Smolander *et al. BMC Cancer* (2019) 19:1176
https://doi.org/10.1186/s12885-019-6338-1

BMC Cancer

RESEARCH ARTICLE Open Access

Comparing biological information contained in mRNA and non-coding RNAs for classification of lung cancer patients

Johannes Smolander[1,2], Alexey Stupnikov[3], Galina Glazko[4], Matthias Dehmer[5,6,7] and Frank Emmert-Streib[1,8*]

Abstract

Background: Deciphering the meaning of the human DNA is an outstanding goal which would revolutionize medicine and our way for treating diseases. In recent years, non-coding RNAs have attracted much attention and shown to be functional in part. Yet the importance of these RNAs especially for higher biological functions remains under investigation.

Methods: In this paper, we analyze RNA-seq data, including non-coding and protein coding RNAs, from lung adenocarcinoma patients, a histologic subtype of non-small-cell lung cancer, with deep learning neural networks and other state-of-the-art classification methods. The purpose of our paper is three-fold. First, we compare the classification performance of different versions of deep belief networks with SVMs, decision trees and random forests. Second, we compare the classification capabilities of protein coding and non-coding RNAs. Third, we study the influence of feature selection on the classification performance.

Results: As a result, we find that deep belief networks perform at least competitively to other state-of-the-art classifiers. Second, data from non-coding RNAs perform better than coding RNAs across a number of different classification methods. This demonstrates the equivalence of predictive information as captured by non-coding RNAs compared to protein coding RNAs, conventionally used in computational diagnostics tasks. Third, we find that feature selection has in general a negative effect on the classification performance which means that unfiltered data with all features give the best classification results.

Conclusions: Our study is the first to use ncRNAs beyond miRNAs for the computational classification of cancer and for performing a direct comparison of the classification capabilities of protein coding RNAs and non-coding RNAs.

Keywords: Deep learning, Deep belief network, Classification, Non-coding RNA, Lung cancer and Machine learning

Background

Lung cancer is one of the most common cancers in humans worldwide among both men and women, as well as the leading cause of cancer-related deaths [1]. There are two major types of lung cancer, non-small-cell lung cancer (NSCLC) and small-cell lung cancer and adenocarcinoma is a subtype of NSCLC and the most common type in patients who never smoked [1]. In recent years, next-generation sequencing (NGS) technologies have opened an experimental door for the systematic study of complex diseases, including lung cancer, by allowing the generation of high-throughput data on all relevant molecular levels [2, 3]. However, one problem we are facing in this context, is that we are still discovering new variables that might be of crucial importance in understanding the organizational principles of the molecular machinery. For this reason is it not surprising that molecular and genomic medicine are still at its infancy [4–6]. One example for such players are non-coding RNAs (ncRNAs) [7–9].

Non-coding RNAs (ncRNAs) is a broad class of transcripts, consisting of well known transcripts with structural (rRNAs, tRNAs, snRNAs, snoRNAs, etc.) and regulatory (miRMAs, siRNA, piRNAs, etc.) roles, and

*Correspondence: v@bio-complexity.com
[1]Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland
[8]Institute of Biosciences and Medical Technology, Tampere, Finland
Full list of author information is available at the end of the article

transcripts whose functions remain largely unknown [10–12]. The latter includes sense/antisense transcripts, ranging in length from 200 bp to 100 kb. Collectively they are called long non-coding RNAs (lncRNAs) (Wang et al., 2011) and sometimes referred to as genomic 'dark matter' [13, 14]. Large-scale evolutionary properties of the bulk of lncRNAs [15] and the existence of hundreds of experimentally characterized lncRNAs [16, 17] suggest that many of them have a well-defined biological function [12, 13]. The catalogue of the functionally annotated part of the non-coding transcriptome (70-90% of transcribed matter [10, 18, 19] is constantly growing, however, at the moment the total number of non-coding RNAs is unknown. Recent estimates suggest that there are thousands in the human genome [20, 21].

Among the many categories of ncRNAs the most well understood are miRNAs (also called microRNAs), siRNAs and piRNAs, which guide effector Argonaute proteins to genomic loci or target RNAs in a sequence-specific manner. lncRNAs, on the other hand, are implicated mostly in the regulation of gene expression and many are functionally validated by now to be involved in different cellular and developmental pathways [22, 23]. Dysregulation of lncRNAs is observed in many human diseases, including colon cancer, breast cancer, leukemia, ischaemic heart disease, Alzheimer's disease and some others (see [20] for a review). The ever-growing experimental evidence implicating ncRNAs regulatory roles in many biological processes has led to the idea of using ncRNAs as disease biomarkers, e.g., for diagnostic purposes [24].

Regarding the classification of disorders, lncRNAs have been used to distinguish different subtypes in human breast cancer and glioblastoma [25, 26]. However, all of these studies used unsupervised hierarchical clustering or similar methodology for obtaining the class predictions rather than automized supervised methods [27].

Our paper contributes to these diagnostic investigations by studying the classification capabilities of protein coding and non-coding RNAs from lung cancer. Specifically, we are using RNA-seq data from lung adenocarcinoma patients [28] generated with an Illumina HiSeq 2000 platform containing information about patients with lung cancer and matched control samples from adjacent normal tissue. The availability of RNA-seq data allows us to obtain information about protein coding RNAs and non-coding RNAs. We utilize this opportunity investigating the predictive abilities of both data sources by studying the classification of the lung cancer patients.

We perform this analysis for a number of different state-of-the-art classification methods, including deep learning neural networks, decision trees, random forrests and support vector machines [29–34]. We study the dependency of these methods on a multitude of model parameters, e.g., the neural network architecture, the learning algorithms or the kernels. In addition, we study the influence of feature selection on the results. Our results will shed light on the discriminatory information content of ncRNAs in comparison with coding RNAs.

For our classification task, we make use of recent progress in deep learning models based on neural networks [31]. Despite the fact that neural network models are known since many decades [35–40] recent advances in the learning methodology revived them [31]. Specifically, in contrast to classic neural network models, deep neural networks can have a large number of hidden layers. Each of these layers builds a complex representation of the previous layers as a result from nonlinear transformations [41].

Deep learning models have been successfully used in many application areas, most notably in image recognition [41, 42] and speech recognition [43]. In computational biology, deep learning is still at an early stage and the studied data come mostly from the DNA-level. For instance, alternative splicing and protein binding patterns have been studied [44–46]. For analyzing gene expression data and especially for the classification of cancer very little is know. One of the few studies in this area is from [47]. They used data from DNA microarrays to classify lung adenocarcinoma and squamous cell carcinoma. However, they used not only the classification data set but additional data from further disease stages to train their deep learning model in the unsupervised phase. This is possible since during the unsupervised training phase no labels are needed nor used. A more conventional analysis has been conducted by [30]. They used deep forest models for the classification of various cancer types based on gene expression data from both DNA microarray and RNA-Seq. For this Stacked Denoising Autoencoders in combination with ANNs and SVMs have been used. However, neither of these studies investigated ncRNAs, only coding RNAs have been analyzed.

Our paper is organized as follows. In the next section we present details about the lung cancer data and the methods for our analysis. Then we present our results and a discussion of these. We finish the paper with concluding remarks.

Methods

Lung cancer data

For our analysis, we are using RNA-seq data from lung cancer in Koreans [28]. The data were generated with Illumina HiSeq 2000 and contain matched information about 62 subjects with lung adenocarcinoma and 62 subjects from adjacent normal tissues (control samples). These samples are taken from cancer tissues whose driver mutations were not detected by screening tests (Sanger sequencing for EGFR and for KRAS point mutations and fluorescence in situ hybridization for EML4-ALK fusion);

see [28] for details. Access to the data is provided via Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/), accession number GSE40419.

The processing pipeline for the data include the following steps [48]: The dataset [28] was extracted from SRA archive [49]. The samples were aligned using Bowtie2 [50], with 1 mismatch to the hg38 human genome [51]. The reads were summarized to count vectors with samExploreR [48, 52] and normalized with RPKM (Reads Per Kilobase Million - also used in [28]) . RefSeq annotation was taken for reads mapping procedure. Then transcripts that were not expressed or at a very low intensity were removed. On a technical note, we would like to remark that we repeated our analysis by using TPM (Transcripts Per Kilobase Million) instead of RPKM but found no noticeable differences in our results.

TopHat and STAR would be alternative choices instead of Bowti2, which are quite commonly used in transcriptomics studies. Their major difference from BowTie2 is an opportunity to account for reads in splice-affected regions. However, in [53] it was argued that these reads make no impact on transcript abundance quantification, and therefore, for the purpose of Differential Gene Expression Bowtie2 alignment is an applicable procedure.

