Integrating multiple microarray dataset analysis and machine learning methods to reveal the key genes and regulatory mechanisms underlying human intervertebral disc degeneration

Hongze Chang, Xiaolong Yang, Kemin You, Mingwei Jiang, Feng Cai,
Yan Zhang, Liang Liu, Hui Liu and Xiaodong Liu

Department of orthopedics, Shanghai Yangpu Hospital Affiliated to Tongji University, Shanghai, China

ABSTRACT

Intervertebral disc degeneration (IDD), a major cause of lower back pain, has multiple contributing factors including genetics, environment, age, and loading history. Bioinformatics analysis has been extensively used to identify diagnostic biomarkers and therapeutic targets for IDD diagnosis and treatment. However, multiple microarray dataset analysis and machine learning methods have not been integrated. In this study, we downloaded the mRNA, microRNA (miRNA), long noncoding RNA (lncRNA), and circular RNA (circRNA) expression profiles (GSE34095, GSE15227, GSE63492 GSE116726, GSE56081 and GSE67566) associated with IDD from the GEO database. Using differential expression analysis and recursive feature elimination, we extracted four optimal feature genes. We then used the support vector machine (SVM) to make a classification model with the four optimal feature genes. The ROC curve was used to evaluate the model's performance, and the expression profiles (GSE63492, GSE116726, GSE56081, and GSE67566) were used to construct a competitive endogenous RNA (ceRNA) regulatory network and explore the underlying mechanisms of the feature genes. We found that three miRNAs (hsa-miR-4728-5p, hsa-miR-5196-5p, and hsa-miR-185-5p) and three circRNAs (hsa_circRNA_100723, hsa_circRNA_104471, and hsa_circRNA_100750) were important regulators with more interactions than the other RNAs across the whole network. The expression level analysis of the three datasets revealed that BCAS4 and SCRG1 were key genes involved in IDD development. Ultimately, our study proposes a novel approach to determining reliable and effective targets in IDD diagnosis and treatment.

Submitted 13 May 2020
Accepted 16 September 2020
Published 13 October 2020

Corresponding author
Xiaodong Liu,
xiaodong.liu@tongji.edu.cn

Academic editor
Jianye Ge

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj.10120**

Subjects Bioinformatics, Genomics, Orthopedics, Medical Genetics
Keywords Intervertebral disc degeneration, Integrated analysis, Machine learning methods, Key genes

INTRODUCTION

Lower back pain, commonly caused by intervertebral disc degeneration (IDD), can be a significant socioeconomic burden on patients (*Vergroesen et al., 2015*). IDD is characterized by the apoptosis of nucleus pulposus (NP) cells, the degradation of extracellular matrix

How to cite this article Chang H, Yang X, You K, Jiang M, Cai F, Zhang Y, Liu L, Liu H, Liu X. 2020. Integrating multiple microarray dataset analysis and machine learning methods to reveal the key genes and regulatory mechanisms underlying human intervertebral disc degeneration. *PeerJ* **8**:e10120 http://doi.org/10.7717/peerj.10120

(ECM) components, and several contributing factors including genetics and environment (*Battie et al., 2008*; *Feng, Egan & Wang, 2016*). However, the precise etiology of IDD remains largely unknown. Diagnosing degenerative disc disease is difficult, and common IDD treatment and management strategies primarily consist of conservative management or surgical treatment to relieve pain, without resolving the underlying tissue pathology (*An et al., 2003*; *Zaina et al., 2016*). To identify potential biomarkers and specific therapeutic targets, a more detailed understanding of the molecular and cellular events underlying IDD formation is needed.

The etiology of IDD is complex, but over the past several decades, it has become clear that genetic factors are most dominant. Multiple candidate genes, including *thrombospondin-2*, vitamin D receptor, *COL2A1*, *ACAN*, interleukins (*IL1α, IL 1*β, and *IL6*), matrix metalloproteinases (*MMP-3* and *MMP-9*), and growth/differentiation factor 5, have been associated with the pathophysiological process of IDD development (*Feng, Egan & Wang, 2016*; *Kalb et al., 2012*; *Wang et al., 2018b*; *Yuan et al., 2018*). Genome-wide association studies (GWAS) have been used to help identify novel variants. Several GWAS found that the genetic polymorphisms of *PARK2* and *CHST3* were relevant to IDD etiology (*Song et al., 2013*; *Williams et al., 2013*). Additionally, some bioinformatics analyses based on gene expression profiles revealed that *FYN*, *PRKCD*, *YWHAB*, *YWHAZ*, *AR*, *Fibronectin 1*, *COL2A1*, *β-catenin*, *COL6A2*, *IBSP*, *RAP1A*, and *FOXF2* genes may play key roles in IDD development (*Chen et al., 2013*; *Guo et al., 2017*; *Ji et al., 2015*). Although a considerable number of genes associated with IDD development have been found, early IDD diagnosis and precise treatment remain difficult and require further study.

