Deep Genomic Signature for early metastasis prediction in prostate cancer

Hossein Sharifi-Noghabi [1,3], Yang Liu [2], Nicholas Erho [2], Raunak Shrestha [3,4], Mohammed Alshalalfa [2], Elai Davicioni [2], Colin C. Collins [1,3,4], and Martin Ester [1,3,*]

[1] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada.
[2] GenomeDx Biosciences, Vancouver, BC, Canada.
[3] Vancouver Prostate Centre, Vancouver, BC, Canada.
[4] Department of Urologic Sciences, University of British Columbia, Vancouver, BC, Canada

[*] To whom correspondence should be addressed.

Abstract

Motivation: For Prostate Cancer (PCa) patients, timing and intensity of the therapy is adjusted based on their prognosis. This can be predicted from clinical/pathological information and, recently, gene expression signatures. One major challenge in developing such signatures is that all of them are based on cohorts which have limited number of patients with complete clinical outcomes (labelled), especially for slow progressing cancers such as PCa. This poses a challenge to the model development in conjunction with high dimensionality of the transcriptomic data.

Results: In this study we aim to exploit the previously untapped potential of a large cohort (n=15,136) with genomic data but no clinical outcome (unlabelled), to improve the performance of the genomic classifiers for predicting PCa metastasis. We propose, Deep Genomic Signature (DGS), based on Denoising Auto-Encoder (DAE) for feature extraction and selection. In order to capture information from the unlabelled and labelled data we train two DAEs separately and apply transfer learning to bridge the gap between them. We show that DGS captures information from these cohorts that can be utilized to build a logistic regression model to predict metastasis. Results on five validation cohorts indicate that this classifier, which is based on high weight genes in the DAEs, outperforms state-of-the-art signatures for metastatic PCa in terms of prediction accuracy. Survival analysis demonstrate the clinical utility of our signature which adds information to the well-established clinical factors and state-of-the-art signatures. Furthermore, pathway analysis reveals that the signature discovered by our DGS captures the hallmarks of PCa metastasis.

Availability of the implemented codes and supplementary materials:
https://github.com/hosseinshn/Deep-Genomic-Signature

Contact: ester@cs.sfu.ca

1 Introduction

Prostate Cancer (PCa) is the most prevalent cancer type among men, where roughly one in six men will be diagnosed in their lifetime (NIH, 2017). A majority of detectable PCa is entirely quiescent and can be successfully managed without intervention. However, a fraction of PCa is aggressive and responsible for the disease being ranked as the third leading cause of cancer death among men. This wide range of oncologic outcomes makes managing patients with PCa challenging and creates a need to stratify patients into clinically meaningful risk groups. Accurately determining a patient's risk of developing metastatic disease (a surrogate for lethal PCa (Xie *et al.*, 2017)) allows patients with more aggressive forms of PCa to be provided with more intense therapy and patients with relatively indolent cancer to avoid the serious side-effects of overtreatment.

Identification of clinically significant and aggressive PCa has been historically achieved through the interpretation of clinical and pathological risk factors, such as Prostate-Specific Antigen (PSA) blood concentration, Gleason Score (GS), and tumor stage. Recently it has been demonstrated that gene expression patterns within a patient's tumor can be used to identify higher risk PCa (Ramaswamy *et al.*, 2003; Penney *et al.*, 2011). Moreover, genomic signatures built with gene expression have been shown to add prognostic information over contemporary clinical and pathological variables (Erho *et al.*, 2013; Karnes *et al.*, 2017).

There are several genomic signatures for PCa. For example, Cuzick *et al.* (2011) developed Cell Cycle Proliferation score, based on 31 genes identified to be predictive of biochemical recurrence and prostate cancer specific mortality. Penney *et al.* (2011) developed a 157-gene signature by

comparing the gene expression profiles of low (≤ 6) GS and high (≥ 8) GS PCa. They used this signature to build a nearest shrunken centroid (Tibshirani *et al.*, 2002) classifier to build a parsimonious model for these two groups and a logistic regression model to predict the lethal cases among intermediate GS ($= 7$) tumors using this signature. Erho *et al.* (2013) identified a gene expression based signature to predict early metastasis following radical prostatectomy using a random forest. Moreover, it was recently demonstrated that gene expression signatures derived from the stroma or microenvironment of the primary tumor are predictive of the metastatic progression of PCa as well (Mo *et al.*, 2017; Tyekucheva *et al.*, 2017).

Because a long follow-up time is needed to determine if a patient has indolent PCa and high-throughput genomic profiling is costly, all published prognostic signatures were developed on relatively small datasets ($n \leq 545$). The large number of features (genes) relative to the small number of observations poses a great challenge to model development and potentially limits model performance. However, the rise of high-throughput, commercially available, whole-transcriptome, clinical testing, such as the Decipher test, is making large amounts of unlabelled tumor genomic data available without longitudinal outcomes (clinicaltrials.gov identifier: NCT02609269). We hypothesize that this large set of unlabelled, high dimensional genomic data can be exploited along with small well-annotated, labelled datasets from retrospective research studies to develop a genomic classifier that captures information that cannot be obtained from the smaller annotated datasets alone and ultimately outperforms existing prognostic models.

To utilize both the labelled and unlabelled datasets for training a genomic classifier, information needs to be captured from both the large unlabelled cohort and the small labelled one. Auto-Encoders (AE), a type of unsupervised Deep Neural Networks, can be used to reduce the dimensionality of the unlabelled cohorts, and capture the most salient features. These learned features are transferable and can be applied in different settings (Bengio, 2012; Bengio *et al.*, 2012). Information transferred from a source domain (e.g. a unlabelled cohort) to a target domain (e.g. a labelled cohort) can be fine-tuned if enough data is available in the target domain, or can remain frozen and untrainable if the target domain is rather small (Yosinski *et al.*, 2014). Because of their feature learning capabilities, AEs have been widely utilized in diverse biological problems. For instance, Denoising Auto-Encoders (DAE) has been used to obtain complex patterns from gene expression profiles in breast cancer (Tan *et al.*, 2015). Similarly, Stacked Denoising Auto-Encoders have been used to obtain functional features from breast cancer gene expression profiles (Danaee *et al.*, 2016). Therefore, using AE can be a viable solution to address our hypothesis. A more detailed review of AE and its strength is provided in Section 2.1.

We propose to utilize DAE to extract the most salient features from a large unlabelled cohort. These salient features are transferred in the form of weights, biases, and activation function to the first layer of another DAE, trained on a smaller labelled dataset. During training of the second DAE, the transferred parameters in the first layer are fixed and only parameters of the other layer are trained. Finally, a logistic regression model is built based on the informative genes selected from the second DAE via a standard deviation filter on the obtained weights of this DAE. Since this classifier has both l_1 and l_2 penalties, only a subset of the genes have non-zero coefficients. We propose these genes with non-zero coefficients as a novel signature for PCa metastasis.

