Bioinformatics, 36(5), 2020, 1476–1483
doi: 10.1093/bioinformatics/btz769
Advance Access Publication Date: 11 October 2019
Original Paper

Gene expression

Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data

Runpu Chen[1], Le Yang[1], Steve Goodison[2] and Yijun Sun[1,3,4,*]

[1]Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14214, USA, [2]Department of Health Sciences Research, Mayo Clinic, Jacksonville, FL 32224, USA, [3]Department of Microbiology and Immunology and [4]Department of Biostatistics, University at Buffalo, The State University of New York, Buffalo, NY 14214, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on January 16, 2019; revised on August 24, 2019; editorial decision on October 4, 2019; accepted on October 8, 2019

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 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 http://www.acsu.buffalo.edu/~yijunsun/lab/DeepType.html.

Contact: yijunsun@buffalo.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

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 toward personalized treatment regimes, cancers of specific tissues have been divided into subtypes based on the molecular profiles of primary tumors (Curtis *et al.*, 2012; Parker *et al.*, 2009; Sørlie *et al.*, 2001). The premise is that patients of the same molecular subtypes are likely to have similar disease etiology, responses to therapy and clinical outcomes. Thus, 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 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 (Parker *et al.*, 2009; Sørlie *et al.*, 2003). These early studies completely altered our views 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 subtyping systems induces uncertainty in the classification of new patients (Mackay *et al.*, 2011; Weigelt *et al.*, 2010).

The desire for levels of accuracy that can ultimately lead to clinical utility continues to drive the field to refine breast cancer subtypes (Haibe-Kains *et al.*, 2012; Parker *et al.*, 2009; Shen *et al.*, 2013; Sun *et al.*, 2014, 2017) and to identify molecular subtypes in other cancers (Abeshouse *et al.*, 2015). The recent establishment of international cancer genome consortia (Abeshouse *et al.*, 2015; Cancer Genome Atlas Network, 2012; Curtis *et al.*, 2012) has generally overcome the sample size issue. In this article, 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 gene selection and construct a model for the classification of new patients. However, as mentioned above, current subtyping systems provide only a rough stratification of cancer, and supervised learning-based approaches may not enable us to identify novel subtypes. This is because the primary goal of supervised learning is to identify genes to achieve the maximum separation of samples from different subtypes, and genes that support novel subtypes can be considered irrelevant and removed. 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). A major issue with existing methods is that there is no guarantee that subtypes identified through *de novo* clustering are biologically relevant. Presumably, genomics data record all ongoing biological processes in a cell or tissue, where multiple factors interact with each other in a complex and entangled manner. Therefore, tumor samples can be grouped based on factors that are not related to the actual disease (e.g. race and eye color). A possible way to address the issue is to use previously established results to guide the detection of new subtypes. However, as the name suggests, *de novo* clustering completely ignores results from previous efforts. 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. Finally, 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 pre-processing 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 are usually set somehow arbitrarily.

The above observations motivated us to develop a novel deep-learning-based approach, referred to as DeepType, that performs cancer subtyping through joint supervised and unsupervised learning but addresses their respective limitations. Due to the ability to learn good data representation, deep learning has recently achieved state-of-the-art performance in computer vision, pattern recognition and bioinformatics (LeCun *et al.*, 2015; Zheng *et al.*, 2019). 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 genomic data.

2 Materials and methods

In this section, we present 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.

2.1 Deep learning for cancer subtype identification

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ be a cohort of tumor samples and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ be a rough stratification of the samples (e.g. subtyping results from previous studies), where $\mathbf{x}_n \in \mathbb{R}^D$ is the nth 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 jth 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 multilayer 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 Mth 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:

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

where $\mathbf{W}_m$, $\mathbf{b}_m$ and $\mathbf{o}_m$ are the weight matrix, bias term and output of the mth layer, respectively, and $\bar{\mathbf{y}}$ is the output of the classification layer. For the purpose of this study, we use sigmoid and softmax as the activation functions for the hidden and classification layers, respectively. For notational convenience, let $\mathbf{\Theta} = \{(\mathbf{W}_m, \mathbf{b}_m)\}_{m=1}^{\mathrm{M}}$ and denote $f(\mathbf{x}|\mathbf{\Theta}) : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}^{\mathrm{D_M}}$ 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 $D_M \ll D$.

