	**0.70**	**0.69**	**0.15**	**0.16**	**0.14**	**0.13**
	SparseK	**0.69**	**0.70**	0.70	**0.69**	0.12	0.14	0.13	0.12
	iCluster	0.68	0.67	/	/	0.04	0.11	/	/
Oncotype DX	DeepType	**0.88**	**0.85**	**0.87**	**0.86**	**0.25**	**0.26**	**0.27**	**0.24**
	SparseK	0.75	0.78	0.78	0.76	0.14	0.16	0.15	0.13
	iCluster	0.64	0.74	/	/	0.06	0.13	/	/

Table 2: Internal evaluation of subtypes identified by three methods applied to datasets with a various number of input genes. The Davies–Bouldin index results in a value in $[0, \inf)$, and a smaller value suggests a better clustering scheme. DeepType significantly outperformed SparseK (p-value $\leq$ 7.8e-5) and iCluster (p-value $\leq$ 7.8e-5, Wilcoxon rank-sum test).

	Silhouette width			Davies-Bouldin index		
	DeepType	SparseK	iCluster	DeepType	SparseK	iCluster
5000	**0.48**	0.17	0.33	**1.01**	1.88	1.79
10000	**0.48**	0.22	0.33	**0.87**	1.94	1.23
15000	**0.44**	0.19	/	**0.69**	1.92	/
20000	**0.63**	0.15	/	**0.67**	2.31	/

Table 3: The numbers of genes selected by DeepType, iCluster and SparseK applied to datasets containing a various number of input genes.

# of input genes	DeepType	SparseK	iCluster
5000	**182**	949	521
10000	**239**	982	728
15000	**250**	918	/
20000	**218**	886	/

nal evaluation utilizes only the intrinsic information of cluster assignments to assess the quality of obtained clusters, and compactness and separability are the two most important considerations (Halkidi *et al.*, 2001). A compact and separable clustering structure means that samples in each cluster are homogeneous and different clusters are far away from each other, allowing new patients to be assigned with high certainty and low ambiguity. For the purpose of this study, we used the silhouette width (Wiwie *et al.*, 2015) and the Davies-Bouldin index (Davies and Bouldin, 1979) to quantify the cluster compactness and separability. The results are reported in Table 2. In all cases, DeepType resulted in the highest silhouette width and the lowest Davies-Bouldin index, which is consistent with the visualization result presented in Figure 4. To compare the overall performance, the Wilcoxon rank-sum test was performed. Deeptype significantly outperformed SparseK (p-value $\leq$ 7.8e-5) and iCluster (p-value $\leq$ 7.8e-5). Our analysis suggested that our method resulted in subtypes with significantly higher cluster quality than the competing methods.

Finally, we compared the ability of the three methods to select relevant genes from high-dimensional data for clustering analysis. Table 3 reports the numbers of genes selected by the three methods applied to the data with a various number of input genes. Notably, while DeepType achieved the best result in terms of both internal and external criteria, it selected the fewest genes in all cases. For clinical applications, the ability to select fewer genes can help to develop a more economic clinical assay for breast cancer subtype identification.

4 Conclusion

In this paper, we developed a deep-learning based approach that addresses some technical limitations of existing methods for cancer subtype identification. We demonstrated that the new method performed significantly better than two commonly used approaches in terms of both internal and external evaluation criteria. Although in our numerical study we considered only mRNA data, cross-platform data may provide more comprehensive information about cancer. Thus, as the future work, we will apply the proposed model to datasets of larger sample sizes from multiple genomic sources (mRNA, copy number, somatic mutation, methylation). It is expected that more accurate and robust cancer subtypes would be revealed.

References

Abeshouse, A., Ahn, J., Akbani, R., Ally, A., Amin, S., Andry, C. D., Annala, M., Aprikian, A., Armenia, J., Arora, A., *et al.* (2015). The molecular taxonomy of primary prostate cancer. *Cell*, **163**(4), 1011–1025.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418), 61–70.