Our contributions: To the best of our knowledge, this is the first study of deep transfer learning from a large unlabelled cohort without clinical outcomes to a small labelled cohort with clinical outcomes in the context of cancer research. We propose the Deep Genomic Signature (DGS) method based on two Denoising Auto-Encoders (DAEs), one for the labelled cohort and the other one for the unlabelled cohort. Learned parameters from the first DAE are transferred to the second DAE. High weight genes from the second DAE are used as features of a logistic regression classifier to predict metastasis. The final outcome of DGS is the set of features (genes) with non-zero coefficients which is used as a genomic signature for PCa metastasis prognosis. We applied DGS to six labelled datasets from retrospective research studies and one large unlabelled dataset obtained from the Decipher test of GenomeDx Biosciences Laboratory. We compared the accuracy of the signature discovered by DGS against that of the state-of-the-art signatures for PCa metastasis and showed that our signature significantly outperforms the existing signatures in terms of Area Under the receiver operating characteristic Curve (AUC). Further, we performed Uni/Multivariable Analyses and found that our signature adds novel prognostic information to well-established clinical risk factors and the state-of-the-art signatures (separately and combined). Finally, we performed Gene Set Enrichment Analysis (GSEA) and showed that our signature is highly relevant to metastasis and PCa biology. These experimental findings confirmed that our proposed DGS method succeeds in transferring relevant information from the unlabelled cohort to the labelled ones.

2 Materials and methods

the proposed DGS method consists of four steps: 1) Similar to the methods of (Danaee *et al.*, 2016; Tan *et al.*, 2015, 2016, 2017), a DAE is trained with a large unlabelled gene expression dataset to extract salient features. The output of this step is the learned parameters to be transferred to the next DAE. 2) in this step another DAE is trained with relatively small labelled dataset. This DAE has the transferred parameters from the previous step. Since it is very difficult to retrain those transferred parameters because of the small size of this labelled dataset, these transferred parameters are not trainable and stay fixed. The other parameters are trained. We adopted this method of transfer learning from (Yosinski *et al.*, 2014). Transferring parameters from the first DAE trained with the unlabelled data to the second DAE trained with the labelled data is the core of our DGS model. 3) After training the second DAE, genes with high weights in this DAE are selected based on a standard deviation filter on their connectivity weights. This selection approach is similar to (Danaee *et al.*, 2016; Tan *et al.*, 2015, 2016, 2017; Way and Greene, 2018). The idea of this gene selection step is to provide the classifier with a few rich genes with the strongest signals based on both the labelled and unlabelled datasets. 4)The selected genes from the previous step are used to train a classifier on the labelled dataset to predict metastasis. After training the classifier, those genes with non-zero coefficients are considered as the signature for PCa metastasis.

2.1 Background

In the past decade, Deep Learning proved to be significantly beneficial for unstructured problems such as audio/image processing, machine vision, and natural language processing (Goodfellow *et al.*, 2016). Deep Learning also has found its role in biological challenges such as regulatory genomics, cellular imaging (Angermueller *et al.*, 2016), Gene expression inference (Chen *et al.*, 2016), and sequence analysis (Alipanahi *et al.*, 2015; Zhou and Troyanskaya, 2015). The art of Deep Learning is to extract higher levels of abstraction or build more complex representations based on simpler ones. For example, in object recognition task first layers of the network only detect edges and contour-like shapes based on the output of the previous layers in a hierarchical order (Goodfellow *et al.*, 2016).

AEs are a type of the unsupervised deep networks which are designed to reduce the dimensionality of the data by learning the most salient features of it. An AE has two major components: an encoder which is capable of learning a latent space from the input data, and a decoder,

which can generate (reconstruct) an output vector similar to the input from the latent representation. Usually, an AE is constructed as a feed forward neural network and parameters e.g. connectivity weights between layers are optimized via backpropagation similar to the ordinary deep networks. The latent space usually has lower dimensionality than the input, so it captures most salient features from the input space. The ultimate goal of an AE is to generate an output identical to the input, but this task is required to take place first by obtaining and extracting informative features and then accurately decoding and reconstituting the input data from them rather than just learning an identity matrix and copying the input (Goodfellow *et al.*, 2016). There are various types of AE including DAE (Vincent *et al.*, 2010), Sparse Auto-Encoder (Ng, 2011), Contractive Auto-Encoder (Rifai *et al.*, 2011), Variational Auto-Encoder (VAE) (Kingma and Welling, 2013), and Adversarial Auto-Encoder (AAE) (Makhzani *et al.*, 2015).

Due to their feature learning capabilities, AEs have been widely utilized in diverse biological challenges. For instance, DAE was applied in the Analysis Using Denoising Auto-Encoders of Gene Expression (ADAGE) method on compendium of *Pseudomonas aeruginosa* gene expression profiling experiments to identify biological patterns (Tan *et al.*, 2016). Further, in order to enhance robustness and generate better signatures (more consistent with biological pathways), an ensemble version of ADAGE (eADAGE) was developed that integrated stable signatures across models (Tan *et al.*, 2017). Deep Patient, an unsupervised deep feature learning method based on Stacked DAE, was developed to predict future of patients with different cancers from Electronic Health Records data (Miotto *et al.*, 2016). VAE and AAE were also successfully utilized in designing new molecules with desired properties for drug discovery purposes (Gómez-Bombarelli *et al.*, 2016; Kadurin *et al.*, 2017). Moreover, VAE was also able to capture patterns in the gene expression pan-cancer data for specific tissues (Way and Greene, 2018).

Therefore, Deep Learning in general and AEs specifically, have been proved to be effective in extracting salient features, circumventing curse of dimentionality, and dealing with uncertainty in the input data without considering known biological knowledge.

We present DGS model in Section 2.2 and 2.3. Section 2.4 describes selection of high weight genes for classification. The datasets and statistical methods used for model training and evaluation are described in Section 2.5 and 2.6. Figure 1 illustrates the general idea of the DGS method.

2.2 Prospective Auto-Encoder

We denote the first DAE as the Prospective Auto-Encoder (PAE) because it is trained with the unlabelled data (prospectively collected). Since gene expression data is generally corrupted with noise, a DAE is a natural choice to deal with this issue and also it is a way to avoid over-fitting. PAE has three main components: 1) an encoder $\mathscr{F}(.)$, 2) a decoder $\mathscr{G}(.)$, and 3) a cost function $J(.)$. For M patients, N features (i.e. genes) and P nodes in the hidden layer: the encoder receives the input data $x \in \Re^N$, and encodes it to a hidden representation in lower dimensions $h \in \Re^P$ via a weight matrix $W \in \Re^{N \times P}$ and a bias vector $b \in \Re^{P \times 1}$:

$$\mathscr{F}_\theta(\overline{x})_{Pro} = \sigma_{Pro}(\overline{x}_{Pro} * W_{Pro} + b_{Pro}), \tag{1}$$

where $\overline{x} \in \Re^N$ is the corrupted version of input data x obtained from a corruption process like $\bar{x} = q(x)$, i.e., $\bar{x} = x * z$ and z is a binary variable from a binomial distribution. $\sigma(.)$ is the activation function like the Sigmoid or Relu functions. Decoder is another mapping function which maps back this hidden code or representation to the original input space:

$$\mathscr{G}_\theta(h)_{Pro} = \sigma_{Pro}(h_{Pro} * W_{Pro}^T + b_{Pro}). \tag{2}$$