We optimize network parameters $\mathbf{\Theta}$ 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 jth element of $\bar{\mathbf{y}}_n$ can be interpreted as the probability of $\mathbf{x}_n$ belonging to the jth 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 loss function optimized by K-means is given by

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

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 ijth 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

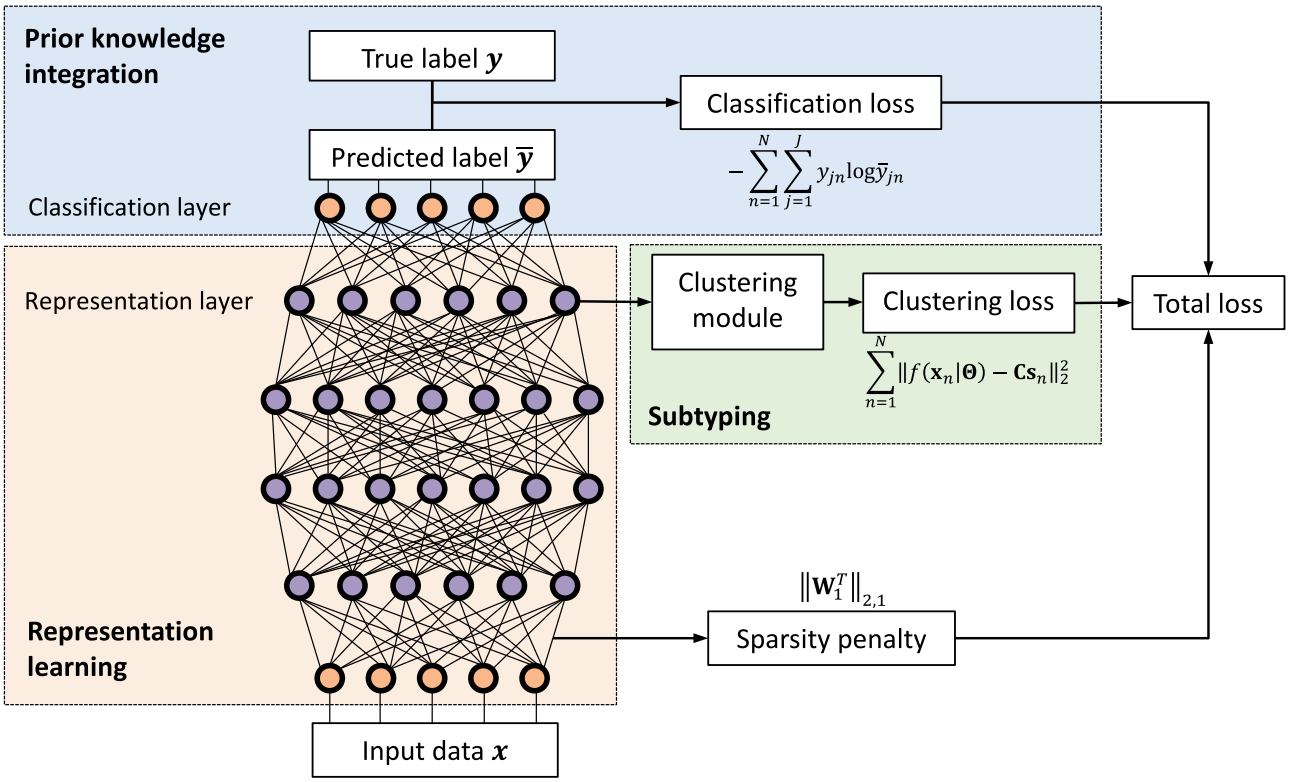


Fig. 1. Overview of the proposed deep-learning-based method for cancer molecular subtyping. It consists of three major components: representation learning, prior knowledge integration and subtyping. The first part maps raw genomics data onto a representation space, the second part incorporates prior 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

input layer, and thus the number of genes used in downstream subtyping analysis.