Using the human genome annotation databases Reference Sequence (hg38) as reference to map RNA sequencing reads to protein coding RNAs, we find expression levels for 36,742 RefSeq genes. After filtering out redundant entries this results in 22,427 gene transcripts. Removing genes with low expression values (maximum expression smaller than 3 RPKM) or genes with a small standard variation results in 12,360 genes corresponding to protein coding RNAs we used for our analysis.

For a similar analysis for non-coding transcripts we find 3124 non-coding RNAs (ncRNAs) for RPKM and 1398 ncRNAs for TPM (16828 before filtering). That means the majority of ncRNAs is expressed at a very low level. Among these the most important ncRNAs are [54]:

- microRNA: miRNA
- ribosomal RNA: rRNA
- small interfering RNA: siRNA
- long non-coding RNAs: long ncRNAs or lncRNAs
- small nuclear ribonucleic acid: snRNA

Error measures

For our analysis, we need to assess the performance of a binary classification. The results form such a classification can be summarized by a confusion matrix shown in Table 1.

From the confusion matrix in Table 1 one can derive the following three performance metrics [55, 56].

- Accuracy $(A) = \frac{TP+TN}{TP+FP+FN+TN}$

Table 1 Confusion matrix summarizing the results for binary classifications

		True class	
		Positive	Negative
Predicted class	Positive	True positives (TP)	False positives (FP)
	Negative	False negatives (FN)	True negatives (TN)

- True positive rate (TPR) (also called sensitivity) $= \frac{TP}{TP+FN}$
- True negative rate (TNR) (also called specificity) $= \frac{TN}{TN+FP}$

For our analysis a true positive (TP) indicates a correctly predicted lung adenocarcinoma sample and a true negative (TN) a correctly predicted control sample. Hence, the true positive rate is with respect to lung adenocarcinoma and the true negative rate for the control samples. The true positive rate (TPR), also called sensitivity, evaluates the proportion of all positives correctly identified and the true negative rate (TNR), also called specificity, evaluates the proportion of all negatives correctly identified. With respect to the confusion matrix the TPR and the TNR are symmetrically defined by exchanging the class labels. Overall, the TPR has a focus on positive labels which correspond in our case to lung adenocarcinoma patients and the TNR has a focus on negative labels corresponding to control patients. In contrast, the accuracy assesses the overall classification performance for both classes in a weighted manner. In addition, we evaluate the area under the receiver operator characteristics (AUROC) curve [57]. For the construction of a ROC curve one needs pairs of TPR and FPR values. These values are obtained for different threshold parameters of the classifier. Practically, it is sufficient to rank all values for the samples and use successively different threshold values. Technical details are described in [55].

Due to the fact that the number of samples in both classes is exactly the same, none of our measures suffers from negative consequences of imbalanced classes [58].

For assessing the variability of our results, we use a 10-fold cross validation (CV) in order to estimate the standard errors of the performance measures. CV splits the data into 10-folds whereby one fold is used for training (estimation of parameters of the models) and 9-folds are used for testing. CV is a resampling method that is the gold standard for error estimations in order to avoid a high bias [59, 60].

Deep belief networks

For our analysis, we are using Deep Belief Networks (DBNs) [31]. DBN models are trained in two separate phases. In the first phase, a Restricted Boltzmann Machine (RBM) is used to initialize the model, and in the second phase a supervised method is used for tuning of

the parameters [61]. These steps are called pre-training phase and fine-tuning phase. For the fine-tuning we are using the *stochastic gradient descent* either in combination with the *basic backpropagation* (Bprop) algorithm or the *resilient backpropagation* (Rprop) algorithm, whereas the *resilient backpropagation* (Rprop) algorithm is a more efficient, faster variant of the Bprop.

In the following, we describe the pre-training and fine-tuning steps briefly.

Unsupervised pre-training

In principle, it is possible to learn neural network models solely by supervised learning methods skipping a pre-training step entirely. However, it has been shown that pre-training with a suitable initialization of the model parameters, i.e., the weights of the network, can make the supervised learning step much faster and improve the overall performance [41]. The Restricted Boltzmann Machine (RBM) has been introduced for this pre-training step providing for this an unsupervised initialization of the parameters [31, 62]. This allows the training of deep architectures, i.e., networks with many hidden layers, that can achieve better performances than shallow architectures with only one hidden layer.

Technically, the pre-training step of DBNs consists in stacking RBMs, so that the next RBM in a chain is trained by using the previous hidden layer as its input layer, in order to initialize parameters for each layer. In previous studies this approach has been found to be efficient [63]. Furthermore, this allows the order of the layers to be trained to be chosen freely. For example, one can train the last layer first and after a certain number of epochs the preceding layers can be trained [31]. For our analysis, we are using a Restricted Boltzmann Machine model with binary units and the contrastive divergence (CD) algorithm for approximating the log-likelihood of the RBM.

Supervised fine-tuning

In the supervised training step the parameters of the model are optimized by fine-tuning the model. For this step, the class labels of the training data are used, which makes this step supervised. In contrast, the pre-training step does not make use of these labels and for this reason is unsupervised.

The *resilient backpropagation* (Rprop) algorithm is a motified version of the backpropagation algorithm. The purpose for introdicing this algorithm was to speed-up the backpropagation (Bprop) algorithm [64]. There are several realizations of Rprop available [65]. However, for our study, we are using the iRprop$^+$ algorithm. iRprop$^+$, as well as all other realizations, are available in the darch package [66]. In addition to iRprop$^+$, we are using the basic backpropagation algorithm (Bprop) for reasons of comparison.

Network architecture

The architecture of the neural network is a parameter of the model that needs to be defined. From previous studies it is known that there is not one type of network architecture that is best under all conditions but the choice of the architecture is data and problem dependent. For instance, some studies use a decreasing architecture (the number of neurons in the hidden layers decreases) [63], whereas others use an increasing architecture [67] or even a constant architecture [68]. This implies that there is no consensus for deep learning networks about the shape of the architecture. For this reason, we were testing a vast number of different network architectures to find the best one for our problem. In the "Results" section, we provide information about the architectures we were testing.

Results

In the following, we will analyze RNA-seq data from lung cancer patients in two ways. First, we will only use gene expression values from protein-coding RNAs corresponding to mRNAs. Second, we will only use gene expression values from non-coding RNAs (ncRNAs) including miRNAs.

Protein-coding genes

The RNA-seq data for our analysis consist of 62 samples from lung adenocarcinoma and 62 samples from adjacent normal tissues corresponding to control samples. For our first analysis will only use gene expression data from protein-coding genes. For our data set this corresponds to 12360 mRNAs (genes). We will use these data for a binary classification separating adenocarcinoma samples from control samples. The results of this classification are summarized in Table 2.

Our results are interesting for several reasons. First, the DBN classifier in combination with SVM or alone outperform the SVM, decision tree (DT) and random forrest (RF) for almost all combinations. Second, Bprop and Rprop perform similarly with only a slight advantage for Rprop. This is somehow surprising because it is know that Rprop performs in general better than Bprop. Third, the deep learning classifiers outperform the SVM but only slightly.

In Table 3 we show some examples for the further configurations we studied for DBNs and SVMs. The best results are color highlighted. Overall, the architecture of the DBN seems to tolerate a large variability in the number of hidden layers as well as their sizes. This hold for both algorithms Bprop and Rprop. Interestingly, repeating the above analysis by using various feature selection mechanisms we did not find a beneficial effect for the deep learning model, in contrast, the performance decreases. Similar results hold also for the SVM. Only the decision tree and random forrest classifiers benefit somewhat

Table 2 Summary of the best classification results for lung cancer

Classifier	A %	TPR %	TNR %	AUROC %
Task: AC vs N (protein-coding)				
DBN + Bprop	95.65 ± 0.13	97.90 ± 0.25	93.39 ± 0.16	95.64 ± 0.19
DBN + Bprop + SVM	94.19 ± 0.11	95.16 ± 0.50	93.23 ± 0.22	94.19 ± 0.36
DBN + Rprop	95.73 ± 0.17	98.39 ± 0.50	93.06 ± 0.34	95.72 ± 0.42
DBN + Rprop + SVM	**95.97** ± 0.50	98.39 ± 0.52	93.55 ± 0.50	95.97 ± 0.51
SVM	93.66 ± 0.81	92.66 ± 0.87	94.66 ± 0.91	93.66 ± 0.89
DT (100 genes)	89.33 ± 1.27	88.00 ± 1.31	88.16 ± 0.93	88.08 ± 1.12
RF (500 genes)	94.50 ± 0.62	92.33 ± 0.67	96.16 ± 0.59	94.24 ± 0.63
Literature	A %	features	samples	
SVM [69]	95.30	44 genes	445 AC & 19 N	
SVM [70]	91.00	12 genes	73 AC & 80 N	

Only RNA-seq data from protein-coding genes were used. The best classifier is highlighted in green. In addition, for comparison reference results from the literature are shown, highlighted in blue

from a feature selection and we obtain for 100 genes the best results for a decision tree and for 500 gene the best results for a random forrest classifier. However, the benefit for both methods is moderate because without feature selection the accuracy of both classifiers drops by only 2.5%.

