# Novel human lncRNA–disease association inference based on lncRNA expression profiles

Xing Chen[1,2,*] and Gui-Ying Yan[1,2,*]

[1]National Center for Mathematics and Interdisciplinary Sciences and [2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, P.R. China

Associate Editor: Martin Bishop

## ABSTRACT

**Motivation:** More and more evidences have indicated that long–non-coding RNAs (lncRNAs) play critical roles in many important biological processes. Therefore, mutations and dysregulations of these lncRNAs would contribute to the development of various complex diseases. Developing powerful computational models for potential disease-related lncRNAs identification would benefit biomarker identification and drug discovery for human disease diagnosis, treatment, prognosis and prevention.

**Results**: In this article, we proposed the assumption that similar diseases tend to be associated with functionally similar lncRNAs. Then, we further developed the method of Laplacian Regularized Least Squares for LncRNA–Disease Association (LRLSLDA) in the semisupervised learning framework. Although known disease–lncRNA associations in the database are rare, LRLSLDA still obtained an AUC of 0.7760 in the leave-one-out cross validation, significantly improving the performance of previous methods. We also illustrated the performance of LRLSLDA is not sensitive (even robust) to the parameters selection and it can obtain a reliable performance in all the test classes. Plenty of potential disease–lncRNA associations were publicly released and some of them have been confirmed by recent results in biological experiments. It is anticipated that LRLSLDA could be an effective and important biological tool for biomedical research.

**Availability:** The code of LRLSLDA is freely available at http://asdcd.amss.ac.cn/Software/Details/2.

**Contact:** xingchen@amss.ac.cn or yangy@amt.ac.cn

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

## 1 INTRODUCTION

It is well known that genetic information is stored in protein-coding genes, which are referred to as the central dogma of molecular biology (Crick *et al.*, 1961; Yanofsky, 2007). Therefore, RNA is only considered to be an intermediary between a DNA sequence and its encoded protein during a considerable long period (Mattick and Makunin, 2006). Recent studies have shown that protein-coding genes account for only a small fraction of the human genome (~1.5%). In other words, >98% of the human genome does not encode protein sequences (Bertone *et al.*, 2004; Birney *et al.*, 2007; Carninci *et al.*, 2006; Claverie,

2005; Core *et al.*, 2008; Kapranov *et al.*, 2007; Lander *et al.*, 2001; Taft *et al.*, 2010; Wilusz *et al.*, 2009). Especially, it has been observed that the proportion of non–protein-coding sequence increases with the complexity of organisms (Taft *et al.*, 2007). These facts challenge forementioned traditional view of RNA. Furthermore, accumulating evidences have shown that non-coding RNAs (ncRNAs) normally play a critical role in various biological processes. Specially, long–non-coding RNAs (lncRNAs) are a class of important ncRNAs with the length >200 nt (Kapranov *et al.*, 2007; Mercer *et al.*, 2009; Wapinski and Chang, 2011). In the past few years, increasing number of lncRNAs have been discovered in eukaryotic organisms ranging from nematodes to humans with the rapid development of both experimental technology and computational methods (Amaral *et al.*, 2011). It has also been shown that the expression levels of lncRNAs appear to be lower than protein-coding genes (Babak *et al.*, 2005; Bono *et al.*, 2003; Gibb *et al.*, 2011; Guttman *et al.*, 2010; Ramsköld *et al.*, 2009), and some lncRNAs have high tissue specificity (Gibb *et al.*, 2011; Mercer *et al.*, 2008; Pauli *et al.*, 2012; Ponting *et al.*, 2009).

Accumulating evidences have indicated that plenty of lncRNAs play critical roles in many important biological processes, including transcription, translation, splicing, differentiation, epigenetic regulation, immune responses, cell cycle control and so on (Bu *et al.*, 2012; Chen *et al.*, 2013; Lander *et al.*, 2001; Managadze *et al.*, 2011; Mattick, 2009; Mattick and Makunin, 2006; Mitchell Guttman *et al.*, 2009; Qureshi *et al.*, 2010; Wapinski and Chang, 2011; Wilusz *et al.*, 2009). Therefore, mutations and dysregulations of lncRNAs are associated with a broad range of human diseases (Mercer *et al.*, 2009; Ponting *et al.*, 2009; Taft *et al.*, 2010; Wapinski and Chang, 2011), such as cancers (Chung *et al.*, 2011; Gupta *et al.*, 2010; Spizzo *et al.*, 2012; van Poppel *et al.*, 2011; Yang *et al.*, 2011; Zhang *et al.*, 2012), cardiovascular diseases (Congrains *et al.*, 2011) and neurodegeneration diseases (Johnson, 2011). For example, lncRNA HOTAIR, PCA3 and UCA1 have been treated as potential biomarker of hepatocellular carcinoma recurrence (Yang *et al.*, 2011), prostate cancer aggressiveness (van Poppel *et al.*, 2011) and bladder cancer diagnosis, respectively (Zhang *et al.*, 2012). Therefore, identifying potential human disease-related lncRNAs can facilitate not only the understanding of molecular mechanisms of human disease at lncRNA level, but also biomarker identification for human disease diagnosis, treatment, prognosis and prevention (Chen *et al.*, 2013). So far, plenty of studies have generated a large amount of lncRNA-related biological data about sequence, expression, function and so on. These datasets

---

*To whom correspondence should be addressed.

have been stored in some publicly available databases, such as NRED (Dinger *et al.*, 2009), lncRNAdb (Amaral *et al.*, 2011), NONCODE (Bu *et al.*, 2012). However, only relatively few lncRNA–disease associations have been reported. Developing powerful computational models based on these datasets to predict potential disease–lncRNA associations on a large scale has been treated as one of the most important topics of lncRNAs and diseases. Computational model can provide most promising lncRNA–disease associations for further experimental validation, hence decrease the time and cost of biological experiments.