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

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

where $\mathbf{S}=[\mathbf{s}_1,\ldots,\mathbf{s}_N]$ and λ is a regularization parameter that controls the sparseness of weight matrix $\mathbf{W}_1$. 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:

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

where α is a trade-off 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

$$\min_{\{\mathbf{C},\mathbf{S}\}}\sum_{n=1}^{N}||f(\mathbf{x}_n|\Theta)-\mathbf{C}\mathbf{s}_n||_2^2,\tag{7}$$

subject to $\sum_{k=1}^{K}s_{kn}=1,\ s_{kn}\in\{0,1\},\ \forall k,\forall n$, which can be readily solved by using the standard K-means method. Then, we fix ($\mathbf{S}$, $\mathbf{C}$) and write the problem as

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

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.

2.3 Parameter estimation

We describe how to estimate the three hyper-parameters of the proposed method, namely regularization parameter λ, trade-off 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 10 equally sized sub-datasets, perform 10-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 , 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, is likely to be the lower bound of the true value. Let $K_i = \tilde{K} + i, 0 \leq i \leq T$. For each K_i, we train a deep-learning model by using different α values and record the corresponding 10-fold cross-validation classification errors. By design, α controls the trad-eoff 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 trade-off parameter α as the pair that yields the largest average silhouette width. The pseudo-code of the proposed procedure is given in Supplementary Algorithm S1, and the proposed procedure performed quite well in our numerical experiment (see Supplementary Fig. S1).

3 Results

We conducted a large-scale experiment on breast and bladder cancers to demonstrate the effectiveness of the proposed method. Due to space limit, here we report only the results of the breast cancer study and present the bladder cancer results in Supplementary Material.

3.1 Experiment setting

The breast cancer dataset was obtained from the METABRIC study (Curtis *et al.*, 2012), which contains the expression profiles of 25 160 genes from 1989 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, 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 for the joint supervised and unsupervised learning. The number of the nodes in the input layer, the two hidden layers and the output layer were set to 20 000, 1024, 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 number of training epochs for model initialization and the joint supervised and unsupervised training were set to 300 and 1500, respectively, and the batch size was set to 256. By using the method proposed in Section 2.3, the number of clusters K, the trade-off parameter α and the regularization parameter λ were estimated to be 11, 1.2 and 0.006, respectively (see Supplementary Fig. S1). To ensure that the constructed model did not overfit the data, we tracked the training and validation losses in the training process (see Supplementary Fig. S2) and no sign of over-fitting was observed.

3.2 Clinically relevant subtypes revealed by DeepType

By applying the proposed 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 2a and b presents 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 10 well-separated clusters, labeled as DeepType 1-10. To demonstrate the clinical relevance of the identified tumor subtypes, a disease-specific survival data analysis was performed. Figure 2c shows that the 10 subtypes were 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 Sections 3.3 and 3.4.

Figure 2d represents the heatmap of the 218 selected genes. The descriptions of the genes are given in Supplementary Table S1. The detected 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 normal samples. The luminal A samples were divided into 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 were partitioned 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 was highest in DeepType 10, intermediate in DeepType 2 and lowest in DeepType 7. DeepTypes 5 and 6, which were dominated by mixed HER2+/basal and HER2+ samples, respectively, had very high expression of *ERBB2* and *CDH1* and low expression of *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. The distinct expression patterns and prognostic outcomes of the detected clusters suggest that the proposed method is able to detect new breast cancer subtypes beyond the PAM50 classification, and a further analysis could reveal information on the breast cancer molecular taxonomy at a higher level of resolution.