In order to compare our results with previous findings, we performed a literature search. From this we found two related studied, the results are shown in Table 3 color highlighted in blue. Specifically, in [69] a SVM was used to classify 445 adenocarcinoma (AC) and 19 normal (N) samples from RNA-seq data using 44 gene features. Also in [70] a SVM was used. In their case, 73 adenocarcinoma (AC) and 80 normal (N) samples from RNA-seq data were classified using 12 gene features. Overall, we observe that our results are competitive and even provide slightly better results.

There are further studies about the classification of lung cancer, but they are less close to our setting. For instance, in [71] lung adenocarcinoma with normal lung tissue samples have been compared. However, the data set they used contained only 5 normal samples in total. This is from a statistical point of view highly problematic because the number of normal samples in the training sets seems far too small. The accuracy values they obtained were 97.2% (SVM), 97.2% (Radial basis function Neural Nets), 97.2% (Multi-layer perceptron), 95.8% (Bayesian network), 94.4% (J48 decision tree) and 95.8% (random

Table 3 A: DBN results for the RNA-Seq data set. B: SVM results for the RNA-Seq data set

A.

		DBN			DBN and SVM		
Model	Architecture	A %	TPR %	TNR %	A %	TPR %	TNR %
DBN + Bprop	A-500-250-100-1	95.48 ± 0.25	98.06 ± 0.40	92.90 ± 0.36	94.13 ± 0.27	97.58 ± 0.27	91.94 ± 0.34
DBN + Bprop	A-100-1	95.65 ± 0.13	97.90 ± 0.25	93.39 ± 0.16	94.19 ± 0.11	95.16 ± 0.50	93.23 ± 0.22
DBN + Rprop	A-5-10-1	95.73 ± 0.17	98.39 ± 0.50	93.06 ± 0.34	95.89 ± 0.08	98.39 ± 0.50	93.39 ± 0.16
DBN + Rprop	A-50-1	95.16 ± 0.50	98.39 ± 0.50	91.94 ± 0.50	**95.97 ± 0.50**	98.39 ± 0.50	93.55 ± 0.50
Task: AC vs N; non-coding, A = 3124							
DBN + Bprop	A-2000-1000-500-1	**96.77 ± 0.50**	100 ± 0.50	93.55 ± 0.50	95.16 ± 0.50	96.94 ± 0.16	93.39 ± 0.16
DBN + Bprop	A-100-1	95.97 ± 0.50	100 ± 0.50	91.94 ± 0.50	95.48 ± 0.13	97.42 ± 0.26	93.55 ± 0.50
DBN + Rprop	A-5-10-1	96.05 ± 0.22	98.39 ± 0.50	93.71 ± 0.45	96.45 ± 0.13	98.06 ± 0.22	94.84 ± 0.22
DBN + Rprop	A-50-1	94.35 ± 0.50	100 ± 0.50	88.71 ± 0.50	95.48 ± 0.13	98.39 ± 0.50	92.58 ± 0.26

B.

		Radial			Linear		
Data	Features	A %	TPR %	TNR %	A %	TPR %	TNR %
Protein-coding	12360	**93.66 ± 0.81**	92.66 ± 0.87	94.66 ± 0.91	93.00 ± 0.89	90.00 ± 0.89	96.00 ± 0.85
Non-coding	3124	91.41 ± 0.97	88.00 ± 0.99	94.83 ± 0.56	**93.91 ± 0.89**	90.50 ± 0.73	97.33 ± 0.72

The best results are shown in bold

forest). These results were obtained for 917 features. By using features selection mechanism reducing the dimension to 50 they were able to further improve the results for some methods. However, also these improvements are only very moderate.

Another study by [72] used 86 lung adenocarcinomas samples and 10 non-neoplastic samples, but no controls. They tested two different methods and achieved for both 100% accuracy using 813 and 9 features respectively. Also the study by [73] did not use normal samples, but compared 150 lung adenocarcinomas samples with 31 malignant pleural mesothelioma for expression data from DNA microarrays. Comparing seven different methods they found accuracy values between 96.13% (decision tree) and 99.45% (SVM). For the decision tree 3 features have been used and for the SVM 12533 (no feature selection has been applied). The same data set has been used in several other studies, e.g., in [74] in comparison with 9 different classifiers and they found a SVM with linear kernel to perform best with an accuracy of 98.13% for 20 feature genes and in [75] with a two-gene classifier (TGC) based on finding optimal cuts they reached 93% with 2 genes as features.

We want to highlight that in addition to the differences mentioned above, all of these less close studies used DNA microarray data but no RNA-seq data.

Regarding the application of deep learning classifiers, in [47] samples from adenocarcinoma and squamous cell carcinoma were compared with various deep learning methods. They obtained accuracy values ranging from 87.5 to 93.33%. The best performing method utilized a PCA for dimension reduction as input for an one or multi-layered sparse autoencoder connected to a SVM (Gaussian kernel) as classifier. It is important to highlight that in addition to the data mentioned, they used further gene expression data from lung cancer patients for their unsupervised learning phase to improve the general learning behavior. This included also lung cancer patients with other tumor types than adenocarcinoma or squamous cell carcinoma. Such data could be used because in the unsupervised phase for learning the autoencoder the labels are ignored. We found only one further study that applied deep learning to gene expression data from DNA microarray [76], but not from lung cancer. However, also in [76] a feature selection mechanism was used (Infinite Feature Selection [77]) selecting 500 genes as input for the deep learning models.

As a result from this literature overview its seems that our study is the first to investigate the classification of deep learning classifiers without feature selection.

Non-coding RNAs

In this section we are repeating a similar analysis as in the previous section, however, with one important difference. Instead of using RNA-seq data from genes coding for proteins we will use RNA-seq data from genes that do not code for proteins (non-coding genes) leading to non-coding RNAs. In our data set, we have in total 3124 non-coding RNAs.

Our results are summarized in Table 4. This time all configurations of the DBNs outperform the SVM, decision tree and random forrest classifier clearly. Interestingly, learning with Bprop is more beneficial than using Rprop. A possible explanation for this is that the non-coding data set has distinctly less features than the data set for coding RNAs. The difference is about a factor of 4 ($3.96 = 12360/3124$). Hence, one can use more complex network architectures that are learnable.

In Table 4, we included also three results from the literature (highlighted in blue) that performed a comparable analysis. In [78] a K-nearest neighbors (KNN) classifier was used for 19 miRNAs to classify 12 adenocarcinoma (AC) patients from 10 controls (N). In [79] a nearest shrunken centroids (NSC) classifier has been used for 38 miRNAs. For their analysis they use 123 carcinoma (C) samples and 123 controls (N). The carcinomas class included adenocarcinoma, squamous-cell carcinoma and small-cell carcinoma. Similarly, in [24] a compound covariate predictor (CCP) has been applied to 43 miRNAs found to be differentially expressed. However, as one can see from Table 4 all of their classification results were worse than ours.

Recently, it was shown that information about the presence and absence of miRNA isoforms (isomiRs) can be used to discriminate between 32 cancer subtypes [80]. The average sensitivity of a SVM classifier trained on RNAs-seq data was 90%, and between 80-100% for data sets from diverse platforms (Affimetrix miRNA Array, AVI SOLID sequencing) [81]. Of note, the classifier differentiated between lung adenocarcinoma and lung squamous cell carcinoma [80].

All of the results from the literature have in common that they all used feature selection and they all performed worse than our best results. Also to the best of our knowledge there is no study that included ncRNAs beyond miRNAs in their analysis.

Comparison and feature selection

Next, we are comparing the results we obtained for different classification methods (here RF: random forest, DT: decision tree). In Fig. 1a-c we show a summary of results for the coding RNAs (red lines) and non-coding RNAs (green lines) for accuracy, true positive rate and true negative rate.