In the previous work, we have manually collected experimentally reported disease–lncRNA associations and constructed the first lncRNA–disease association database, LncRNADisease (http://cmbi.bjmu.edu.cn/lncrnadisease) (Chen *et al.*, 2013). This database has included >480 lncRNA–disease associations, ~208 lncRNAs and 166 diseases and laid the solid data fundament for lncRNA-related predictive research. Furthermore, we obtained an important conclusion that lncRNAs tend to be related to the same disease as their genomic neighbor genes and developed a simple lncRNA–disease association prediction method based on the genomic context of a given lncRNA (Chen *et al.*, 2013). The conclusion obtained in this study laid the solid theoretical fundament for disease–lncRNA association prediction research. Based on this conclusion, various disease-related lncRNA prediction methods can be developed.

In this article, we logically extended the basic assumption in the previous disease-microRNA (miRNA) association prediction research (Chen *et al.*, 2012a, b) and proposed the following assumption for disease-related lncRNA prediction: similar diseases tend to be associated with functionally similar lncRNAs. Based on this assumption and the fact that selecting lncRNAs that are not related to the given disease is currently difficult or even impossible, we developed a computational model of Laplacian Regularized Least Squares for LncRNA–Disease Association (LRLSLDA) in the semisupervised learning framework. This method prioritizes the entire lncRNAome for disease of interest by integrating known phenome-lncRNAome network obtained from the database of LncRNADisease, disease similarity network and lncRNA similarity network. LRLSLDA is a global approach that can rank candidate disease–lncRNA pairs for all the diseases simultaneously. In the leave-one-out cross validation (LOOCV), LRLSLDA obtained the reliable AUC of 0.7760, demonstrating superiority performance of LRLSLDA to previous methods and potential value for disease-related lncRNA prediction and biomarker detection in the diagnosis, treatment, prognosis and prevention of human disease. We also classified test samples of lncRNA–disease associations into distinct classes, and LRLSLDA obtained reliable performance in different test classes. Plenty of potential disease–lncRNA associations were publicly released for experimental verification. Some of the associations have been confirmed by recent results in biological experiments.

## 2 MATERIALS

### 2.1 LncRNA–disease associations

We downloaded known lncRNA–disease association dataset from the LncRNADisease database in October, 2012. This dataset is used as gold standard dataset in the cross validation and training dataset in potential disease–lncRNA association prediction. After getting rid of duplicate associations, 293 distinct experimentally confirmed lncRNA–disease associations were obtained, including 118 lncRNAs and 167 diseases (Supplementary Table S1). We denoted variable *nl* as the number of lncRNAs, *nd* as the number of diseases, matrix *A* as the adjacency matrix of lncRNA–disease associations, where *A(i,j)* in row *i* column *j* is 1 if lncRNA *l(j)* is related to the disease *d(i)*, otherwise 0.

### 2.2 LncRNA expression similarity

Considering the current situation that comprehensive expression data of lncRNA is still unavailable and the fact that long intergenic non-coding RNA (lincRNA) accounts for a large fraction of the whole lncRNA set, we downloaded lincRNA expression profiles from UCSC Genome Bioinformatics (http://genome.ucsc.edu/) in October, 2012, including the expression profiles of 21 626 lincRNAs in 22 human tissues or cell types (Supplementary Table S2). Then, we defined the lincRNA expression similarity as the Spearman correlation coefficient between the expression profiles of each lincRNA pair. Matrix *SPC* is denoted as the lncRNA expression similarity matrix, where *SPC(i,j)* in row *i* column *j* is the expression similarity between lncRNA *l(i)* and *l(j)* if they are both lincRNA, otherwise 0. LRLSLDA developed in this article can be applied to interactions prediction between all the lncRNAs (not only lincRNAs, but also other members of lncRNA) and diseases by integrating lncRNA–disease association data and lncRNA expression data. Making use of Spearman correlation coefficient between the expression profiles of each pair is the general method in bioinformatics research. Hence, it is likely that this similarity measure would still obtain reliable performance for lncRNA expression data, as already shown for lincRNAs in this article.

## 3 METHODS

### 3.1 Gaussian interaction profile kernel similarity for diseases

Based on the assumption that similar diseases tend to show a similar interaction and non-interaction pattern with the lncRNAs, we constructed Gaussian interaction profile kernel similarity for disease from known lncRNA–disease associations [motivated by van Laarhoven *et al.* (2011)]. The procedures of Gaussian interaction profile kernel similarity have been illustrated in Figure 1. Firstly, we denoted the interaction profile *IP(d(i))* of disease *d(i)* as the binary vector encoding the presence or absence of association between disease *d(i)* and each lncRNA in the known disease–lncRNA association dataset, i.e. the *ith* row of the adjacency matrix *A*. Then, we introduced Gaussian kernel for the interaction profiles of diseases. Kernel for disease *d(i)* and *d(j)* was defined as follows and used as the similarity score between these two diseases.