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 study (Curtis *et al.*, 2012). The source code of the two methods was downloaded from the CRAN website: https://cran.r-project.org/web/packages/iCluster/index.html and https://cran.r-project.org/web/packages/sparcl/index.html. Following Shen *et al.* (2013), we tuned the parameters of iCluster (i.e. the number of clusters K and the sparsity 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), 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 5000, 10 000, 15 000 and 20 000. Although we herein considered only gene expression data, it is possible to perform cancer subtyping by integrating genomics data from different platforms. Therefore, the ability to handle high-dimensional data is an important consideration in algorithm development. Below, 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 (Fig. 3). 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. We can see that DeepType identified 11 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.

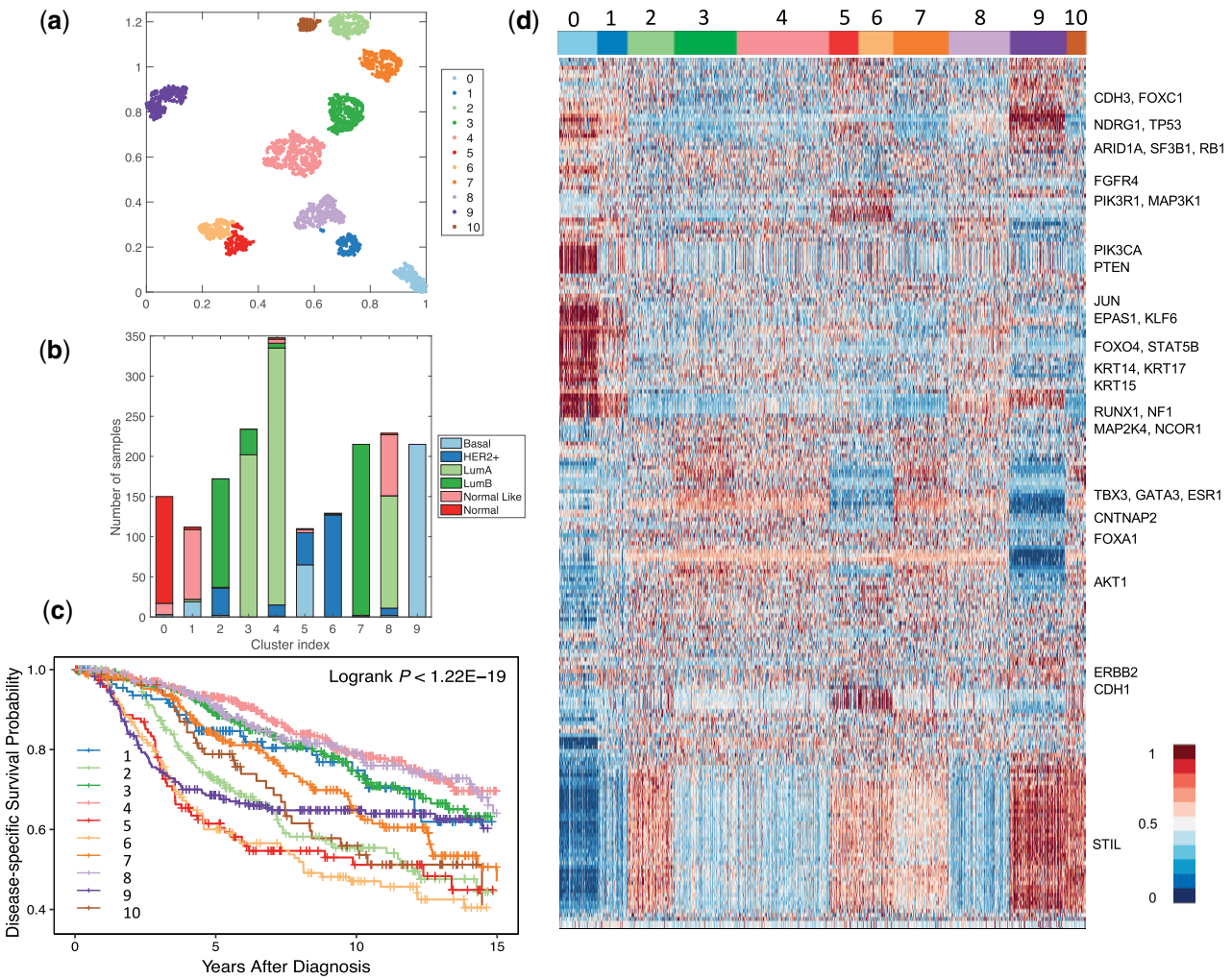


Fig. 2. DeepType identified 10 clinically relevant breast cancer subtypes. (**a**) The sample distributions of the identified clusters visualized by t-SNE. Nearly all of the normal tissue samples were grouped into a single cluster (i.e. Cluster 0), and the tumor samples were grouped into 10 well-separated clusters, labeled as DeepType 1-10. (**b**) The PAM50 composition of the identified clusters. (**c**) Survival data analysis showed that the 10 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

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 (Sparano *et al.*, 2018) (see Supplementary Table S2 for a detailed description). Specifically, we used average purity and normalized mutual information (NMI) to evaluate the extent to which the identified subtypes matched the above described characteristics. The results are reported in 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 versus SparseK) and 1.3e-19 (DeepType versus iCluster).

We next performed internal evaluation of the subtypes identified by the three methods. Internal 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 3. 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.