Overall, the accuracy values for data from non-coding RNAs are for all compared classification methods higher than for data from coding RNAs (Fig. 1a). The differences are not large but sufficient to demonstrate that the predictive abilities of non-coding RNAs are at least as good

Table 4 Summary of the best classification results for lung cancer

Classifier	A %	TPR %	TNR %	AUROC %
Task: AC vs N (non-coding)				
DBN + Bprop	**96.77** ± 0.50	100 ± 0.50	93.55 ± 0.52	96.77 ± 0.51
DBN + Bprop + SVM	95.16 ± 0.12	96.94 ± 0.16	93.39 ± 0.16	95.15 ± 0.16
DBN + Rprop	96.05 ± 0.22	98.39 ± 0.50	93.71 ± 0.45	96.05 ± 0.48
DBN + Rprop + SVM	96.45 ± 0.13	98.06 ± 0.22	94.84 ± 0.22	96.45 ± 0.22
SVM	93.91 ± 0.89	90.50 ± 0.73	97.33 ± 0.72	93.91 ± 0.73
DT	85.83 ± 0.97	84.66 ± 0.93	87.00 ± 0.91	85.83 ± 0.93
RF (500 genes)	89.03 ± 0.77	86.16 ± 0.72	91.83 ± 0.89	88.99 ± 0.81
Literature	A %	Features	Samples	
KNN [78]	85.00	19 miRNAs	12 AC & 10 N	
NSC [79]	69.00	38 miRNAs	123 C & 123 N	
CCP [24]	91.00	43 miRNAs	104 C & 104 N	

Only RNA-seq data from non-coding RNAs were used. The best classifier is highlighted in green. In addition, for comparison reference results from the literature are shown, highlighted in blue

as for coding RNAs for the diagnostics of lung adenocarcinoma. The fact that this holds independently of the used classification method is an indicator for the robustness of this finding irrespectively of the specific statistical methodology.

Due to the fact that deep learning networks do not require a feature selection mechanisms for reducing the dimension of the input data we did not use such a mechanism as a preprocessing step for our analysis so far. However, now we want to study the effect of feature selection [82, 83] in a systematic way.

In Figs. 2, 3, 4, and 5 we show results for the effect of different feature selection mechanisms. Specifically, we used three different feature selection mechanisms producing gene-scores that can be used for rank ordering the genes. We used the variance (A), JIM (joint impurity filter) (B) and JMI (joint mutual information) (C).

The joint impurity filter (JIM) is defined as,

$$J(X_j) = \sum_{W_k \in S} G(X_j, W_k; Y_j). \quad (1)$$

Here X_j is the expression value of feature j, Y_j is its the outcome variable (class label), S is the set of already selected features and $G(X_j, W_k; Y_j)$ is the Gini impurity gain,

$$G(X; Y) = \sum_{xy} \frac{p_{xy}^2}{p_x} - \sum_{y} p_y^2. \quad (2)$$

The method starts with the feature that maximizes the impurity gain and then greedily adds new features that maximize $J(X_j)$.

The joint mutual information (JMI) is defined as [84],

$$J(X_j) = \sum_{W_i \in S} I(X_j, W_i; Y_j). \quad (3)$$

Here S is again the set of already selected features and I is the joint mutual information between X_j, W_i and Y_j. The method adds new features $X_j \notin S$ in a greedy way by maximizing $J(X_j)$.

In Fig. 2 we show results for coding RNAs and in Fig. 3 for non-coding RNAs. The label 'all' corresponds to 12360 coding RNAs (Fig. 2) and 3124 non-coding RNAs (Fig. 3) respectively and FS indicates the applied feature selection mechanism. As one can see, regardless of the chosen feature selection mechanism or the kernel of the SVM reducing the number of features/RNAs is not beneficial for the performance of the classification. In Fig. 5 we show also results for combined data, i.e., we used the coding and non-coding RNAs together. In this case 'all' corresponds to 15484 RNAs. Also for this, no benefit is gained from reducing the number of features (only results for JIM are shown). All of these results have been obtained for RPKM normalized data. In order to demonstrate that the normalization has no effect on our results we show in Fig. 4 results for TPM normalized data. Results for other feature selection mechanisms are similar (not shown).

Overall, these investigations demonstrate that a feature selection does not have a positive effect on the classifiers.

Discussion

From analyzing the classification of lung adenocarcinoma patients by using a number of different state-of-the-art classification methods, we found that data from non-coding RNAs have a comparable classification performance as data from coding RNAs, whereas for DBN we found an even better performance. This demonstrates that (I) both data sources (coding and non-coding RNAs) contain a comparable amount of information regarding

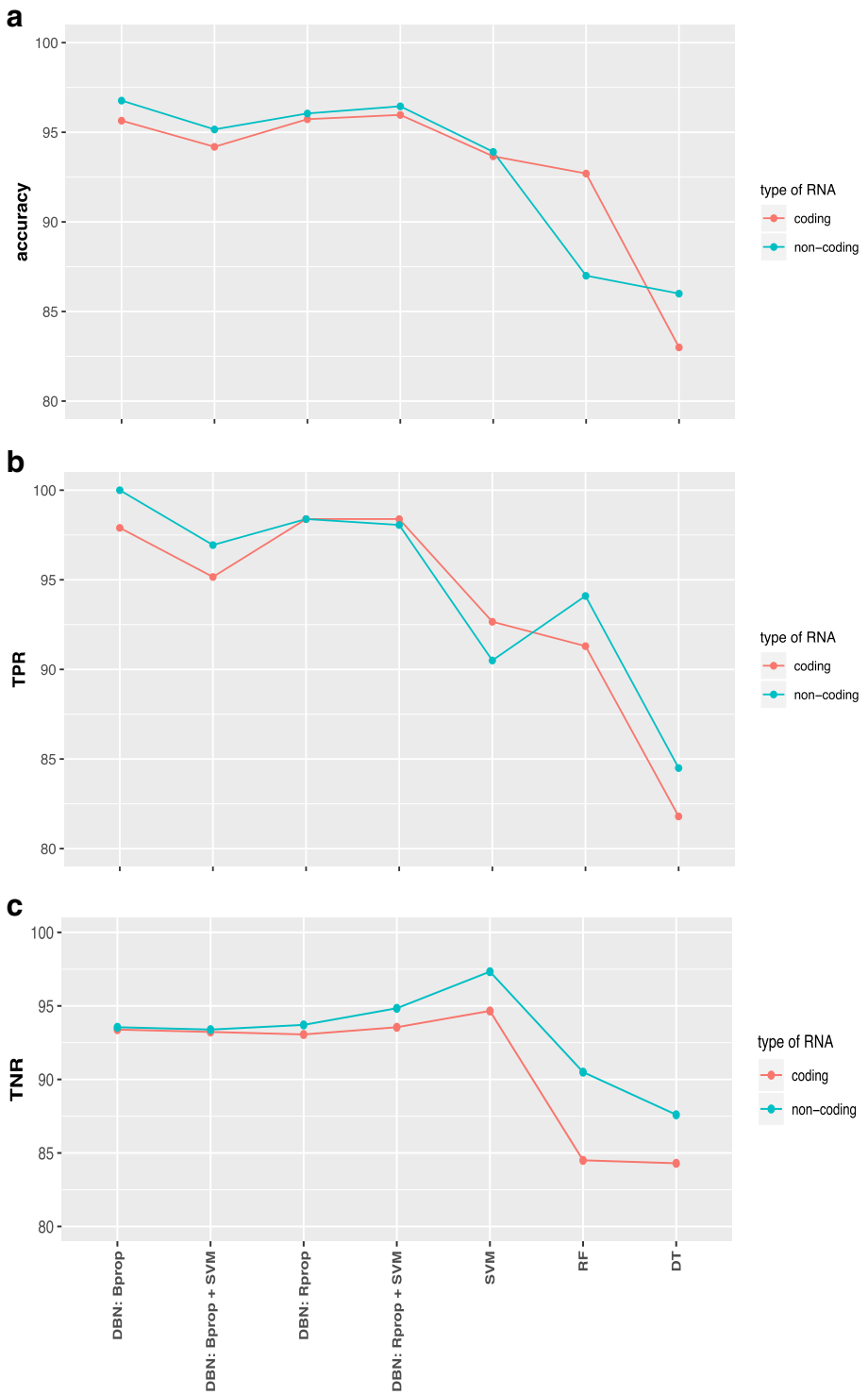

Fig. 1 Comparison of the results for different classification methods (see x-axis) and data from coding RNAs (red) and non-coding RNAs (green) for RPKM normalization. **a** Accuracy of the classification. **b** True positive rate. **c** True negative rate

the underlying disease and (II) the results are robust and do not depend on a particular classification method. From this we conclude the equivalence of predictive information as captured by non-coding RNAs and protein coding RNAs and their utility for computational diagnostics tasks in lung adenocarcinoma.