Finally, the following cost function measures the discrepancy between the reconstructed version and the original input:

$$\mathcal{J}(\theta)_{Pro} = \sum_t {}_{q(x_{Pro}^{(t)})} \left[\mathcal{L}_{\text{MSE}}(x_{Pro}^{(t)}, \mathscr{G}_\theta(\mathscr{F}_\theta(\overline{x}_{Pro}^{(t)}))) \right], \tag{3}$$

where, "*Pro*" indicates the unlabelled cohort indexing, $x^{(t)}$ denotes the input vector for sample t, ${}_{q(x^{(t)})}$ denotes the expectation over the corrupted observation $\overline{x}$ via the corruption process q, and $\mathcal{L}_{\text{MSE}}(.)$ is the Mean Squared Error loss function. PAE is Denoising, which means the expression values for some randomly selected patients are changed to zero, therefore, PAE has to learn how to deal with this intentionally added noise in the corruption process as well. $(.)^T$ is the transpose operator because we used tied weights. Therefore, the set of parameters of the PAE to be optimized is $\theta_{Pro} = \{W_{Pro}, b_{Pro}\}$.

2.3 Retrospective Auto-Encoder

We denote the second DAE as the Retrospective Auto-Encoder (RAE) because it is trained with the labelled data (retrospective studies). RAE has two main differences comparing to the previous one, the first difference is the structure because it has two hidden layers. Second, parameters of the first layer i.e., weights and bias are not initialized randomly, these parameters are transferred along with the activation function from the PAE. These parameters are not trainable and stay frozen during training of the RAE with the labelled data because number of samples in the labelled data is not sufficient to train both of the layers. Therefore, the first layer stays fixed, and only parameters of the second layer are trained:

$$\mathscr{F}_\theta(\overline{x})_{Ret} = \sigma_{Ret}(\sigma_{Pro}(\overline{x}_{Ret} * W_{Pro} + b_{Pro}) * W_{Ret} + b_{Ret}), \tag{4}$$

$$\mathscr{G}_\theta(h)_{Ret} = \sigma_{Pro}(\sigma_{Ret}(h_{Ret} * W_{Ret}^T + b_{Ret}) * W_{Pro} + b_{Pro}), \tag{5}$$

$$\mathcal{J}_{Ret}(\theta) = \sum_t {}_{q(x_{Ret}^{(t)})} \left[\mathcal{L}_{\text{MSE}}(x_{Ret}^{(t)}, \mathscr{G}_\theta(\mathscr{F}_\theta(\overline{x}_{Ret}^{(t)}))) \right], \tag{6}$$

where, descriptions of symbols are the same as the previous one, and "*Ret*" denotes labelled cohort indexing. The set of parameters for the RAE is $\theta_{Ret} = \{W_{Ret}, b_{Ret}, W_{Pro},\ b_{Pro}\}$.

All codes for the PAE and the RAE are implemented by TensorFlow framework in Python 2.

2.4 Gene Selection and Building Classifier

A gene is considered to be a Deep High Weight (DHW) gene, if the value of its connectivity weight falls into the tails of the distribution of the input weights of a node in the second hidden layer. This selection is based on a standard deviation filter on those weights and regardless of their sign. After training the RAE with the labelled data, we have two weight matrices. The first weight matrix is obtained from training the PAE with the unlabelled data and the second one is obtained by training the RAE with the labelled data, here we denote them $W_{Pro} \in \Re^{N \times P}$ and $W_{Ret} \in \Re^{P \times Q}$ for learned weights by PAE and RAE respectively. Q is the number of nodes in the second hidden layer of the RAE. We define $W_{Final} \in \Re^{N \times Q}$, as a result of matrix product as follows:

$$W_{Final} = W_{Pro} * W_{Ret}, \tag{7}$$

In W_{Final} each gene has a vector of size Q which shows how strongly it is connecting/contributing to a node in the hidden layer. The standard deviation filter is applied on this matrix. In another words, only those genes that are in certain distance from the mean of the distribution of nodes in the hidden layer are considered to be DHW. Some of the DHW genes might be irrelevant to metastasis because their selection is unsupervised.

After selection of DHW genes, they will be the input of a regularized logistic regression in order to build the classifier to predict the desired phenotype i.e. metastasis. Signature obtained by our DGS method is a set of genes with non-zero coefficients in this classifier. Training and parameter tuning of this model is implemented in R 3.4.

2.5 Datasets

A total of 16,838 de-identified and anonymized radical prostatectomy tumor expression profiles using the Human Exon 1.0 ST microarray (ThermoFisher, Carlsbad, CA) were retrieved from the Decipher GRID. This included expression data from six published retrospective cohorts (n=1,702) with complete treatment and long-term outcomes data (we referred to them as the labelled datasets) and 1 prospective cohort (n = 15,136) with baseline pathological information only (we referred to it as the unlabelled dataset). labelled cohorts are as follows: Mayo I and II (GSE46691 and GSE62116) (Nakagawa *et al.*, 2008; Erho *et al.*, 2013; Karnes *et al.*, 2017); CCF (GSE62667) (Prensner *et al.*, 2014; Klein *et al.*, 2015); JHMI (GSE79957 and GSE79956) (Ross *et al.*, 2016); TJU (GSE72291) (Den *et al.*, 2014); MetaSet (Spratt *et al.*, 2017)[1]. The characteristics of these labelled datasets are brought in Table S1 in the Supplementary Materials. The unlabelled cohort was from clinical use of the Decipher test (GenomeDx Biosciences Laboratory, San Diego, CA; clinicaltrials.gov identifier: NCT02609269). The microarrays were normalized using Single Channel Array Normalization (SCAN) algorithm and summarized by gene annotation (Piccolo *et al.*, 2012). The normalized expression data had 46,050 features which we reduced to 13,856 via a standard deviation filter (sd > 0.11) to remove features with low levels of detection and signal.

2.6 Clinical and Biological Evaluation

We evaluate performance of DGS against five competitor methods: (Cuzick *et al.*, 2011), (Penney *et al.*, 2011), (Erho *et al.*, 2013), (Mo *et al.*, 2017), and (Tyekucheva *et al.*, 2017). Area Under the receiver operating characteristic Curve (AUC), with the Delong method (DeLong *et al.*, 1988; Robin *et al.*, 2011) for confidence interval calculation, was used to compare the genomic signature's ability to predict metastasis in five validation datasets. We further assess the clinical utility of our signature with survival analyses on a large multi-institutional cohort (MetaSet). Similarly, both univariable and multivariable (UVA and MVA) Cox proportional hazard models with mixed effects (CoxPHME) (Therneau *et al.*, 2003) are used to assess the association of the signature obtained by the DGS method with the time to metastasis. Here the mixed effect models are chosen to account for variation of underlying hazard functions and referral patterns across different institutions.