$$KD(d(i), d(j)) = exp(-\gamma_d \parallel IP(d(i)) - IP(d(j)) \parallel^2)$$

where the parameter $\gamma_d$ controls the kernel bandwidth. It was normally defined as a new bandwidth parameter $\gamma'_d$ normalized by the average number of associations with lncRNA per disease. Although this new bandwidth parameter can be better selected through further cross validation, here for simplicity we set $\gamma'_d = 1$ according to the choice in the
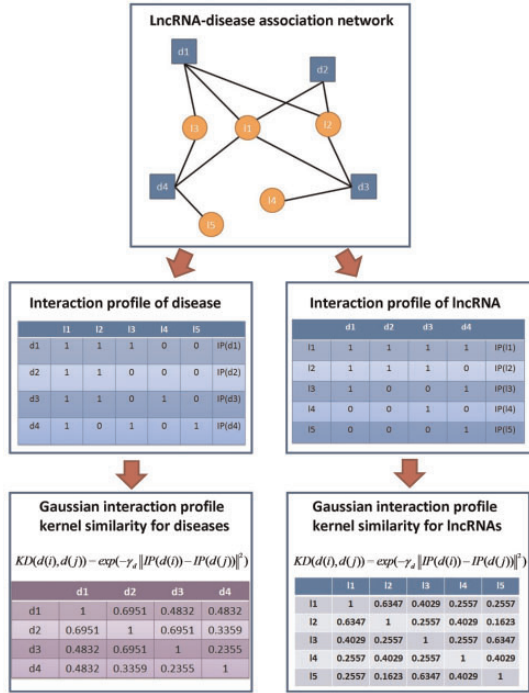
**Fig. 1.** The procedures of Gaussian interaction profile kernel similarity calculation based on known disease–lncRNA association network have been divided into three steps: constructing known disease–lncRNA association network; obtain interaction profiles of diseases and lncRNAs, respectively; calculate Gaussian interaction profile kernel similarity for diseases and lncRNAs, respectively. Similarity used in the table of bottom panels is Gaussian interaction profile kernel similarity for diseases and lncRNAs, respectively

previous work (van Laarhoven *et al.*, 2011). The formula for the calculation of $\gamma_d$ is

$$\gamma_d = \gamma'_d / (\frac{1}{nd} \sum_{i=1}^{nd} \| IP(d(i)) \|^2)$$

Finally, *KD* is denoted as Gaussian interaction profile kernel similarity matrix for diseases, where the entity *KD(i,j)* in row *i* column *j* is the Gaussian interaction profile kernel for disease *d(i)* and *d(j)*. From relevant research (Vanunu *et al.*, 2010), we could obtain the conclusion that disease similarity after logistic function transformation can improve predictive accuracy of disease-related problems. Therefore, we used the logistic function in the previous study (Vanunu *et al.*, 2010) as follows:

$$SD(d(i), d(j)) = \frac{1}{1 + e^{c \cdot KD(d(i), d(j)) + d}}$$

For the two parameters contained in this formula, we adopt the same parameter selection as previous study (Vanunu *et al.*, 2010), i.e. $c = -15, d = \log(9999)$. Disease similarity matrix *SD* will be used in the optimal classifier construction in the following sections.

## 3.2 Gaussian interaction profile kernel similarity for lncRNAs

Gaussian interaction profile kernel similarity matrix for lncRNAs, *KL*, can be constructed in a similar way as follows:

$$KL(l(i), l(j)) = exp(-\gamma_l \| IP(l(i)) - IP(l(j)) \|^2)$$

where $IP(l(i))$ for lncRNA *l(i)* is the binary vector encoding the presence or absence of association between lncRNA *l(i)* and each disease and $\gamma_l$ controls the kernel bandwidth, which can be obtained as follows by normalizing a new bandwidth parameter $\gamma'_l$ [$\gamma'_l = 1$, according to the choice in the previous work (van Laarhoven *et al.*, 2011)].

$$\gamma_l = \gamma'_l / (\frac{1}{nl} \sum_{i=1}^{nl} \| IP(l(i)) \|^2)$$

Based on the forementioned lncRNA expression similarity and Gaussian interaction profile kernel similarity, we constructed lncRNA integrated similarity matrix *SL*, where the entity *SL(i,j)* in row *i* column *j* is integrated similarity between lncRNA *l(i)* and l *(j)* defined as follows and *ew* is the weight coefficient of lncRNA expression similarity.

$$SL(i,j) = \begin{cases} ew \cdot SPC(i,j) + (1-ew) \cdot KL(i,j) & \text{both} l(i) \text{ and } l(j) \text{ are lincRNAs} \\ KL(i,j) & otherwise \end{cases}$$

## 3.3 Laplacian Regularized Least Squares for LncRNA−Disease Association

Based on the underlying assumption that similar diseases tend to be associated with similar lncRNAs, here we developed the method of LRLSLDA to predict the potential related lncRNAs for the disease of interest. The flowchart of LRLSLDA has been shown in Figure 2, including the steps of similarity calculation and Laplacian normalization, cost function construction and optimal classifier function calculation, and optimal classifier function combination. Based on the framework of Laplacian Regularized Least Squares (LapRLS), aforementioned assumption will be formulated into two classifiers in the disease space and lncRNA space, respectively. Then these two classifiers will be combined into a single classifier by a simple mean operation to give final prediction about disease–lncRNA association probability. It is anticipated that this classifier would be a continuous classification function, which could reflect the probability that each lncRNA is associated with all the diseases of interest. Hence, following two criterions were used to evaluate constructed classifier: (i) the classifier should comply with known lncRNA–disease associations as accurately as possible; (ii) the classifier should be smooth over disease space and lncRNA space, i.e. the scores for the potential association between similar lncRNAs (diseases) and the same disease (lncRNA) should be similar, which reflect aforementioned basic assumption. Candidate lncRNA–disease pairs with high scores will be expected to have a high priority for biological experiments validation. In this way, we could dramatically reduce the costs and time for potential lncRNA–disease association identification.