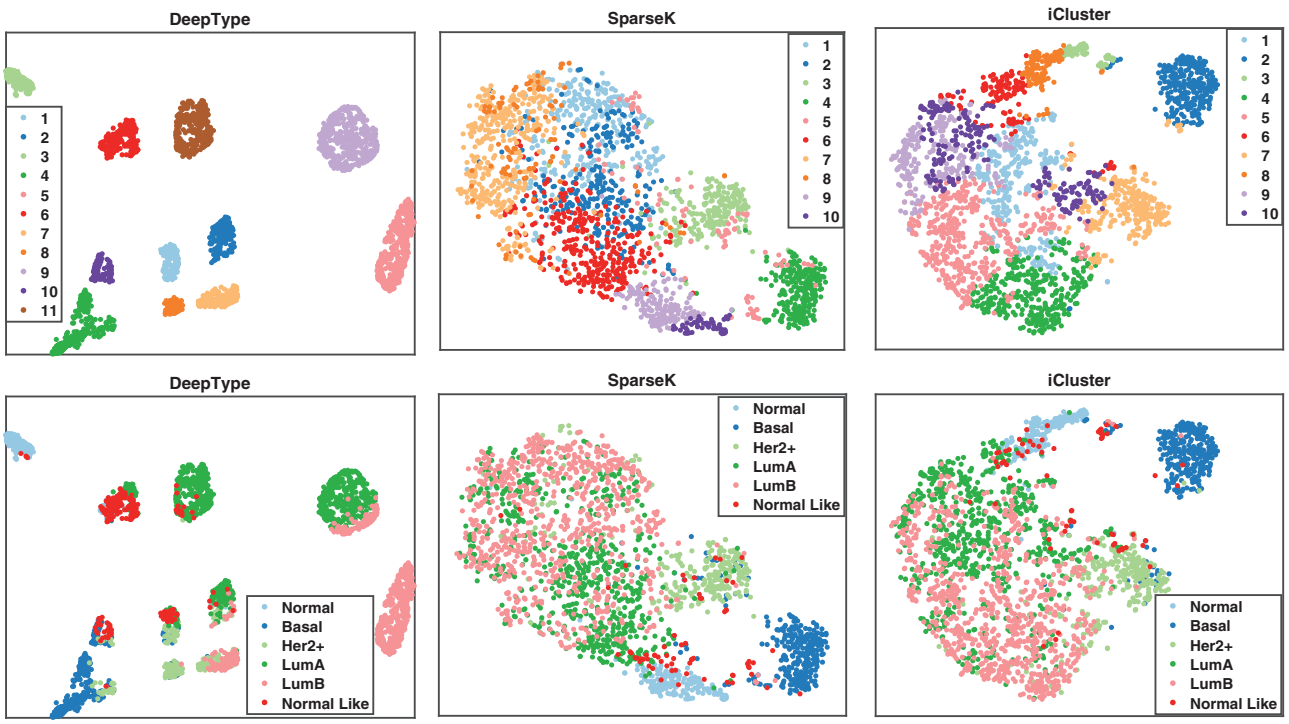


Fig. 3. Visualization of the sample distributions of the clusters detected by three methods applied to data containing 10 000 most variant genes. Each sample was color coded by its clustering assignment (top) and PAM50 label (bottom). DeepType revealed a clear 11-cluster structure including a cluster comprising primarily normal tissue samples

Table 1. External evaluation of subtypes identified by three methods applied to datasets with a various number of input genes

		Average purity				NMI			
		5000	10 000	15 000	20 000	5000	10 000	15 000	20 000
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	—	—
Tumor 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	—	—

Note: 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). The best results are boldfaced.

Finally, we compared the ability of the three methods to select relevant genes from high-dimensional data for clustering analysis. Table 3 reports the number 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.

3.4 Validation study