In order to demonstrate the independent prognostic power for the obtained signature, we adjusted for the common clinical variables that have shown to be prognostic in PCa in the MVA. In addition, we run multivariable CoxPHME with the signature obtained by DGS adjusting for one or more competitor methods. We divide the prediction scores from all five competitor models[2] as well as our DGS by 0.1, such that the reported Hazard Ratio (HR) should be interpreted for each 10% increase in the predictor score.

We studied co-expression network of the entire DHW genes to find correlated modules and their associations to PCa. We calculated correlation among the DHW genes in the unlabelled dataset and selected interactions with greater than 0.5 correlation.

[1] A subset of patients in this cohort were included in Mayo II, JHMI, TJU cohorts.

[2] prediction scores of Penney *et al.* (2011) and Cuzick *et al.* (2011) are scaled to $(0, 1)$ first.

To investigate the pathway and functional enrichment in those genes that had non-zero coefficients in the classifier, we performed a hypergeometric test based Gene Set Enrichment Analysis (GSEA) (https://github.com/raunakms/GSEAFisher). Gene sets of pathways were obtained from Molecular Signature Database v6.0 (Liberzon *et al.*, 2015). A cut-off threshold of false discovery rate < 0.01 was used to obtain the significantly enriched pathways. We selected only gene-sets/pathways with at least three genes enriched from our query list. Analysis for biological and clinical evaluation of this signature is performed in RStudio 1.0.143.

3 Results

In this section, first experiments for tuning the hyper-parameters of PAE, RAE, and the classifier are described. Then, we compared our method to the state-of-the-art signatures for PCa and interpreted the results. Finally, biological significance of the signature obtained by DGS method is discussed. To train the PAE, the RAE, and the classifier, we used the unlabelled and the Mayo I (one of the labelled datasets) cohorts. The PAE was trained using the unlabelled cohort, while the RAE and the final classifier were trained using the Mayo I cohort.

3.1 Experiments

3.1.1 PAE

In order to train the PAE, three randomly selected sample subsets of size 12000, 2570, and 566 were dedicated to train, development, and test sets, respectively. PAE was trained using the training set, validated on the development set, and tested with the unseen test set. For the PAE, we investigated diverse sets of hyper-parameters to determine the size of hidden layer, learning rate, batch size, corruption rate, number of epochs, and activation function. Assessing the cost function of the model, and minimize it using Adagrad optimization method (Duchi *et al.*, 2011), we selected the following values for the hyper-parameters used in the PAE: 256 nodes in the hidden layer (from a set of 256, 512, 1024, 2048, and 4096); learning rate of 0.05 (from a set of 0.05, 0.01, and 0.1); batch size of 100 (from a set of 10, 100, 500, and 1000); corruption level of 0.2 (from a set of 0.1, 0.2, 0.3 and 0.5); and number of epoch was set to be 1000 (from a set of 100, 200, 300, 500, and 1000). Finally, the Sigmoid activation function was used in the PAE from the set of Tanh, Relu, and Sigmoid activation functions. Obtained costs for train, development, and test for this DAE are 0.007, 0.008, and 0.008 respectively.

3.1.2 RAE

We used Mayo I cohort for training. A set of 359 randomly selected samples were dedicated for training, and the rest for development and testing (186 samples) similar to (Erho *et al.*, 2013). We selected the following values for hyper-parameters of the RAE after tuning: 64 nodes in the second hidden layer of the RAE (selected from a set of 32, 64, and 128); learning rate of 0.01 (from a set of 0.01, 0.05, and 0.1); batch size of 359 (from a set of 5, 10, 50, 100, and 359); number of epochs of 1000 (from a set of 300, 500, and 1000). Although we examined the same corruption rates as for the PAE, the input data without noise had the best performance. Activation function, cost function, and optimizer were also the same as the PAE. Train and test costs for the RAE are 0.031 and 0.036 respectively.

3.1.3 DHW genes

In order to select DHW genes, we set the threshold to be 3.94 which means only genes whose weights are 3.94 standard deviations smaller or greater than the mean (average value of all of the input weights of a node) are considered as DHW genes. In the previous studies the value of this threshold for similar purposes was typically set to 2 (Danaee *et al.*, 2016; Tan *et al.*, 2015, 2016) or 2.5 (Tan *et al.*, 2017; Way and Greene, 2018).

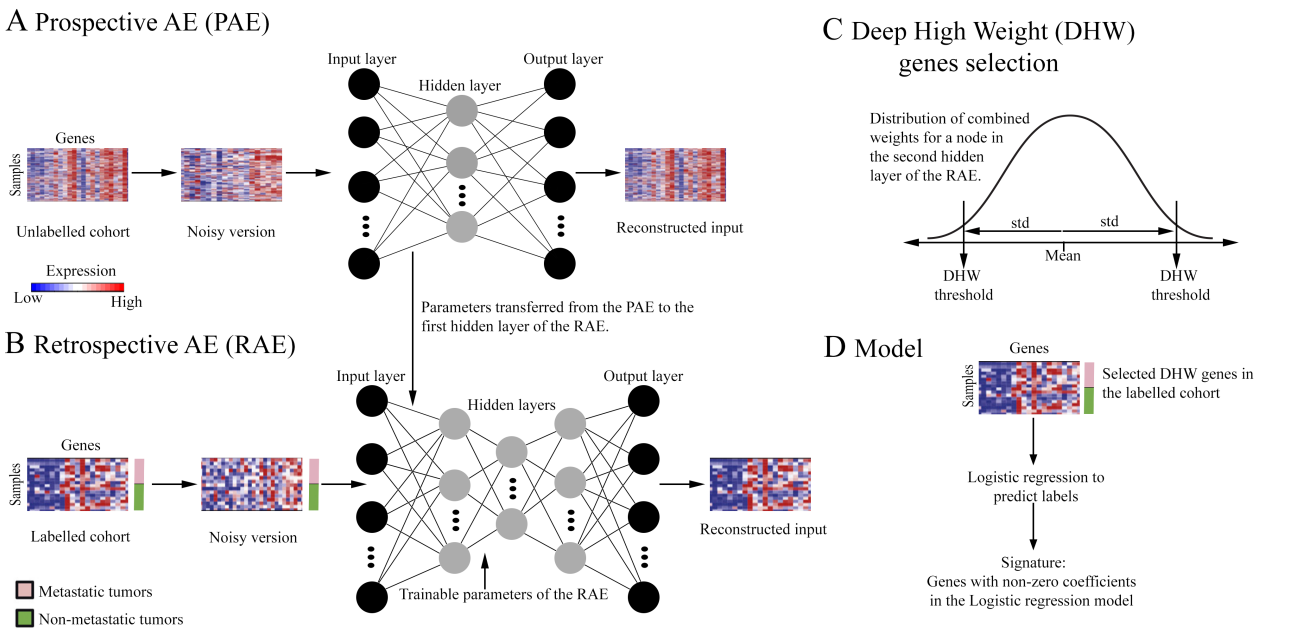

Fig. 1. Schematic overview of Deep Genomic Signature (A) Training the PAE with the unlabelled data to extract salient features from this cohort. PAE has one hidden layer. (B) Training the RAE with the labelled data. RAE has two hidden layers and parameters of the first hidden layer, which are transferred from the PAE, remain frozen and only parameters of the second hidden layer are trainable. (C) Deep High Weight (DHW) genes are selected based on the weights learned by the RAE (section 2.4). DHW genes are in the tails of the weight distribution of nodes in the second hidden layer of the RAE trained with the labelled data (D) DHW genes are used to train a logistic regression classifier with l_1 and l_2 penalties to predict metastasis. The signature is consist of all those genes with non-zero coefficients in this classifier.