In the framework of LapRLS, Laplacian operation will be firstly implemented to normalize the similar matrix used in the classifier construction as follows:

$$LD = (DD)^{-1/2}(DD - SD)(DD)^{-1/2}$$

$$LL = (DL)^{-1/2}(DL - SL)(DL)^{-1/2}$$

where the diagonal matrices *DD* and *DL* are defined such that *DD(i,i)* and *DL(i,i)* are the sum of the *ith* row of *SD* and *SL*, respectively. Then, cost functions will be defined in lncRNA space and disease space, respectively. The optimal classifier meeting above criterions will be obtained by minimizing this cost function. In the lncRNA space, optimal classifier can be obtained by solving the following optimization problem:

$$\min_{FL}[\| A^T - FL \|_F^2 + \eta L \| FL \cdot LL \cdot FL^T \|_F^2]$$

where $\| \cdot \|_F$ is the Frobenius norm and $\eta L$ is the trade-off parameter. We solved this optimization problem by calculating the derivative of this
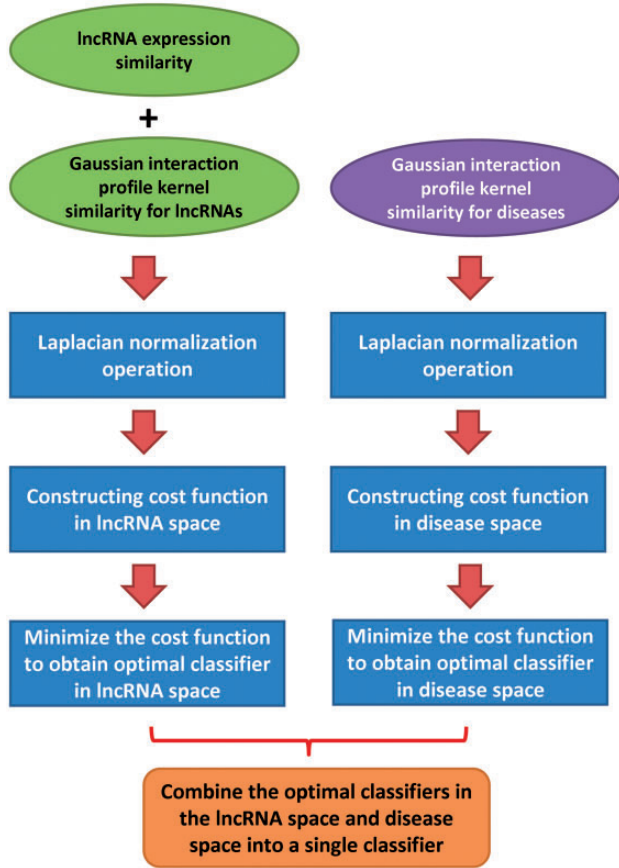
**Fig. 2.** The flowchart of LRLSLDA is shown here, including the basic steps to predict potential disease-related lncRNAs based on LRLSLDA

objective function (Belkin *et al.*, 2006; Xia *et al.*, 2010). The optimal classification function can be obtained as follows:

$$FL^* = SL(SL + \eta L \cdot LL \cdot SL)A^T$$

We can also obtain the optimal classification function in the disease space in a similar way by solving the following optimization problem:

$$\min_{FD}[\|A - FD\|_F^2 + \eta D\|FD \cdot LD \cdot FD^T\|_F^2]$$

$$FD^* = SD(SD + \eta D \cdot LD \cdot SD)A$$

where $\eta D$ is also the trade-off parameter. According to the choice in the previous work (van Laarhoven *et al.*, 2011), we set $\eta L = 1, \eta D = 1$.

Finally, we combined the optimal classifiers in the lncRNA space and disease space into a single classifier by a simple mean operation:

$$F^* = lw \cdot FL^{*T} + (1 - lw) \cdot FD^*$$

where $lw$ is denoted as the weight coefficient of the classification function in the lncRNA space and the entity $F^*(i,j)$ in row $i$ column $j$ reflect the probability that lncRNA $l(j)$ is related to the disease $d(i)$.

## 4 RESULTS

### 4.1 Leave-one-out cross validation

LOOCV was implemented on the known experimentally verified lncRNA–disease associations to evaluate the performance of

LRLSLDA. Here, we combined the lncRNA expression similarity and Gaussian interaction profile kernel similarity matrix for lncRNAs into the integrated similarity by a simple mean operation according to the previous studies (Chen *et al.*, 2012a, c) and will discuss the parameter effect on the predictive performance in the next section. For the weight coefficients in the final optimal classifier function combination, we implemented similar average operation for classifier combination according to previous successful studies about drug-target prediction and disease-related miRNA environmental factor (EF) interaction prediction (Chen *et al.*, 2012a; Xia *et al.*, 2010). It has been observed from Figure 3 that combined classifier can significantly improve the predictive accuracy of classifier in the single space. Also we will discuss whether this parameter selection would have a great influence on the predictive performance in the next section. To our knowledge, this is the first work making use of known lncRNA–disease associations to predict potential ones. Therefore, no previous methods can be compared with our method. We will compare LRLSLDA with the predictive result of classifiers in the single space to show the reasonability of combining classifiers in different spaces into final predictive results.