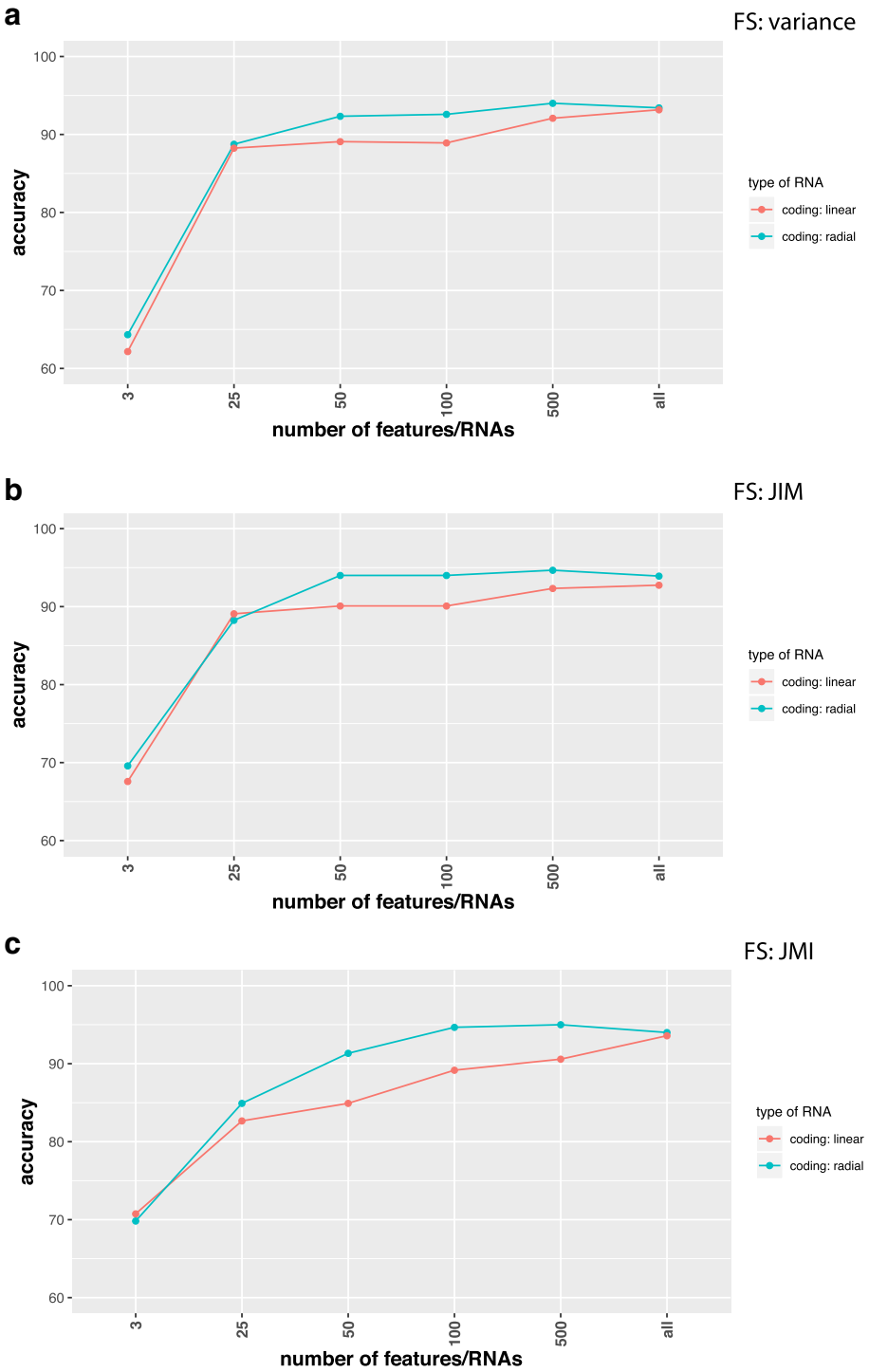

Fig. 2 Comparison of classification results for SVMs with a linear (red) and radial (green) basis kernel in dependence on the number of input features/RNAs (x-axis). Data are from coding RNAs for RPKM normalization and the label 'all' corresponds to 12360 RNAs. Feature selection methods used are **a** Variance, **b** JIM and **c** JMI

An intuitive biological explanation for this observation is given by the idea underlying the epigenetic landscape [85] or the model for transcription regulation by [86]. In both studies it has been realized that molecular mechanisms are organized by networks and these are containing feedback loops [87–89] connecting all system variables. Despite the fact that ncRNAs do not code for proteins, their activity is concerted alongside ordinary cell

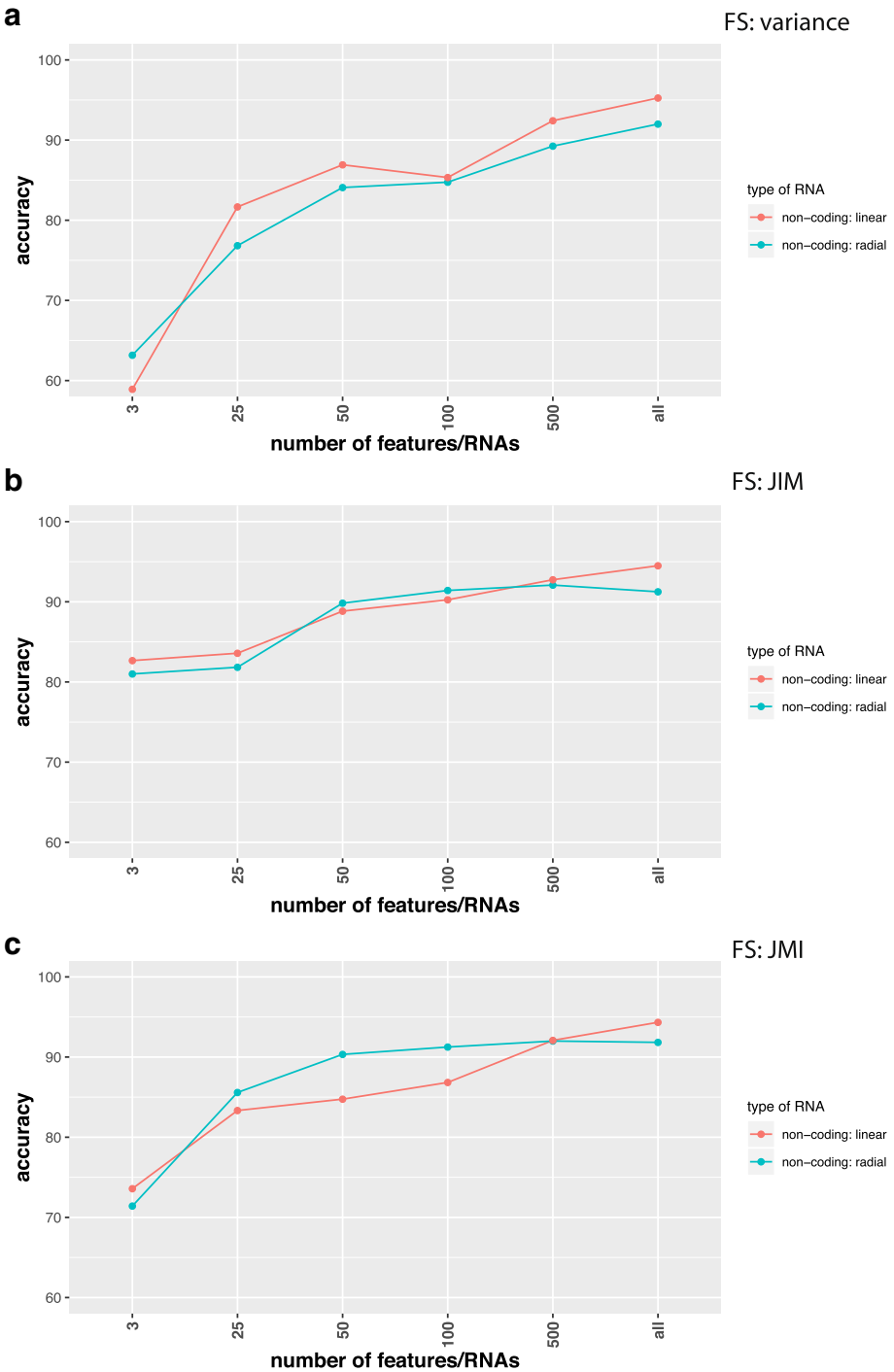

Fig. 3 Comparison of classification results for SVMs with a linear (red) and radial (green) basis kernel in dependence on the number of input features/RNAs (x-axis). Data are from non-coding RNAs for RPKM normalization and the label 'all' corresponds to 3124 RNAs. Feature selection methods used are **a** Variance, **b** JIM and **c** JMI

activities and, hence, their expression levels reflect ordinary cell functioning. What is more interesting is the fact that the signal captured by the ncRNAs is equally strong for diagnostic purposes as of coding RNAs.