However, due to possible different ranges of the weights or normalization, in our experiment these thresholds were too loose and resulted in too many features (genes) for a logistic regression classifier and a rather small training dataset. On the other hand, thresholds of 4 or 5 were too tight and resulted in a too small subset of genes. We decreased the value of the threshold by steps of 0.01 starting from 4 and observed the first jump in the number of DHW genes at 3.94 and set it as our threshold. It is important to note that this threshold was chosen without considering prior knowledge about the biology of the selected genes and without training separate classifiers for the considered thresholds. By applying this threshold, we obtained 141 DHW genes (Table S2 in the Supplementary Materials). We further used these genes to build a logistic regression model to predict metastasis.

3.1.4 Classifier

To study the performance of the selected DHW genes for the prediction of metastasis, a logistic regression model with l_1 and l_2 regularization was built based on Mayo I cohort (with same train and validation split). Results (training and validation AUCs) for training the classifier are available in Table S3. Two hyper-parameters γ, the complexity parameter ($\gamma > 0$), and α, the compromise between l_2 ($\alpha = 0$) and l_1 ($\alpha = 1$) penalties ($0 \leq \alpha \leq 1$), were optimized under cross validation ($\alpha = 0.6$ and $\gamma = 0.032$). 38 gene features had non-zero coefficients in the logistic regression model (Table S4). Scores from this model predict the probability of a PCa patient developing metastasis after surgery.

3.2 Clinical Validation

In order to study the clinical significance of DGS and highlight the unity of our method over other development methods, we assessed its ability to predict metastatic PCa when adjusted for established clinicopathologic risk factors and previously published genomic signatures.

3.2.1 DGS predicts PCa metastasis

In the blinded validation we found that our DGS method is highly significant for predicting metastatic disease across our five test datasets (CCF, Mayo II, JHMI, TJU, and MetaSet) with AUCs ranging from 0.67 to 0.83 (Table 1). Our signature outperformed all previously published signatures in four testing cohorts (CCF, Mayo II, TJU, and MetaSet). Erho *et al.* (2013)'s signature was found to be the most predictive classifier in JHMI. Likewise, Mo *et al.* (2017)'s signature tied DGS for highest AUC in Mayo II. Erho *et al.* (2013)'s signature had the second highest AUC in three datasets (Mayo II, TJU, and MetaSet). Penney *et al.* (2011)'s signature was the second most predictive classifier in JHMI and MetaSet (tied with Erho *et al.* (2013)). Mo *et al.* (2017)'s signature had the second best performance in one testing dataset (CCF). Penney *et al.* (2011)'s signature had the third highest AUC in three datasets (CCF, Mayo II, and TJU). Cuzick *et al.* (2011)'s signature had the third highest AUC in MetaSet cohort. Likewise, our signature was the third most predictive classifier in JHMI testing dataset.

3.2.2 DGS captures novel information in addition to clinicopatholgic risk factors alone

In order to evaluate DGS for clinical use, we preformed UVA and MVA survival analysis in the Metaset against clinical and pathologic risk factors (Table 2). UVA revealed that, except for surgical margins and PSA, DGS (HR 1.23, $p < 0.001$) along with all the clinicopatholgic variables were significantly associated with metastasis. Likewise, in MVA, when DGS is adjusted for the same clinicopathologic variables, DGS still remains significantly associated with metastasis (HR of 1.14, p-value<0.001). This indicates that DGS captures additional information for prognosis beyond these standard clinicopatholgic risk factors.

Table 1. Comparison of signatures based on AUC (with confidence interval) of prediction of metastasis on independent cohorts.

Study/Dataset	CCF	Mayo II	JHMI	TJU	MetaSet
(Penney *et al.*, 2011)	*0.80 (0.74-0.87)*	*0.71 (0.65-0.78)*	0.72 (0.65-0.78)	*0.68 (0.53-0.83)*	0.70 (0.65-0.76)
(Cuzick *et al.*, 2011)	0.73 (0.65-0.81)	0.64 (0.57-0.72)	0.63 (0.55-0.70)	0.67 (0.51-0.82)	*0.67 (0.60-0.73)*
(Erho *et al.*, 2013)	0.78 (0.70-0.85)	0.74 (0.67-0.81)	**0.73 (0.67-0.79)**	0.74 (0.61-0.87)	0.70 (0.64-0.76)
(Mo *et al.*, 2017)	0.81 (0.74-0.88)	**0.79 (0.73-0.85)**	0.58 (0.51-0.66)	0.59 (0.35-0.83)	0.65 (0.59-0.72)
(Tyekucheva *et al.*, 2017)	0.71 (0.62-0.79)	0.64 (0.56-0.71)	0.63 (0.56-0.70)	0.55 (0.39-0.72)	0.58 (0.51-0.64)
DGS	**0.83 (0.76-0.89)**	**0.79 (0.73-0.85)**	*0.67 (0.61-0.74)*	**0.79 (0.69-0.88)**	**0.73 (0.68-0.79)**

All methods are trained on Mayo I; **boldface** indicates the best method of that corresponding dataset; underline indicates the second best method(s) for that dataset, the third best method(s) are indicated in *italic* form.

Table 2. UVA and MVA Cox Model with random effects on the MetaSet.

Variable	UHR	P-value	AHR	P-value
Age	0.97 (0.94-0.99)	0.020*	0.96 (0.93-0.99)	0.007*
PSA	1.23 (0.99-1.52)	0.065	1.02 (0.80-1.29)	0.89
GS	4.73 (2.76-8.10)	<0.001*	3.31 (1.88-5.81)	<0.001
EPE	3.86 (2.23-6.68)	<0.001*	1.90 (1.02-3.52)	0.042*
SM	1.61 (0.99-2.62)	0.055	1.39 (0.86-2.26)	0.179
SVI	3.21 (2.06-5.00)	<0.001*	1.78 (1.10-2.90)	0.020*
LNI	4.65 (2.71-8.00)	<0.001*	2.19 (1.21-3.95)	0.009*
DGS	1.23 (1.14-1.32)	<0.001*	1.14 (1.05-1.22)	<0.001*

UHR: Unadjusted Hazard Ratio; AHR: Adjusted Hazard Ratio; *:statistically significant; PSA: $log2$ transformed PSA level; GS: pathological Gleason group (binarized by group 3, 4, and 5 against 1 and 2); EPE: extent of extracapsular extension; SM: surgical margin; SVI: semivacal invasion; LNI: lymph node invasion; .