Because there were 167 diseases and 293 disease–lncRNA associations in the known golden standard dataset, i.e. less than two associations per disease, it is inappropriate and infeasible to implement LOOCV for a given disease $d$. Further, taking into account the fact that LRLSLDA is a global method (i.e. it can prioritize candidate lncRNAs for all the disease simultaneously and can compare the scores of different lncRNA–disease pairs), we implemented LOOCV for all the diseases simultaneously. We left out each known disease–lncRNA association in turn as test sample and further evaluate how well this association was ranked relative to the candidate samples. Here, all other known disease–lncRNA associations were regarded as training samples, and all the disease–lncRNA pairs without confirmed associations were regarded as candidate samples. Receiver-operating characteristics (ROC) curve was used to evaluate the predictive performance, which plots true-positive rate (TPR, sensitivity) versus false-positive rate (FPR, 1-specificity) at different rank cutoffs. Here, sensitivity means the percentage of the left-out associations obtaining the ranking higher than a given rank cutoff; Specificity means the percentage of candidate associations obtaining the ranking lower than this given rank cutoff. When we vary the rank cutoffs of successful prediction, we can obtain the corresponding TPR and FPR. In this way, ROC was drawn and AUC was calculated.

The similarity measure defined in this article relies on Gaussian interaction profile kernel similarity, which is calculated from known disease–lncRNA associations. When LOOCV was implemented, each known disease–lncRNA association was considered as test sample in turn. Therefore, we would obtain different similarity matrice because of different training samples in the each step of LOOCV (training samples were different when different known associations were considered as test samples, hence obtaining different similarity matrice). Therefore, it is necessary to recalculate Gaussian interaction profile kernel similarity for disease and lncRNA (not a fixed similarity matrix in the whole process of LOOCV) in the each step of LOOCV when each known lncRNA–disease association is left out as test sample. As a result, LRLSLDA achieved an AUC of 0.7760
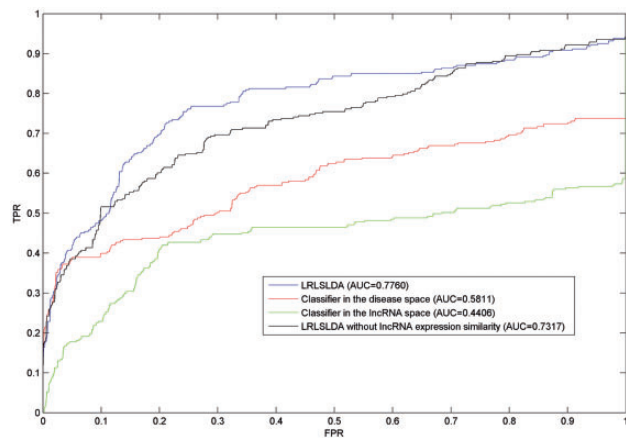
**Fig. 3.** Comparison between LRLSLDA and the classifier in the single space was shown, which was implemented based on LOOCV schema in terms of ROC curve and AUC. The result demonstrates LRLSLDA has a reliable performance for potential lncRNA–disease inference and shows the benefit from combining the classifiers from different spaces into the single one. We also implemented LOOCV based on the LRLSLDA without introducing the information of expression profiles. A slightly lower AUC has been obtained, which shows the benefit from the introduction of expression profiles and relative reliable predictive ability even if the information of expression profiles can not be obtained for lncRNAs of interest

and the comparison between LRLSLDA and the predictive results in the single space (disease space or lncRNA space) was shown in Figure 3. The conclusion can be reached that predictive accuracy has been significantly improved by the operation of combining the classifiers in different spaces. We further implemented LOOCV based on the LRLSLDA without introducing the information of expression profiles. The AUC of 0.7317 has been obtained (Fig. 3), which was slightly lower than LRLSLDA with expression information. From this comparison, we can reach the following two conclusions. For one thing, the performance of LRLSLDA could be further improved by introducing the information of expression profiles. For another, we still can obtain the relative reliable predictive results even if the information of expression profiles can not be obtained for lncRNAs of interest. Considering that each disease is only associated with less than two diseases on average in the known disease–lncRNA associations, the performance of LRLSLDA is reliable and would be further improved after obtaining more known disease–lncRNAs associations.

### 4.2 The effect of weight coefficients on LRLSLDA performance

There are two kinds of weight coefficients in LRLSLDA: combinatorial coefficients in integrated lncRNA similarity and the final classifier, respectively.

In the previous researches of disease-related miRNA-EF interactions prediction and drug target prediction, simple average was adopted to combine different similarity measures of drug, protein target, miRNA and EF, respectively, where reliable performance have been obtained (Chen *et al.*, 2012a, c). Even in the study about drug target interactions prediction, the robustness of

predictive accuracy to weight parameters selection has been illustrated. No accurate and practical methods for similarity integration have been developed and applied well to bioinformatics research so far. In the Supplementary Figure S1, we showed the AUC values under different lncRNA similarity weights in the framework of LOOCV. We again observed the predictive accuracy of LRLSLDA is robust to the selection of lncRNA similarity weight coefficients.