Cancer Genome Atlas Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, **513**(7517), 202–209.

Cover, T. M. and Thomas, J. A. (2012). Elements of Information Theory. John Wiley & Sons.

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.* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**(7403), 346–352.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**(2), 224–227.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., and Sotiriou, C. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*, **104**(4), 311–325.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, **17**(2-3), 107–145.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, **144**(5), 646–674.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning. Springer, New York.

Haybittle, J., Blamey, R., Elston, C., Johnson, J., Doyle, P., Campbell, F., Nicholson, R., and Griffiths, K. (1982). A prognostic index in primary breast cancer. *British Journal of Cancer*, **45**(3), 361.

Kingma, D. P. and Ba, J. (2014). Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–13.

Kormaksson, M., Booth, J. G., Figueroa, M. E., and Melnick, A. (2012). Integrative model-based clustering of microarray methylation and expression data. *The Annals of Applied Statistics*, **6**(3), 1327–1347.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, **521**(7553), 436–444.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, **28**(2), 129–137.

Mackay, A., Weigelt, B., Grigoriadis, A., Kreike, B., Natrajan, R., A'hern, R., Tan, D. S., Dowsett, M., Ashworth, A., and Reis-Filho, J. S. (2011). Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. *Journal of the National Cancer Institute*, **103**(8), 662–673.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814.

Nie, F., Huang, H., Cai, X., and Ding, C. H. (2010). Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in Neural Information Processing Systems*, pages 1813–1821.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., *et al.* (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, **351**(27), 2817–2826.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., *et al.* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, **27**(8), 1160–1167.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *The British Machine Vision Conference*, page 6.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**(22), 2906–2912.

Shen, R., Wang, S., and Mo, Q. (2013). Sparse integrative clustering of multiple omics data sets. *The Annals of Applied Statistics*, **7**(1), 269 – 294.

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.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(19), 10869–10874.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., *et al.* (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(14), 8418–8423.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., *et al.* (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, **98**(4), 262–272.

Sparano, J. A., Gray, R. J., Makower, D. F., Pritchard, K. I., Albain, K. S., Hayes, D. F., Geyer Jr, C. E., Dees, E. C., Goetz, M. P., Olson Jr, J. A., *et al.* (2018). Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *New England Journal of Medicine*, **379**(2), 111–121.

Sun, Y., Yao, J., Nowak, N., and Goodison, S. (2014). Cancer progression modeling using static sample data. *Genome Biology*, **15**(8), 440.

Sun, Y., Yao, J., Yang, L., Chen, R., Nowak, N. J., and Goodison, S. (2017). Computational approach for deriving cancer progression roadmaps from static sample data. *Nucleic Acids Research*, **45**(9), e69.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(2), 411–423.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D. S., Dowsett, M., Ashworth, A., and Reis-Filho, J. S. (2010). Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *The Lancet Oncology*, **11**(4), 339–349.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, **105**(490), 713–726.

Wiwie, C., Baumbach, J., and Röttger, R. (2015). Comparing the performance of biomedical clustering methods. *Nature Methods*, **12**(11), 1033–1038.

Xie, J., Girshick, R., and Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pages 478–487.

Zheng, W., Yang, L., Genco, R. J., Wactawski-Wende, J., Buck, M., and Sun, Y. (in press). SENSE: Siamese neural network for sequence embedding and alignment-free comparison. *Bioinformatics*.

Supplementary Material

Table S1: DeepType identified 218 genes to be informative for breast cancer subtyping

See the attached Excel file

Table S2: Clinical and prognostic characteristics of breast cancer

Characteristics	Class label
PAM50 subtype	basal, HER2+, luminal A/B, normal-like
Histological grade	1, 2, 3
NPI	1, 2, 3, 4
GGI	low risk, high risk
OncotypeDX	low risk, intermediate risk, high risk