To demonstrate the generalization capability of the proposed method, we performed a validation study using the METABRIC data for training and SUPERTAM data (Haibe-Kains *et al.*, 2012) for testing. The SUPERTAM dataset contains the expression profiles of 13 092 genes from 856 breast tumor samples. Prior to the analysis, we identified 10 087 genes present in both datasets and used ComBat (Johnson *et al.*, 2007) to remove batch effects. Using the expression measures of the selected genes, we trained a deep-learning model using the METABRIC dataset and identified 11 clusters

Table 2. Internal evaluation of subtypes identified by three methods applied to datasets with a various number of input genes

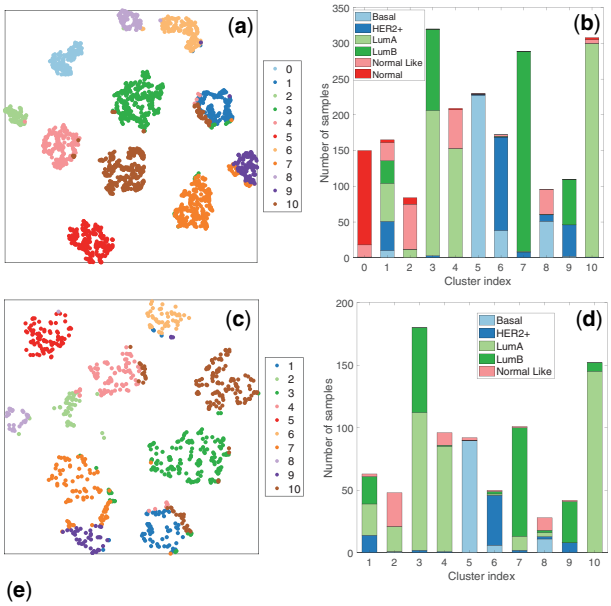
	Silhouette width			Davies–Bouldin index		
	DeepType	SparseK	iCluster	DeepType	SparseK	iCluster
5000	**0.48**	0.17	0.33	**1.01**	1.88	1.79
10 000	**0.48**	0.22	0.33	**0.87**	1.94	1.23
15 000	**0.44**	0.19	—	**0.69**	1.92	—
20 000	**0.63**	0.15	—	**0.67**	2.31	—

Note: The Davies–Bouldin index is 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). The best results are boldfaced.

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

No. of input genes	DeepType	SparseK	iCluster
5000	**182**	949	521
10 000	**239**	982	728
15 000	**250**	918	—
20 000	**218**	886	—

Note: The best results are boldfaced.