Currently, there are no good methods for the selection of weight coefficients to combine different classification functions into the final optimal classification function. In the previous studies of drug–target prediction and disease-related miRNA-EF prediction (Chen *et al.*, 2012a; Xia *et al.*, 2010), reliable performance has been obtained based on simple average operation for classifier combination. Especially, it has been shown that predictive accuracy of miREFScan in disease-related miRNA-EF prediction is not sensitive to the selection of weight parameter (Chen *et al.*, 2012a). Here, we assigned 0.1–0.9 to weight coefficients and calculated corresponding AUC values in the framework of LOOCV (Supplementary Fig. S2). It has been observed that predictive performance of LRLSLDA is not sensitive to weight coefficient selection. Also we can conclude that it is improper to give too much weight to classifier function in the lncRNA space. It is not difficult to understand this conclusion. There were only 293 disease–lncRNA associations for 167 diseases in the known golden standard dataset, i.e. less than two associations per disease. When one association was left out as test sample in the process of LOOCV, it is much likely that there were no lncRNAs associated with this disease in the known training samples. Therefore, only lncRNA similarity is not enough and disease similarity must be introduced to make full use of the information of similar diseases.

### 4.3 Compared with other methods

As mentioned above, there is no method to predict potential lncRNA–disease associations in the previous studies. However, there are some similar problems in other fields of computational biology and many methods have been already developed to solve these problems. Some of these methods can be applied to predict lncRNA–disease associations. Taking into account the fact there are less than two associations for each disease on average, it is infeasible to implement LOOCV for a given disease, and LOOCV must be implemented for all the diseases simultaneously as we did before. Therefore, only global methods can be used in LOOCV, which can prioritize candidate lncRNAs for all the disease simultaneously, and compare the scores of different lncRNA–disease pairs.

Hence, we compared LRLSLDA with three previous published methods as follows in LOOCV based on the same dataset: (i) RLS-kron (van Laarhoven *et al.*, 2011), combining different kernels based on Kronecker product in the drug–target interaction prediction; (ii) hypergeometric distribution method (Jiang *et al.*, 2010), predicting disease-related miRNAs based on the hypergeometric distribution; (iii) network-consistency-based inference (Chen and Zhang, 2013), inferring potential disease–miRNA associations based on the idea of network consistency. The comparison between LRLSLDA and three previous methods in the LOOCV was shown in Figure 4, which
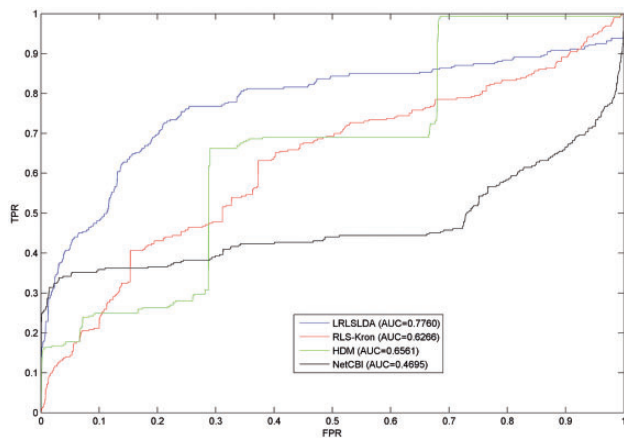
**Fig. 4.** Here, we compared the performance of LRLSLDA with three previous state-of-the-art methods in LOOCV. It has been indicated that LRLSLDA significantly improved previous methods LOOCV by at least 0.1199 in the term of AUC values



**Fig. 5.** Here, we implemented LOOCV for the test samples in C1, C2 and C3, respectively, and the performance have been shown, with AUC of 0.7880 in C1, 0.8071 in C2, 0.8736 in C3. Results indicate that LRLSLDA has a reliable predictive performance in different test classes

significantly improved the performance of previous method by at least 0.1199 in the term of AUC values and fully demonstrated the superiority performance of LRLSLDA.

### 4.4 LOOCV in the new validation framework

Recently, Park and Marcotte (2012) have pointed out that the flaw of evaluation procedure based on cross validation for the pair-input computational prediction problems has affected a number of previous studies. They have demonstrated that the paired nature of inputs leads to a natural partitioning of test pairs, and pair-input methods achieve significantly different predictive performances for distinct test classes (Park and Marcotte, 2012). They further performed experiments for protein–protein interactions prediction based on seven state-of-the-art methods and observed that the performance of each method differs significantly in different test classes (Park and Marcotte, 2012).

According to the evaluation methods proposed in this article, we classified test samples of lncRNA–disease associations into four distinct classes: C1 is composed of the test samples sharing both diseases and lncRNAs with the training samples; C2 is composed of the test samples sharing only diseases with the training samples; C3 is composed of the test samples sharing only lncRNAs with the training samples; C4 is composed of the test samples sharing neither diseases nor lncRNAs with the training samples. To be honest, LRLSLDA strongly relies on the topology structures in known disease–lncRNA association network, so it can not be applied to predict test samples in C4. We implemented LOOCV for the test samples in C1, C2 and C3, respectively. The performance of LRLSLDA in these three classes test samples have been shown in Figure 5 (AUC of 0.7880 in C1, 0.8071 in C2, 0.8736 in C3), which has illustrated that LRLSLDA has a reliable predictive performance in different test classes.