In our study, we showed that for the classification of lung adenocarcinoma patients a feature selection is not beneficial, but reduces the prediction accuracy for the DNN and SVM (see Figs. 2, 3, 4, and 5). In contrast, the

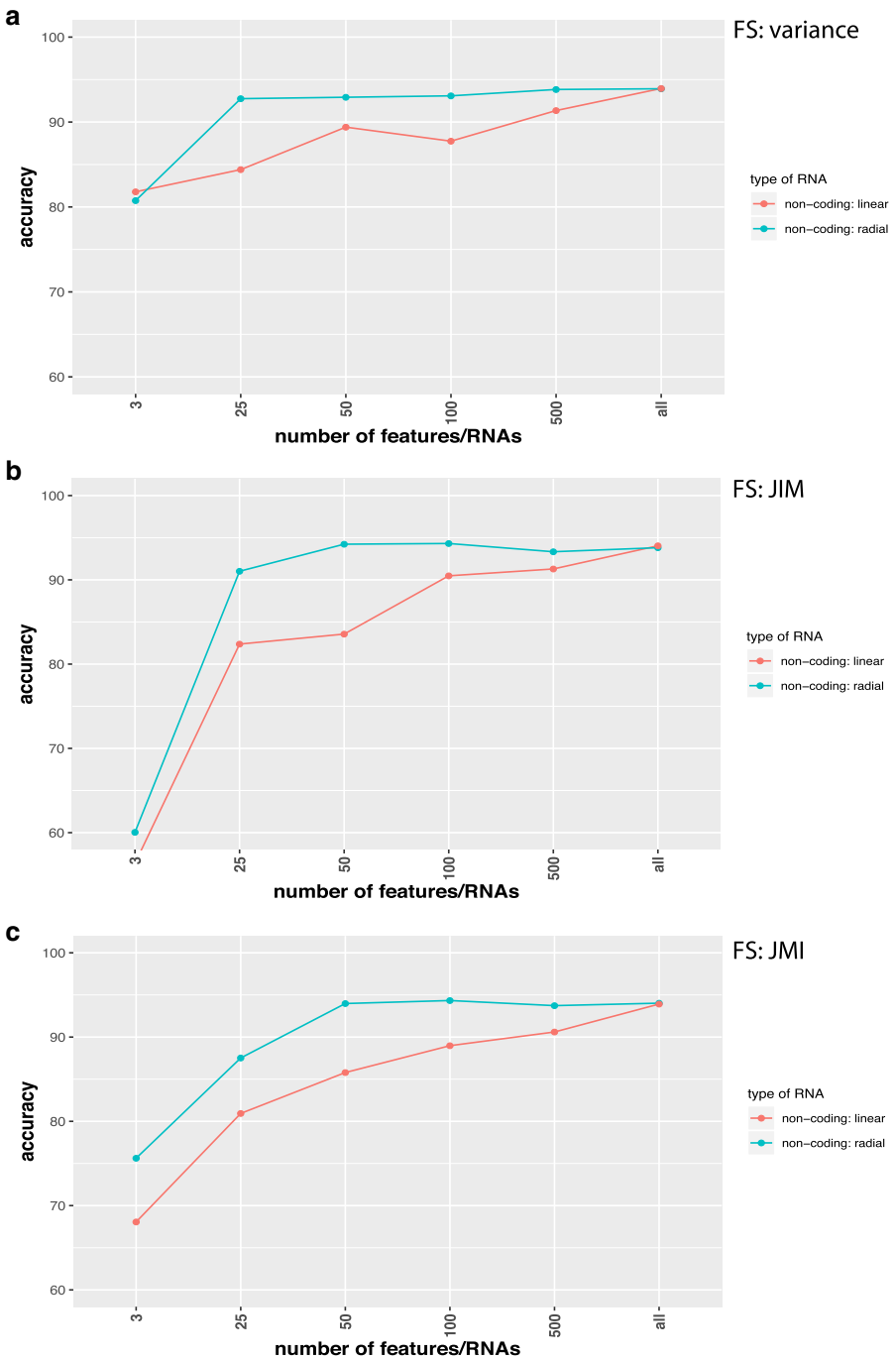

Fig. 4 Comparison of classification results for SVMs with a linear (red) and radial (green) basis kernel in dependence on the number of input features/RNAs (x-axis). Data are from non-coding RNAs for TPM normalization and the label 'all' corresponds to 1398 ncRNAs. Feature selection methods used are **a** Variance, **b** JIM and **c** JMI

best classifications were obtained without using feature selection. We demonstrated this for different feature selection methods. In this respect, we want to clarify the difference between a feature filtering and a feature selection. A filtering of features is used to remove variables (in our case either mRNAs or ncRNAs) from the analysis that do not carry any information. For instance, for biological cells it is know that not all genes are expressed in all cell types. For this reason, the corresponding mRNAs are not present in such cell types and any measured counts

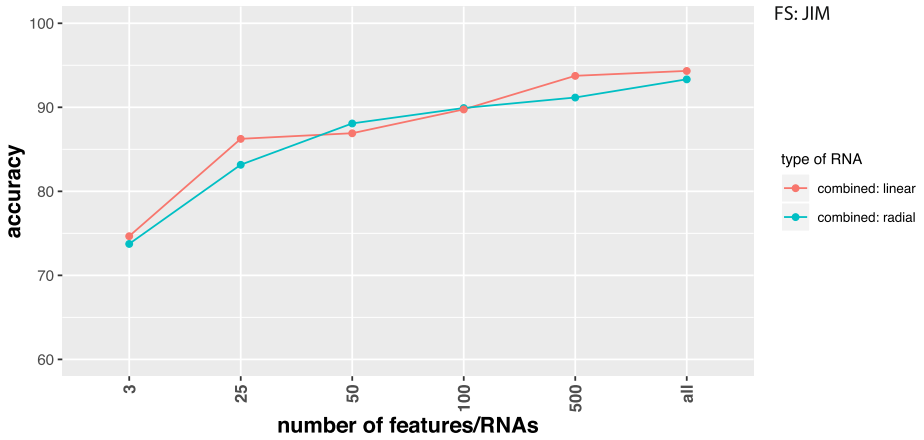

Fig. 5 Comparison of classification results for SVMs with a linear (red) and radial (green) basis kernel in dependence on the number of input features/RNAs (x-axis). Data combine coding and non-coding RNAs and the label 'all' corresponds to 15484 RNAs. Feature selection method used is JIM. Data were RPKM normalized

are due to pure noise of the high-throughput device. In order to remove such features we filtered inactive RNAs. In contrast, a feature selection operates on *valid* features, which are all supposed to carry information, and select from these a small subset to which the analysis is limited. Even for the data sets we analyzed without using any feature selection mechanism we applied a feature filtering to remove noise from the data.

For general deep learning networks one could state that these models perform also feature selection, but in an implicit manner. This is certainly true but the key point is this is part of the model itself and, hence, the feature selection and the classifier are merged with each other. Furthermore, such *higher order* features, corresponding to representations in deeper layers, are non-linear combinations of the original input features and for this reason they lost their biological interpretability. In other words, even if one would dissect such low dimensional features from the model, they would no longer correspond to a few genes or RNAs because information from all of these would be present to a certain extend. For SVMs such an implicit feature selection mechanism is less obvious.

The RNA world population is constantly expanding, e.g., new types of ncRNAs with tissue-, disease-, sex-, population origin- and race- specific expression rates were recently described [90]. These miRNAs isoform (isomiRs) were able to discriminate between 32 TCGA cancer types, probably because of the high expression specificity. It is well understood by now that ncRNAs are key regulators of physiological programs in developmental and disease states and are particularly relevant in cancer, identified as oncogenic drivers and tumor supressors for all major cancer types [91]. ncRNAs link associated genes into regulatory networks, as well as regulate each other [92]. Therefore it is plausible that ncRNAs have at least the same predictive capabilities as mRNAs.

To the best of our knowledge, we are the first to use ncRNAs beyond miRNAs for the computational classification of cancer. All previous studies limited their focus on miRNAs when classifying disease stages, e.g., [24, 78, 79]. Furthermore, we are also the first to perform a direct comparison of the classification capabilities of coding and non-coding RNAs.

Despite the popularity of general deep learning methods in the last years in many fields [41, 42], the analysis of gene expression data is so far understudied. Our results show that different variations of deep belief networks, using either Bprop or Rprop for the fine-tuning phase, lead to competitive results compared to SVMs. Also the combination of deep belief networks with SVMs is fruitful, which even showed for the coding RNAs the best results. These results are encouraging and demonstrate that, at least for large data sets, as used in our study, deep learning classifiers are capable of dealing with gene expression data.