(e)

Cluster ID	1	2	3	4	5	6	7	8	9	10
IGP	0.95	0.96	0.98	0.95	1	1	0.95	0.96	0.92	0.98
p-value	0	0	0	0	0	0	0	0	0	0

Fig. 4. Results of a validation study performed on the METABRIC and SUPERTAM datasets. (**a–d**) The clusters detected in the METABRIC (top row) and SUPERTAM datasets (middle row) were compact and well separated and had very similar PAM50 compositions. (**e**) In-group proportion (IGP) scores and *P*-values (computed based on 1000 permutations) showed that the clusters identified in the METABRIC data were reproducible in the SUPERTAM data

including one comprising dominantly normal samples. We then applied the constructed model to the validation dataset and classified each sample into one of the 11 clusters using the nearest shrunken centroid classifier (Tibshirani *et al.*, 2002). Since the SUPERTAM data does not contain normal samples, only three samples were classified into the normal cluster and thus omitted in the further analysis. Figure 4a–d presents the sample distributions and PAM50 compositions of the identified clusters. We observed that the clusters detected in the two datasets were compact and well-separated and had similar PAM50 compositions. To provide a quantitative analysis of the reproducibility of the detected clusters, we employed the strategy proposed in Kapp and Tibshirani (2007) and calculated the in-group proportion (IGP) score and *P*-value for each cluster (Fig. 4e). Our analysis showed that the identified clusters were reproducible (*P*-value = 0) and that the proposed method generalizes well on independent datasets.

3.5 Robustness analysis

DeepType detects disease molecular subtypes through joint supervised and unsupervised learning, where the class labels from previous studies are usually error prone. To investigate how DeepType performs in the presence of label noise, we performed a robustness analysis where we corrupted the PAM50 labels of a certain percentage of randomly selected samples in the METABRIC training dataset, constructed a deep-learning model using the corrupted data, and applied the model to the test dataset. To assess the performance of the constructed model, we computed the Rand index by comparing the cluster assignments of the test samples with their PAM50 labels and those obtained by using the original training dataset (i.e. no corrupted labels). To remove random variations, the experiment was repeated five times. Supplementary Figure S3 presents the results obtained by using the training data containing a varying percentage of corrupted labels ranging from 0% to 20%. We can see that DeepType performed similarly with up to 10% label errors. Considering that the PAM50 label set itself contains an unknown percentage of errors, our method is very robust against label noise.

4 Discussion

In this article, we developed a deep-learning-based approach for cancer subtype identification that addresses some technical limitations of existing methods. The new method performed significantly better than two commonly used approaches in terms of both internal and external evaluation criteria. By leveraging the power of deep learning, the new method is able to handle data with extremely high dimensionality. We further demonstrated that the method generalizes well on independent datasets and is very robust against label noise.

The proposed method has several limitations. Usually, training a deep-learning model requires a large amount of data. The method is thus not applicable to cancers for which only a small number of samples have been assayed. However, it is expected that the sequencing cost will be significantly reduced in the near future and more tumor samples will be collected. In this study, we applied the method to breast and bladder cancers where molecular subtypes are well established and thus can be used to guide the detection of new subtypes. However, for many other cancers, molecular subtypes have not yet been well established. It is possible to use other clinical variables (e.g. tumor grade) to guide the identification of cancer subtypes and we have shown that our approach performed well in the presence of label noise. Further investigations are warranted to explore such possibilities.

In this article, we presented a proof-of-concept study considering only gene expression data. Several studies have recently demonstrated that combining cross-platform data could provide more information for cancer subtype identification [see, e.g. Shen *et al.* (2013) and Zhang *et al.* (2012)]. It is possible to use deep learning to integrate genomics data from different platforms, including mRNA, gene copy number, somatic DNA mutation and methylation, for cancer subtyping. However, there are ongoing debates about how to design a network to process multiple data types (Wang *et al.*, 2015). In future work, we will perform a large-scale experiment to look into this issue to identify the optimal network structure for genomics data analysis. It is expected that more accurate and robust cancer subtypes would be revealed.