### 4.5 Case studies and novel lncRNA–disease association prediction

We applied LRLSLDA to prioritize all the candidate lncRNAs for each disease investigated in this article. Here, all the known
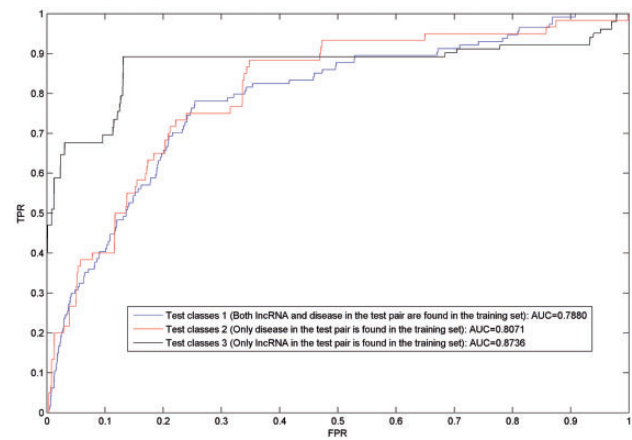
disease–lncRNA associations in the gold standard dataset were used as training samples. Predictive results were publicly released to benefit experimental validation from biologists (Supplementary Table S3). It is anticipated that these potential lncRNA–disease associations predicted by LRLSLDA could be confirmed by biological experiments.

Recent results in biological experiments confirmed that Alzheimer disease is related to gene RELN and its antisense transcript HAR1A and HAR1B (Harries, 2012). These two lncRNAs were both ranked in the top of predictive list for Alzheimer disease (17th and 18th, respectively). Recently, Han *et al.* (2012) confirmed lincRNA TUG1 is up-related in 44 patients with bladder cancer based on Real-Time qPCR. In the potential bladder cancer–related lncRNAs list predicted by LRLSLDA, TUG1 was ranked 18th.

Aforementioned lncRNAs all have known related diseases in the golden standard dataset. One of advantages of LRLSLDA is that it can predict potential related diseases for lncRNA of interest, even if it does not have any related diseases in the training dataset. Flockhart *et al.* (2012) identified the potential functional role of lncRNA BRAF-regulated lncRNA 1 (BANCR) in melanoma cell migration. BANCR has not been associated with any diseases in our golden dataset. We ranked all the candidate diseases for BANCR. In the predictive results, melanoma is ranked 10th out of all the 167 candidate diseases.

As mentioned above, LRLSLDA is a global ranking approach that can prioritize potential lncRNA–disease associations for all the diseases simultaneously. Therefore, we further applied our method to simultaneously rank all the candidate lncRNA–disease associations. We also publicly released predictive results in the Supplementary Table S4. In this global ranking, potential association between lncRNA NEAT1 and breast cancer was ranked fourth out of 19 413 candidate associations. Recent study by Redvers *et al.* (2012) confirmed this potential association. They implemented an integrated *in vivo* genomics screen and revealed significant upregulation of NEAT1 in breast metastasis tumors. This independent and high-ranking evidence

further demonstrates the reliable performance of LRLSLDA and gives a strong support to other predicted lncRNA–disease associations (Supplementary Table S4).

## 5 DISCUSSION AND CONCLUSION

Disease–lncRNA association inference is important in the design of specific molecular tools for human disease diagnosis, treatment, prognosis and prevention. In this article, we proposed the assumption that similar diseases tend to be associated with functionally similar lncRNAs and further developed a novel method of LRLSLDA in the framework of LapRLS. LRLSLDA can effectively identify potential disease–lncRNA associations on a large scale by integrating the information of known disease–lncRNA associations and lncRNA expression profiles. More importantly, LRLSLDA is a semisupervised method that does not need the information of negative samples. Also, it can prioritize lncRNA–disease pairs for all the diseases simultaneously. The method has shown its reliable performance in the term of LOOCV and significantly improved previous methods. According to the evaluation methods proposed in a recent article, we further demonstrated LRLSLDA can work effectively in the different test samples of lncRNA–disease associations. Therefore, we publicly released plenty of potential lncRNA–disease pairs for biological experiments validation, and some of potential associations have been confirmed by recent results in biological experiments. As an effective and important biological tool, we anticipated LRLSLDA can benefit early diagnosis and treatment of diseases and improvement of the human health in the future.

The reliable performance of LRLSLDA could be largely attributed to the following several factors, which are also the reasons for which we integrate LapRLS and Gaussian kernel for potential disease-related lncRNA prediction. Firstly, known disease–lncRNA associations and lncRNA expression profiles could be integrated to capture the potential associations between disease and lncRNA. Especially, this method can be used to predict potential disease–lncRNA associations sharing only diseases or lncRNAs with the known associations in training dataset. Secondly, the classifiers from different spaces would be combined and the predictive ability could be significantly improved in this way, which has been fully demonstrated from the comparison between LRLSLDA and classifier in the single space in Figure 3. More importantly, as a semi-supervised method, the advantage of LRLSLDA over supervised methods has been shown in many previous studies. Especially, semisupervised method could be implemented without any negative disease–lncRNA associations, which are difficult or even impossible to obtain nowadays. Finally, as a global method, LRLSLDA can predict the potential lncRNA–disease associations for all the diseases simultaneously. In conclusion, LRLSLDA represents a novel, important and powerful tool in biomedical research for disease treatment and drug discovery.

Some limitations also exist in the LRLSLDA. Firstly, many parameters appear in our model and how to select the parameter is not still solved well. Secondly, for the same lncRNA–disease pair, two different scores from different spaces will be obtained. How to directly obtain a single classifier or reasonably integrate these two classifiers would be an important problem for future

research. Thirdly, introducing more reliable measure of disease similarity and lncRNA similarity and developing more reliable similarity integration method would improve the performance of LRLSLDA. Especially, disease similarity in this model totally relies on known disease–lncRNA association. We would construct new similarity measures that do not rely on topology structures in the known association network and hence we can predict potential associations sharing neither diseases nor lncRNAs with known associations. Finally, available experimentally verified disease–lncRNA associations are still comparatively rare. The performance of LRLSLDA would be further improved when more known associations can be obtained.