Finally, we think it is important to emphasize that our analysis was only possible due to the general capability of DBNs and LIBSVMs [33] to deal efficiently with high-dimensional input data as a result from omitting feature selection mechanisms. This is certainly not the case for every classification method.

Conclusion

In our study, we assessed the entire information content of coding RNAs (mRNAs) and non-coding RNAs provided in RNA-seq data by studying the data with and without feature selection. Overall, from analyzing a large-scale data set from lung adenocarcinoma patients we found:

1 For the diagnostic classification, ncRNAs have as much predictive abilities as coding RNAs. These results hold for different state-of-the-art classification methods, including deep learning methods (deep belief networks), SVMs and combinations.
⇒ This may point to a new application area for ncRNAs in the computational diagnostics of lung cancer and potentially other disorders.
2 Feature selection reduces this predictive ability in both cases and the best prediction accuracy is obtained without feature selection.
⇒ This eliminates a general problem all biomarker studies suffering from, which is the definition of the optimal biomarker set [93].
3 Deep belief networks perform competitively to SVMs.
⇒ Despite this positive finding, compared to the overwhelming success of DBN for image analysis, the result differences are not large enough to claim a dominating performance.

From reviewing the literature it seems our study is the first to use ncRNAs beyond miRNAs for the computational classification of cancer. All previous studies limited their focus on miRNAs when classifying disease stages, e.g., [24, 78, 79]. Furthermore, we are also the first to perform a direct comparison of the classification capabilities of protein coding RNAs and non-coding RNAs.

For future studies, it would be interesting to expand our analysis to other cancer types and general complex disorders. It would be interesting to see if also other diseases exhibit the same behavior as we observed for lung cancer. We expect similar results to hold for other cancers but for general complex disorders predictions are more difficult.

In summary, our investigations underline the importance of general ncRNAs in understanding the complex etiology of lung cancer and suggest to conduct similar studies for other cancer types and possibly other complex disorders.

Acknowledgements
We would like to thank Shailesh Tripathi for fruitful discussions.

Authors' contributions
FES conceived the study. JS, GG, AS, MD and FES performed the analysis. All authors contributed to the preparation and the writing of the manuscript. All authors read and approved the final manuscript.

Funding
Galina Glazko was supported in part by the NIH IDeA Networks of Biomedical Research Excellence (INBRE) grant P20GM103429, and by Center for Translational Pediatric Research (CTPR) NIH Center of Biomedical Research Excellence award P20GM121293. Matthias Dehmer thanks the Austrian Science Funds for supporting this work (project P 30031). The funding bodies had no role in the design of the study.

Availability of data and materials
Access to the data is provided via Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/), accession number GSE40419.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

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

Author details
[1]Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland. [2]Turku Centre for Biotechnology, University of Turku, Turku, Finland. [3]Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, USA. [4]Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, USA. [5]Institute for Intelligent Production, Faculty for Management, University of Applied Sciences Upper Austria, Steyr, Austria. [6]Department of Mechatronics and Biomedical Computer Science, UMIT, Hall in Tyrol, Austria. [7]College of Artificial Intelligence, Nankai University, China, Tianjin, China. [8]Institute of Biosciences and Medical Technology, Tampere, Finland.

Received: 13 February 2019 Accepted: 6 November 2019
Published online: 03 December 2019

References

1. Herbst RS, Heymach JV, Lippman SM. Lung cancer. New England J Med. 2008;359(13):1367–80. https://doi.org/10.1056/NEJMra0802714. PMID: 18815398.
2. Ansorge WJ. Next-generation dna sequencing techniques. New Biotechnol. 2009;25(4):195–203.
3. Werner T. Next generation sequencing in functional genomics. Brief Bioinformatics. 2010;11(5):499–511.
4. Chen R, Snyder M. Promise of personalized omics to precision medicine. Wiley Interdiscipl Rev: Syst Biol Med. 2013;5(1):73–82.
5. Seo D, Ginsburg GS. Genomic medicine: bringing biomarkers to clinical medicine. Curr Opin Chem Biol. 2005;9(4):381–6.
6. Emmert-Streib F, Tuomisto L, Yli-Harja O. The Need for Formally Defining 'Modern Medicine' by Means of Experimental Design. Frontiers Genet. 2016;7:60. https://doi.org/10.3389/fgene.2016.00060.
7. Anastasiadou E, Jacob LS, Slack FJ. Non-coding rna networks in cancer. Nature Rev Cancer. 2018;18(1):5.
8. Cech TR, Steitz JA. The noncoding rna revolution?trashing old rules to forge new ones. Cell. 2014;157(1):77–94.
9. Fatica A, Bozzoni I. Long non-coding rnas: new players in cell differentiation and development. Nature Rev Genet. 2014;15(1):7.
10. Mercer TR, Dinger ME, Mattick JS. Long non-coding rnas: insights into functions. Nature Rev Genet. 2009;10(3):155.
11. QD Wang X, L Crutchley J, Dostie J. Shaping the genome with non-coding rnas. Curr Genomics. 2011;12(5):307–21.
12. Sacco LD, Baldassarre A, Masotti A. Bioinformatics tools and novel challenges in long non-coding rnas (lncrnas) functional analysis. Int J Mole Sci. 2011;13(1):97–114.
13. Ponting CP, Belgard TG. Transcribed dark matter: meaning or myth? Human Mole Genet. 2010;19(R2):162–8.
14. Robinson R. Dark matter transcripts: sound and fury, signifying nothing? PLoS Biol. 2010;8(5):1000370.
15. Managadze D, Rogozin IB, Chernikova D, Shabalina SA, Koonin EV. Negative correlation between expression level and evolutionary rate of long intergenic noncoding rnas. Genome Biol Evol. 2011;3:1390–1404.
16. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncrnadb: a reference database for long noncoding rnas. Nucleic Acids Res. 2010;39(suppl_1):146–151.
17. Moran VA, Perera RJ, Khalil AM. Emerging functional and mechanistic paradigms of mammalian long non-coding rnas. Nucleic Acids Res. 2012;40(14):6391–400.
18. Carninci P, Kasukawa T, Katayama S, Gough J, Frith M, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. The transcriptional landscape of the mammalian genome. Science. 2005;309(5740):1559–63.
19. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, et al. Rna maps reveal new rna classes and a possible function for pervasive transcription. Science. 2007;316(5830):1484–8.
20. Esteller M. Non-coding rnas in human disease. Nature Rev Genet. 2011;12(12):861.