Table 3. MVA adjusting other competitors as well as clinical variables on Meta Set.

Method	AHR	P-value
Erho *et al.* (2013)	1.21 (1.06-1.38)	0.006
DGS	1.09 (1.00-1.18)	0.039*
Penney *et al.* (2011)	1.42 (1.22-1.65)	<0.001*
DGS	1.14 (1.06-1.23)	<0.001*
Cuzick *et al.* (2011)	1.09 (0.95-1.25)	0.235
DGS	1.10 (1.01-1.20)	0.024*
Tyekucheva *et al.* (2017)	0.99 (0.89-1.10)	0.825
DGS	1.14 (1.05-1.23)	<0.001*
Mo *et al.* (2017)	1.04 (0.98-1.11)	0.178
DGS	1.12 (1.04-1.21)	0.004*

AHR: Adjusted Hazard Ratio; *: statistically significant.

3.2.3 DGS contains novel information not captured by other prognostic models

We also performed multivariable survival analysis of DGS with each of the previously published prognostic genomic signatures. When assessed against these five genomic signatures individually, DGS was found to be significant and hence containing independent information not captured by the previously published signatures (Table 3). In order to determine if DGS captured novel prognostic information beyond what was captured by the previously published signatures in combination, we performed a single MVA of DGS adjusting for all published genomic signatures (6 variables in total). Here we found that again DGS remained significant indicating it captures information that no other signature has captured (Table 4).

Table 4. DGS MVA analysis adjusting for 5 published genomic signatures.

Methods	AHR	P-value
Erho *et al.* (2013)	1.16 (1.01-1.33)	0.040*
Cuzick *et al.* (2011)	1.35 (1.16-1.57)	<0.001*
Penney *et al.* (2011)	1.16 (1.00-1.34)	0.048*
Mo *et al.* (2017)	1.04 (0.97-1.10)	0.261
Tyekucheva *et al.* (2017)	1.00 (0.91-1.11)	0.951
DGS	1.10 (1.01-1.20)	0.031*

AHR: Adjusted Hazard Ratio; (.): Confidence Interval; *: statistically significant

3.3 Biological Evaluation

In addition to providing evidence for the clinical significance of the signature obtained by the DGS method, we evaluated the DHW genes and the final signature to demonstrate that our method was able to select genes which drive PCa disease progression. we explored the co-expression network among the entire 141 DHW genes. Co-expression network analysis showed that these DHW genes fall into 6 highly correlated modules with distinct distribution and distinct biology (Figure 2- part A). The modules are capturing key molecular pathways involved in PCa progression and metastasis including AR-signaling, cell adhesion, ERG-fusion, and prostate specific antigen processing. ERG-fusion is the most common genomic event in PCa that occurs in about 40% of patients. Cell adhesion is a key molecular event that is associated with metastasis in several solid tumors including PCa; loss of expression of cell adhesion genes leads to increased potential of cell invasion and metastasis. This suggest that even DHW genes in DGS method represent the underlying biology of aggressive PCa.

We used GSEA to capture the signaling pathways dysregulated by the 38-gene signature with non-zero coefficients (Figure 2- part B). We obtained pathways such as Muscle Contraction (p = 7x10^{-8}) and Actin-Cytoskeleton Regulation (p = 6x10^{-5}) as our top-hits. These pathways are critical for the migration of metastatic initiating cells from the primary site of the tumor to the circulatory system invading the stroma (Yamazaki *et al.*, 2005; Pollard and Cooper, 2009). this 38-gene signature was also enriched for Androgen Response pathway (p = 4x10^{-5}) which is critical for PCa progression (Chen *et al.*, 2017). We also studied Kaplan-Meier (KM) survival analysis for the DGS model score. KM curve by tertiles in MetaSet (the independent large multi-institutional cohort) demonstrated that the DGS can successfully distinguish primary PCa tumors with metastatic potential from indolent ones (Figure 2- part C).

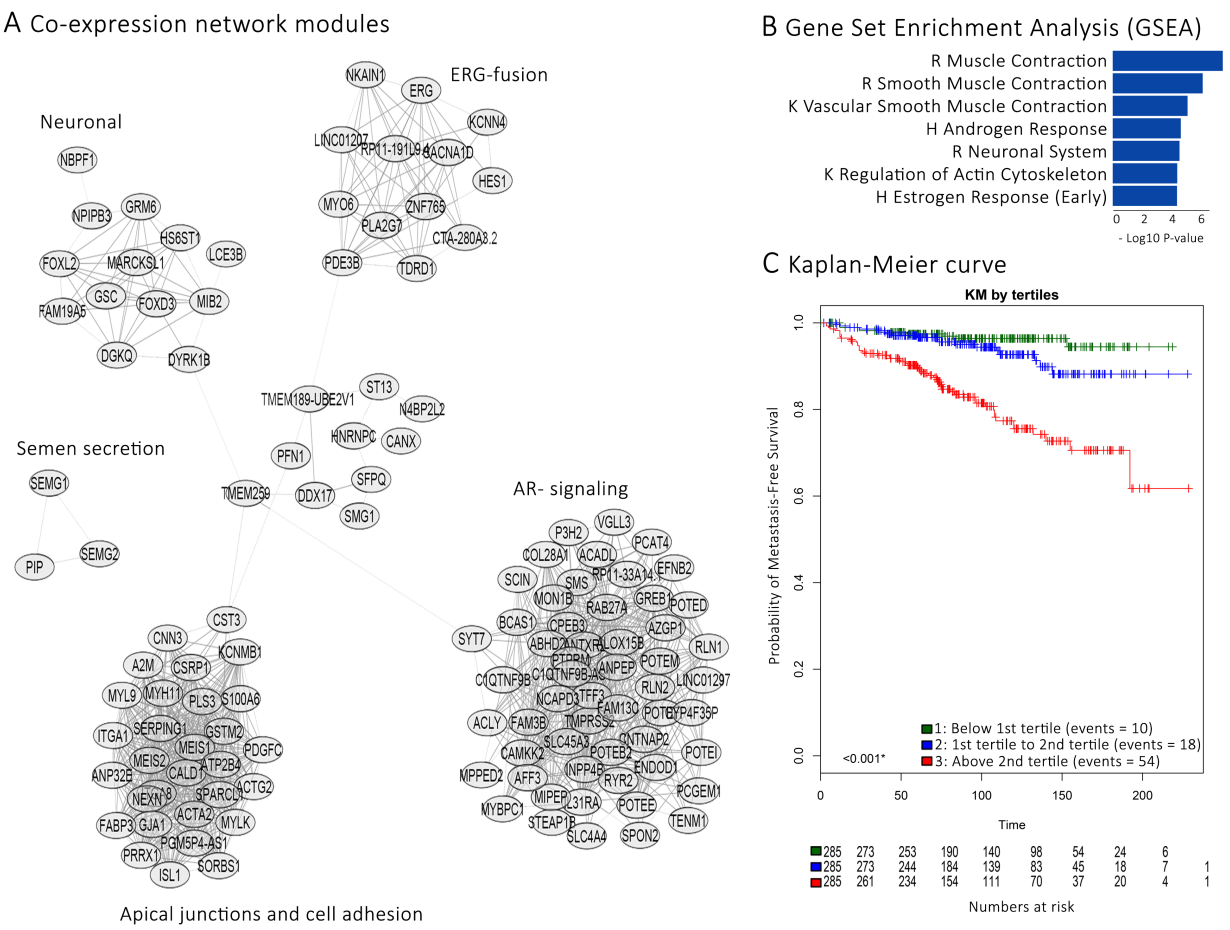

Fig. 2. Biological validation and evaluation of the DHW genes and the signature obtained by DGS method (A) Co-expression modules for 141 DHW genes which shows DAE is capturing different distinct biology related to PCa. (B) GSEA for those genes with non-zero coefficients in the classifier (K, R, and H denotes KEGG, REACTOME, and Hallmarks resources respectively). (C) KM curve for DGS signature which indicates it can successfully distinguish indolent tumors from metastatic ones.