Acknowledgements

We thank the editor and the three reviewers for their valuable comments that have helped us significantly improve the quality of the article. This work was, supported in part, by RO1AI125982 (Y.S.), RO1DE024523 (Y.S.) and RO1CA241123 (S.G.).

Conflict of Interest: none declared.

References

Abeshouse,A. *et al.* (2015) The molecular taxonomy of primary prostate cancer. *Cell*, **163**, 1011–1025.

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

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

Davies,D.L. and Bouldin,D.W. (1979) A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, **1**, 224–227.

Haibe-Kains,B. *et al.* (2012) A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst.*, **104**, 311–325.

Halkidi,M. *et al.* (2001) On clustering validation techniques. *J. Intell. Inform. Syst.*, **17**, 107–145.

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

Hastie,T. *et al.* (2009). *The Elements of Statistical Learning*. Springer, New York.

Haybittle,J. *et al.* (1982) A prognostic index in primary breast cancer. *Br. J. Cancer*, **45**, 361–366.

Johnson,W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Kapp,A.V. and Tibshirani,R. (2007) Are clusters found in one dataset present in another dataset? *Biostatistics*, **8**, 9–31.

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

Kormaksson,M. *et al*. (2012) Integrative model-based clustering of microarray methylation and expression data. *Ann. Appl. Statist*., **6**, 1327–1347.

LeCun,Y. *et al*. (2015) Deep learning. *Nature*, **521**, 436–444.

Lloyd,S. (1982) Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, **28**, 129–137.

Mackay,A. *et al*. (2011) Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. *J. Natl. Cancer Inst*., **103**, 662–673.

Nie,F. *et al*. (2010) Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In: *Advances in Neural Information Processing Systems, Vancouver, Canada,* pp. 1813–1821.

Parker,J.S. *et al*. (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol*., **27**, 1160–1167.

Shen,R. *et al*. (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**, 2906–2912.

Shen,R. *et al*. (2013) Sparse integrative clustering of multiple omics data sets. *Ann. Appl. Statist*., **7**, 269–294.

Sørlie,T. *et al*. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **98**, 10869–10874.

Sørlie,T. *et al*. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA*, **100**, 8418–8423.

Sotiriou,C. *et al*. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst*., **98**, 262–272.

Sparano,J.A. *et al*. (2018) Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N. Engl. J. Med*., **379**, 111–121.

Sun,Y. *et al*. (2014) Cancer progression modeling using static sample data. *Genome Biol*., **15**, 440.

Sun,Y. *et al*. (2017) Computational approach for deriving cancer progression roadmaps from static sample data. *Nucleic Acids Res*., **45**, e69.

Tibshirani,R. *et al*. (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Statist. Soc. Ser. B Statist. Methodol*., **63**, 411–423.

Tibshirani,R. *et al*. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA*, **99**, 6567–6572.

van der Maaten,L. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res*., **9**, 2579–2605.

Wang,W. *et al*. (2015) On deep multi-view representation learning. In: *International Conference on Machine Learning, Lille, France,* pp. 1083–1092.

Weigelt,B. *et al*. (2010) Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol*., **11**, 339–349.

Witten,D.M. and Tibshirani,R. (2010) A framework for feature selection in clustering. *J. Am. Statist. Assoc*., **105**, 713–726.

Wiwie,C. *et al*. (2015) Comparing the performance of biomedical clustering methods. *Nat. Methods*, **12**, 1033–1038.

Xie,J. *et al*. (2016) Unsupervised deep embedding for clustering analysis. In: *International Conference on Machine Learning, New York, USA,* pp. 478–487.

Zhang,S. *et al*. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*., **40**, 9379–9391.

Zheng,W. *et al*. (2019) SENSE: Siamese neural network for sequence embedding and alignment-free comparison. *Bioinformatics*, **35**, 1820–1828.