21. Palazzo AF, Lee ES. Non-coding rna: what is functional and what is junk? Front Genet. 2015;6:2.
22. Mattick JS. The genetic signatures of noncoding rnas. PLoS Genet. 2009;5(4):1000459.
23. Glazko GV, Zybailov BL, Rogozin IB. Computational prediction of polycomb-associated long non-coding rnas. PLoS ONE. 2012;7(9):44878.
24. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, Stephens RM, Okamoto A, Yokota J, Tanaka T, et al. Unique microrna molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell. 2006;9(3):189–98.
25. Su X, Malouf GG, Chen Y, Zhang J, Yao H, Valero V, Weinstein JN, Spano J-P, Meric-Bernstam F, Khayat D, et al. Comprehensive analysis of long non-coding rnas in human breast cancer clinical subtypes. Oncotarget. 2014;5(20):9864.
26. Li R, Qian J, Wang Y-Y, Zhang J-X, You Y-P. Long noncoding rna profiles reveal three molecular subtypes in glioma. CNS Neurosci Therapeu. 2014;20(4):339–43.
27. Flippot R, Malouf GG, Su X, Mouawad R, Spano J-P, Khayat D. Cancer subtypes classification using long non-coding rna. Oncotarget. 2016;7(33):54082.
28. Seo J-S, Ju YS, Lee W-C, Shin J-Y, Lee JK, Bleazard T, Lee J, Jung YJ, Kim J-O, Shin J-Y, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res. 2012;22:2109–19.
29. Cestarelli V, Fiscon G, Felici G, Bertolazzi P, Weitschek E. Camur: Knowledge extraction from rna-seq cancer data through equivalent classification rules. Bioinformatics. 2015;32(5):697–704.
30. Guo Y, Liu S, Li Z, Shang X. BCDForest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. BMC Bioinformatics. 2018;19(5):118.
31. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18(7):1527–54.
32. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.
33. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol. 2011;2:27–12727. http://www.csie.ntu.edu.tw/~cjlin/libsvm.
34. Weitschek E, Di Lauro S, Cappelli E, Bertolazzi P, Felici G. Camurweb: a classification software and a large knowledge base for gene expression data of cancer. BMC Bioinformatics. 2018;19(10):245.
35. Minsky M, Papert S. Perceptrons. Cambridge: MIT Press; 1969.
36. Crick F. The recent excitement about neural networks. Nature. 1989;337:129–32.
37. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Nat Acad Sci USA. 1982;79:2554–8.
38. Emmert-Streib F. Active learning in recurrent neural networks facilitated by an hebb-like learning rule with memory. Neural Inf Process - Lett Rev. 2005;9(2):31–40.
39. Emmert-Streib F. A heterosynaptic learning rule for neural networks. Int J Modern Phys C. 2006;17(10):1501–20.
40. Rosenblatt F. The Perceptron, a Perceiving and Recognizing Automaton Project Para: Cornell Aeronautical Laboratory; 1957.
41. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
42. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks: Curran Associates, Inc; 2012, pp. 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.
43. Graves A, Mohamed A, Hinton GE. Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013;abs/1303.5778:. https://doi.org/10.1109/icassp.2013.6638947.
44. Leung MKK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. Bioinformatics. 2014;30(12):121–9.
45. Zhang S, Zhou J, Hu H, Gong H, Chen L, Cheng C, Zeng J. A deep learning framework for modeling structural features of rna-binding protein targets. Nucleic Acids Res. 2015;43(20):e32.
46. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. Nat Biotechnol. 2015;33:831–8.
47. Fakoor R, Ladhak F, Nazi A, Huber M. Using deep learning to enhance cancer diagnosis and classification. In: Proceedings of the International Conference on Machine Learning, vol. 28; 2013.
48. Stupnikov A, Tripathi S, de Matos Simoes R, McArt D, Salto-Tellez M, Glazko G, Emmert-Streib F. samExploreR: Exploring reproducibility and robustness of RNA-seq results based on SAM files. Bioinformatics. 2016;32:475.
49. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2010;39:19–21.
50. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nature Methods. 2012;9(4):357–9.
51. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. The ucsc genome browser database: 2014 update. Nucleic Acids Res. 2014;42(D1):764–770.
52. Liao Y, Smyth GK, Shi W. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013656. https://academic.oup.com/bioinformatics/article/30/7/923/232889.
53. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, et al. A survey of best practices for rna-seq data analysis. Genome Biol. 2016;17(1):13.
54. Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. lncrnadb v2. 0: expanding the reference database for functional long noncoding rnas. Nucleic Acids Res. 2014;43(D1):168–73.
55. Emmert-Streib F, Moutari S, Dehmer M. A comprehensive survey of error measures for evaluating binary decision making in data science. Wiley Interdiscipl Rev: Data Mining Knowl Disc. 20191303. https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1303.
56. Webb AR, Copsey KD. Statistical Pattern Recognition, 3rd. Rochelle Park: Wiley; 2011.
57. Bradley AP. The use of the area under the roc curve in the evaluation of machine learning algorithms. Patt Recogn. 1997;30(7):1145–59.
58. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intell Data Anal. 2002;6(5):429–49.
59. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005;21(15):3301–07.
60. Emmert-Streib F, Dehmer M. Evaluation of regression models: Model assessment, model selection and generalization error. Mach Learn Knowl Extract. 2019;1(1):521–51.
61. Yoshua B. Learning deep architectures for ai. Foundations Trends Mach Learn. 2009;2(1):1–127. https://doi.org/10.1561/2200000006.
62. Fischer A, Igel C. An introduction to restricted boltzmann machines. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer; 2012. p. 14–36. http://image.diku.dk/igel/paper/AItRBM-proof.pdf.
63. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–7.
64. Riedmiller M, Braun H. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In: Neural Networks, 1993., IEEE International Conference On. IEEE; 1993. p. 586–91. https://doi.org/10.1109/icnn.1993.298623.
65. Igel C, Hüsken M. Improving the rprop learning algorithm. In: Proceedings of the Second International ICSC Symposium on Neural Computation (NC 2000), vol. 2000. Citeseer; 2000. p. 115–21.
66. Drees M. Darch: Package for Deep Architectures and Restricted-Bolzmann-Machines. The Comprehensive R Archive Network (CRAN). 2014. The Comprehensive R Archive Network (CRAN). Version 0.9.1. https://cran.fhcrc.org/web/packages/darch/index.html.
67. Salakhutdinov R, Hinton GE. Deep boltzmann machines. In: International Conference on Artificial Intelligence and Statistics; 2009. p. 448–55.
68. Hinton G. Where do features come from? Cognitive Sci. 2014;38(6):1078–101.
69. Zhao J, Cheng W, He X, Liu Y, Li J, Sun J, Li J, Wang F, Gao Y. Construction of a specific SVM classifier and identification of molecular markers for lung adenocarcinoma based on lncrna-mirna-mRNA network. OncoTargets Therapy. 2018;11:3129.
70. Fan Z, Xue W, Li L, Zhang C, Lu J, Zhai Y, Suo Z, Zhao J. Identification of an early diagnostic biomarker of lung adenocarcinoma based on co-expression similarity and construction of a diagnostic model. J Trans Med. 2018;16(1):205.
71. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics. 2008;9(1):13.
72. Salem H, Attiya G, El-Fishawy N. Gene expression profiles based human cancer diseases classification. In: Computer Engineering Conference (ICENCO), 2015 11th International. IEEE; 2015. p. 181–7. https://doi.org/10.1109/icenco.2015.7416345.

73. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics. 2005;21(20):3896–904.
74. Wei X, Li K-C. Exploring the within-and between-class correlation distributions for tumor classification. Proc Nat Acad Sci. 2010;107(15): 6737–42.
75. Wang X. Robust two-gene classifiers for cancer prediction. Genomics. 2012;99(2):90–5.
76. Liu J, Wang X, Cheng Y, Zhang L. Tumor gene expression data classification via sample expansion-based deep learning. Oncotarget. 2017;8(65):109646.
77. Roffo G, Melzi S, Cristani M. Infinite feature selection. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 4202–10.
78. Xue Z, Wen J, Chu X, Xue X. A microrna gene signature for identification of lung cancer. Surg Oncol. 2014;23(3):126–31.
79. Volinia S, Calin GA, Liu C-G, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, et al. A microrna expression signature of human solid tumors defines cancer gene targets. Proc Nat Acad Sci. 2006;103(7):2257–61.
80. Telonis AG, Magee R, Loher P, Chervoneva I, Londin E, Rigoutsos I. Knowledge about the presence or absence of mirna isoforms (isomirs) can successfully discriminate amongst 32 tcga cancer types. Nucleic Acids Res. 2017;45(6):2973–85.
81. Seow N, Fenati RA, Connolly AR, Ellis AV. Hi-fidelity discrimination of isomiRs using G-quadruplex gatekeepers. PloS one. 2017;12(11):0188163.
82. Brown G, Pocock A, Zhao M-J, Luján M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res. 2012;13(Jan):27–66.
83. Dash M, Liu H. Feature selection for classification. Intell Data Anal. 1997;1(3):131–56.
84. Yang HH, Moody J. Data visualization and feature selection: New algorithms for nongaussian data. In: Advances in Neural Information Processing Systems; 2000. p. 687–93.
85. Waddington CH. The Strategy of the Genes. New York: Geo, Allen Unwin, London; 1957.
86. Kauffman SA. Metabolic stability and epigenesis in randomly constructed genetic nets. J Theoret Biol. 1969;22:437–67.
87. Becskei A, Séraphin B, Serrano L. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. EMBO J. 2001;20(10):2528–35.
88. Chen Y-R, Huang H-C, Lin C-C. Regulatory feedback loops bridge the human gene regulatory network and regulate carcinogenesis. Brief Bioinforma. 2017.
89. Herranz H, Cohen SM. Micrornas and gene regulatory networks: managing the impact of noise in biological systems. Genes Dev. 2010;24(13):1339–44.
90. Telonis AG, Loher P, Jing Y, Londin E, Rigoutsos I. Beyond the one-locus-one-mirna paradigm: microrna isoforms enable deeper insights into breast cancer heterogeneity. Nucleic Acids Res. 2015;43(19):9158–75.
91. Anastasiadou E, Faggioni A, Trivedi P, Slack FJ. The nefarious nexus of noncoding rnas in cancer. Int J Mole Sci. 2018;19(7):. https://doi.org/10.20944/preprints201803.0187.v1.
92. Yamamura S, Imai-Sumida M, Tanaka Y, Dahiya R. Interaction and cross-talk between non-coding rnas. Cell Mole Life Sci. 20171–18. https://link.springer.com/article/10.1007/s00018-017-2626-6.
93. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol. 2011;7(10):1002240.

Publisher's Note