4 Discussion

In this paper, we investigated the problem of signature discovery for PCa metastasis and addressed the challenge of small training datasets with clinical outcomes (labelled). We proposed to extract knowledge from a large unlabelled cohort with DAE. This information was used to extract DHW genes to build an accurate logistic regression model to predict metastasis. We found a novel signature based on those genes that had non-zero coefficients in this classifier. To our knowledge, this is the first study of the applicability of information from an unlabelled cohort to predict a phenotype (metastasis) in cancer research.

The prognostic capability of the discovered signature by DGS method was successfully validated in five independent cohorts involving a total of 1,157 patients and showed superior performance compared to the state-of-the-art signatures for PCa metastasis. More importantly, the signature obtained by our DGS provided independent prognostic information in addition to the clinical variables. Moreover, our signature also provided novel information to each and all of the state-of-the-art signatures (separately and combined) according to MVA. This observation was indicative of effectiveness of the proposed gene signature for prognostic purposes.

Characterizing the selected DHW genes (extracted from the DAEs) and those genes that had non-zero coefficients (in the classifier) revealed several interesting biological events. First, the RAE captured highly correlated gene modules that were related to the metastatic potential of PCa. Each of these modules represented unique distribution that was representative of a molecular pathway. More importantly, the signature obtained by our DGS method showed enriched pathways highly relevant to PCa metastasis and also revealed novel genes (e.g PGM5P4-AS1) that have not been associated with PCa metastasis and thus require further experimental validation.

In addition to the proposed DGS method, we experimented with several other alternative methods based on DAEs. For example, we attempted building a classifier based on the hidden representation of the RAE (h_{Ret}). To our surprise, this model was not able to predict metastasis accurately (AUC=0.64 on the Mayo I test set instead of 0.72 of the current model). From the computational point of view, one possible explanation is that the hidden representation had not converged to a stable state. We argue that if this was the case, we would not have been able to build an accurate logistic regression model based on DHW genes which are essentially extreme cases in capturing factors of variation. We also built another logistic regression classifier based on high weight genes obtained from the PAE (the first DAE trained with the unlabelled data) and its performance was still fairly competitive. Another explanation is that using the RAE deludes salient features extracted from the unlabelled data. We analyzed this by building two logistic classifiers based on hidden representations of the PAE (h_{Pro}) and the RAE (h_{Ret}) separately and observed an increment in AUC (from 0.60 in the PAE to 0.64 in the RAE). This observation indicated that

the RAE made the hidden representation more suitable for prediction of metastasis. We believe the explanation for the poor performance of the classifier based on hidden representation was that the generic learned factors of variation by DAEs were associated with some other unobserved phenotype(s) (not metastasis). In the abstract hidden representation level, signals which were associated with metastasis were deluded with strong factors of variation associated with that unobserved phenotype(s). On the other hand, this delusion was less severe at the input (genes) level because we only focused on those genes that had high weights and were more distinguished than the others. These genes captured most of the factors of variation in both of the unlabelled and labelled cohorts with less delusion by the other genes.

5 Conclusion

In this paper, we proposed the Deep Genomic Signature (DGS) method for transferring knowledge from an unlabelled cohort to a labelled one in order to build an accurate prognostic model in PCa to predict metastasis. DGS consisted of four steps: training a DAE for the unlabelled cohort, training a DAE for the labelled cohort, DHW genes selection from the second DAE, and training a classifier with those selected DHW genes. We showed that 141 DHW genes in the second DAE are carrying significant amount of information that can be used to train a logistic regression model to predict PCa metastasis. after training the classifier in the last step, 38 genes ended up with non-zero coefficients and were considered as a new signature for PCa metastasis. This signature was successfully validated on five independent cohorts and demonstrated superior performance and outperformed state-of-the-art signatures for metastatic PCa in terms of the prediction accuracy (AUC). We further analyzed and interpreted clinical and biological significance of our signature and found mounting amount of evidence that this 38-gene signature for PCa metastasis can provide patient management valuable information.

In further studies we aim to explore the use of both supervised and unsupervised deep transfer learning in PCa and integrate multiple types of biological information to further improve PCa patient metastasis prediction.

Acknowledgements

We would like to thank our collaborators from Johns Hopkins (Ashley Ross, Bruce Trock, and Edward M. Schaeffer), Mayo Clinic (Robert B. Jenkins and R. Jeffrey Karnes), Thomas Jefferson University (Robert Den, Adam Dicker, Leonard Gomella), Cleveland Clinic (Eric Klein), and others at these institutions for supporting this work by providing patient samples with clinical annotations and insights into their interpretation. We would also like to thank Thiago Toman Parrilha and Kevin Gao from GenomeDx for IT support on configuring the Tensorflow environment. Finally, we would like to thank Yen-Yi Lin from the Vancouver Prostate Centre for his support.

Funding

This study was supported by CIHR grant (identifier: PJT-153073) and the NSERC CREATE program, Computational Methods for the Analysis of the Diversity and Dynamics of Genomes.

Conflict of interests

Y.L., N.E., M.A., and E.D. are employees of GenomeDx Biosciences.

References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, **33**(8), 831–838.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology*, **12**(7), 878.

Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36.

Bengio, Y., Courville, A. C., and Vincent, P. (2012). Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR, abs/1206.5538*, **1**.

Chen, W., Tsai, Y., Siu, M., Yeh, H., Chen, C., Yin, J., Huang, J., and Liu, Y. (2017). Inhibition of the androgen receptor induces a novel tumor promoter, zbtb46, for prostate cancer metastasis. *Oncogene*, **36**(45), 6213.

Chen, Y., Li, Y., Narayan, R., Subramanian, A., and Xie, X. (2016). Gene expression inference with deep learning. *Bioinformatics*, **32**(12), 1832–1839.

Cuzick, J., Swanson, G. P., Fisher, G., Brothman, A. R., Berney, D. M., Reid, J. E., Mesher, D., Speights, V., Stankiewicz, E., Foster, C. S., *et al.* (2011). Prognostic value of an rna expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. *The lancet oncology*, **12**(3), 245–255.