## ACKNOWLEDGEMENTS

## REFERENCES

Amaral,P.P. *et al.* (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.*, **39**, D146–D151.

Babak,T. *et al.* (2005) A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. *BMC Genomics*, **6**, 104.

Belkin,M. *et al.* (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, **7**, 2399–2434.

Bertone,P. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Bono,H. *et al.* (2003) Systematic expression profiling of the mouse transcriptome using RIKEN cDNA microarrays. *Genome Res.*, **13**, 1318–1323.

Bu,D. *et al.* (2012) NONCODE v3. 0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.*, **40**, D210–D215.

Carninci,P. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.

Chen,G. *et al.* (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.

Chen,H. and Zhang,Z. (2013) Similarity-based methods for potential human microRNA-disease association prediction. *BMC Med. Genomics*, **6**, 12.

Chen,X. *et al.* (2012a) Prediction of disease-related interactions between microRNAs and environmental factors based on a semi-supervised classifier. *PloS One*, **7**, e43425.

Chen,X. *et al.* (2012b) RWRMDA: predicting novel human microRNA-disease associations. *Mol. Biosyst.*, **8**, 2792–2798.

Chen,X. *et al.* (2012c) Drug-target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.*, **8**, 1970–1978.

Chung,S. *et al.* (2011) Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. *Cancer Sci.*, **102**, 245–252.

Claverie,J.M. (2005) Fewer genes, more noncoding RNA. *Science*, **309**, 1529–1530.

Congrains,A. *et al.* (2011) Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B. *Atherosclerosis*, **220**, 449–455.

Core,L.J. *et al.* (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.

Crick,F. *et al.* (1961) General nature of the genetic code for proteins. *Nature*, **192**, 1227–1232.

Dinger,M.E. *et al.* (2009) NRED: a database of long noncoding RNA expression. *Nucleic Acids Res.*, **37**, D122–D126.

Flockhart,R.J. *et al.* (2012) BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration. *Genome Res.*, **22**, 1006–1014.

Gibb,E.A. *et al.* (2011) The functional role of long non-coding RNA in human carcinomas. *Mol. Cancer*, **10**, 38.

Gupta,R.A. *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, **464**, 1071–1076.

Guttman,M. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.

Han,Y. *et al.* (2012) Long intergenic non-coding RNA TUG1 is overexpressed in urothelial carcinoma of the bladder. *J. Surg. Oncol.*, **107**, 555–559.

Harries,L. (2012) Long non-coding RNAs and human disease. *Biochem. Soc. Trans.*, **40**, 902.

Jiang,Q. *et al.* (2010) Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.*, **4**, S2.

Johnson,R. (2011) Long non-coding RNAs in Huntington's disease neurodegeneration. *Neurobiol. Dis.*, **46**, 245–254.

Kapranov,P. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.

Lander,E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Managadze,D. *et al.* (2011) Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol. Evol.*, **3**, 1390.

Mattick,J.S. (2009) The genetic signatures of noncoding RNAs. *PLoS Genet.*, **5**, e1000459.

Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. *Hum. Mol. Genet.*, **15**, R17–R29.

Mercer,T.R. *et al.* (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.

Mercer,T.R. *et al.* (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl Acad. Sci.*, **105**, 716–721.

Mitchell Guttman,I.A. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.

Park,Y. and Marcotte,E.M. (2012) Flaws in evaluation schemes for pair-input computational predictions. *Nat. Methods*, **9**, 1134–1136.

Pauli,A. *et al.* (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.*, **22**, 577–591.

Ponting,C.P. *et al.* (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.

Qureshi,I.A. *et al.* (2010) Long-non-coding RNAs in nervous system function and disease. *Brain Res.*, **1338**, 20–35.

Ramsköld,D. *et al.* (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, **5**, e1000598.

Redvers,R.P. *et al.* (2012) An integrated in vivo genomics screen implicates long non-coding RNAs H19 and Neat1 in breast cancer metastasis. In: *14th International Biennial Conference on Metastasis Research*. Brisbane, Australia.

Spizzo,R. *et al.* (2012) Long non-coding RNAs and cancer: a new frontier of translational research? *Oncogene*, **31**, 4577–4587.

Taft,R.J. *et al.* (2010) Non-coding RNAs: regulators of disease. *J. Pathol.*, **220**, 126–139.

Taft,R.J. *et al.* (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, **29**, 288–299.

van Laarhoven,T. *et al.* (2011) Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, **27**, 3036–3043.

van Poppel,H. *et al.* (2011) The relationship between Prostate CAncer gene 3 (PCA3) and prostate cancer significance. *BJU Int.*, **109**, 360–366.

Vanunu,O. *et al.* (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.*, **6**, e1000641.

Wapinski,O. and Chang,H.Y. (2011) Long noncoding RNAs and human disease. *Trends Cell Biol.*, **21**, 354–361.

Wilusz,J.E. *et al.* (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.*, **23**, 1494–1504.

Xia,Z. *et al.* (2010) Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.*, **4**, S6.

Yang,Z. *et al.* (2011) Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation. *Ann. Surg. Oncol.*, **18**, 1243–1250.

Yanofsky,C. (2007) Establishing the triplet nature of the genetic code. *Cell*, **128**, 815–818.

Zhang,Z. *et al.* (2012) [Evaluation of novel gene UCA1 as a tumor biomarker for the detection of bladder cancer]. *Zhonghua Yi Xue Za Zhi*, **92**, 384.