Danaee, P., Ghaeini, R., and Hendrix, D. A. (2016). A deep learning approach for cancer detection and relevant gene identification. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 22, page 219. NIH Public Access.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Den, R. B., Feng, F. Y., Showalter, T. N., Mishra, M. V., Trabulsi, E. J., Lallas, C. D., Gomella, L. G., Kelly, W. K., Birbe, R. C., McCue, P. A., *et al.* (2014). Genomic prostate cancer classifier predicts biochemical failure and metastases in patients after postoperative radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, **89**(5), 1038–1046.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, **12**(Jul), 2121–2159.

Erho, N., Crisan, A., Vergara, I. A., Mitra, A. P., Ghadessi, M., Buerki, C., Bergstralh, E. J., Kollmeyer, T., Fink, S., Haddad, Z., *et al.* (2013). Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PloS one*, **8**(6), e66855.

Gómez-Bombarelli, R., Duvenaud, D., Hernández-Lobato, J. M., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2016). Automatic chemical design using a data-driven continuous representation of molecules. *arXiv preprint arXiv:1610.02415*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., and Zhavoronkov, A. (2017). drugan: an advanced generative adversarial autoencoder model for de-novo generation of new molecules with desired molecular properties in silico. *Molecular Pharmaceutics*.

Karnes, R. J., Choeurng, V., Ross, A. E., Schaeffer, E. M., Klein, E. A., Freedland, S. J., Erho, N., Yousefi, K., Takhar, M., Davicioni, E., *et al.* (2017). Validation of a genomic risk classifier to predict prostate cancer-specific mortality in men with adverse pathologic features. *European Urology*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Klein, E. A., Yousefi, K., Haddad, Z., Choeurng, V., Buerki, C., Stephenson, A. J., Li, J., Kattan, M. W., Magi-Galluzzi, C., and Davicioni, E. (2015). A genomic classifier improves prediction of metastatic disease within 5 years after surgery in node-negative high-risk prostate cancer patients managed by radical prostatectomy without adjuvant therapy. *European urology*, **67**(4), 778–786.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems*, **1**(6), 417–425.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2015). Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, **6**, 26094.

Mo, F., Lin, D., Takhar, M., Ramnarine, V. R., Dong, X., Bell, R. H., Volik, S. V., Wang, K., Xue, H., Wang, Y., *et al.* (2017). Stromal gene expression is predictive for metastatic primary prostate cancer. *European Urology*.

Nakagawa, T., Kollmeyer, T. M., Morlan, B. W., Anderson, S. K., Bergstralh, E. J., Davis, B. J., Asmann, Y. W., Klee, G. G., Ballman, K. V., and Jenkins, R. B. (2008). A tissue biomarker panel predicting systemic progression after psa recurrence

post-definitive prostate cancer therapy. *PloS one*, **3**(5), e2318.

Ng, A. (2011). Sparse autoencoder. *CS294A Lecture notes*, **72**(2011), 1–19.

NIH (2017). Prostate cancer: https://seer.cancer.gov/statfacts/html/prost.html.

Penney, K. L., Sinnott, J. A., Fall, K., Pawitan, Y., Hoshida, Y., Kraft, P., Stark, J. R., Fiorentino, M., Perner, S., Finn, S., *et al.* (2011). mrna expression signature of gleason grade predicts lethal prostate cancer. *Journal of Clinical Oncology*, **29**(17), 2391–2396.

Piccolo, S. R., Sun, Y., Campbell, J. D., Lenburg, M. E., Bild, A. H., and Johnson, W. E. (2012). A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*, **100**(6), 337–344.

Pollard, T. D. and Cooper, J. A. (2009). Actin, a central player in cell shape and movement. *Science*, **326**(5957), 1208–1212.

Prensner, J. R., Zhao, S., Erho, N., Schipper, M., Iyer, M. K., Dhanasekaran, S. M., Magi-Galluzzi, C., Mehra, R., Sahu, A., Siddiqui, J., *et al.* (2014). Rna biomarkers associated with metastatic progression in prostate cancer: a multi-institutional high-throughput analysis of schlap1. *The lancet oncology*, **15**(13), 1469–1480.

Ramaswamy, S., Ross, K. N., Lander, E. S., and Golub, T. R. (2003). A molecular signature of metastasis in primary solid tumors. *Nature genetics*, **33**(1), 49–54.

Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011). Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 833–840.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, **12**(1), 77.

Ross, A. E., Johnson, M. H., Yousefi, K., Davicioni, E., Netto, G. J., Marchionni, L., Fedor, H. L., Glavaris, S., Choeurng, V., Buerki, C., *et al.* (2016). Tissue-based genomics augments post-prostatectomy risk stratification in a natural history cohort of intermediate-and high-risk men. *European urology*, **69**(1), 157–165.

Spratt, D. E., Yousefi, K., Deheshi, S., Ross, A. E., Den, R. B., Schaeffer, E. M., Trock, B. J., Zhang, J., Glass, A. G., Dicker, A. P., *et al.* (2017). Individual patient-level meta-analysis of the performance of the decipher genomic classifier in high-risk men after prostatectomy to predict development of metastatic disease. *Journal of Clinical Oncology*, pages JCO–2016.

Tan, J., Ung, M., Cheng, C., and Greene, C. S. (2015). Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 20, page 132. NIH Public Access.

Tan, J., Hammond, J. H., Hogan, D. A., and Greene, C. S. (2016). Adage-based integration of publicly available pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. *mSystems*, **1**(1), e00025–15.

Tan, J., Doing, G., Lewis, K. A., Price, C. E., Chen, K. M., Cady, K. C., Perchuk, B., Laub, M. T., Hogan, D. A., and Greene, C. S. (2017). Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks. *Cell Systems*, **5**(1), 63–71.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of computational and graphical statistics*, **12**(1), 156–175.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, **99**(10), 6567–6572.

Tyekucheva, S., Bowden, M., Bango, C., Giunchi, F., Huang, Y., Zhou, C., Bondi, A., Lis, R., Van Hemelrijck, M., Andrén, O., *et al.* (2017). Stromal and epithelial transcriptional map of initiation progression and metastatic potential of human prostate cancer. *Nature Communications*, **8**.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, **11**, 3371–3408.

Way, G. P. and Greene, C. S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 23, page 80. NIH Public Access.

Xie, W., Regan, M. M., Buyse, M., Halabi, S., Kantoff, P. W., Sartor, O., Soule, H., Clarke, N. W., Collette, L., Dignam, J. J., Fizazi, K., Paruleker, W. R., Sandler, H. M., Sydes, M. R., Tombal, B., Williams, S. G., Sweeney, C. J., and on behalf of the ICECaP Working Group (2017). Metastasis-free survival is a strong surrogate of overall survival in localized prostate cancer. *Journal of Clinical Oncology*, **35**(27), 3097–3104. PMID: 28796587.

Yamazaki, D., Kurisu, S., and Takenawa, T. (2005). Regulation of cancer cell motility through actin reorganization. *Cancer science*, **96**(7), 379–386.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, **12**(10), 931.