Integrated analyses using combined multiple microarray data can provide a more accurate understanding of the interplay across multi-level genomic features and the molecular mechanisms that cause complex diseases (*Momtaz et al., 2018*). Machine learning is a type of artificial intelligence that can "learn" a model using past data in order to predict future data. Machine learning algorithms have been used in key feature training, recognition, and group classification (*Huang et al., 2018*). Modern researchers have unprecedented access to machine learning methods that can elucidate complex molecular mechanisms and predict disease genes from large biomedical datasets (*Libbrecht & Noble, 2015*; *Obermeyer & Emanuel, 2016*). To the best of our knowledge, no investigations have used integrated analysis and machine learning methods to identify key IDD-associated genes.

Accumulating evidence has indicated that noncoding RNAs, including microRNAs (miRNAs), long noncoding RNAs (lncRNAs), and circular RNAs (circRNAs), are important gene expression regulators that influence cellular function and disease states (*Adams et al., 2017*). lncRNAs and circRNAs act as competitive endogenous RNAs (ceRNAs) and competitively bind miRNA response elements (MREs) to construct a regulatory network involved in IDD progression. *Zhao et al. (2016)* used RNA sequencing to identify 1,854 lncRNAs and 2,804 protein-coding genes that were differentially expressed in the IDD group. *Tan et al. (2018)* found that LncSNHG1 promoted NP cell proliferation by suppressing miR-326 expression and upregulating *CCND1* expression. Recent studies have focused on the functional roles of circRNAs during IDD development and found that

Table 1 **Basic information of expression profiles included in the study.**

Type	Series	Platform	Source name	Number of samples (control/degenerative)	Publication year
mRNA	GSE34095	GPL96	Disc tissue	6(3/3)	2012
mRNA	GSE15227	GPL1352	Disc tissue	15 (12/3)	2009
miRNA	GSE63492	GPL19449	Nucleus pulposus	10 (5/5)	2016
miRNA	GSE116726	GPL20712	Nucleus pulposus	6(3/3)	2018
mRNA-lncRNA	GSE56081	GPL15314	Nucleus pulposus	10(5/5)	2014
circRNA	GSE67566	GPL19978	Nucleus pulposus	10(5/5)	2016

circ-4099, circ-GRB10, circVMA21, and circ_001653 play pivotal roles during NP cell proliferation, apoptosis, and extracellular matrix synthesis/degradation (*Cheng et al., 2018*; *Cui & Zhang, 2020*; *Guo et al., 2018*; *Wang et al., 2018a*). The diverse biological functions of ceRNAs deserve further exploration.

In this study, we aimed to identify the key genes and underlying mechanisms of IDD development by constructing an lncRNA/circRNA-miRNA-mRNA network using multiple microarray datasets and machine learning methods. Figure S1 shows the flow chart for this study. Our results present novel biomarkers and therapeutic targets that can be used for IDD diagnosis and treatment.

MATERIALS & METHODS

Microarray datasets

We retrieved a total of six expression profiles from the GEO database (http://www.ncbi.nlm.nih.gov/geo): two mRNA expression profiles (GSE34095 (*Tsai et al., 2013*) and GSE15227 (*Gruber et al., 2009*)), two miRNA expression profiles (GSE63492 (*Lan et al., 2016*)) and GSE116726 (*Ji et al., 2018*), one mRNA-lncRNA expression profile (GSE56081 (*Lan et al., 2016*; *Wan et al., 2014*)), and one circRNA expression profile (GSE67566 (*Lan et al., 2016*; *Liu et al., 2015*)). Table 1 contains the basic information for these expression profiles.

Differential expression analysis

The raw data were annotated, normalized, log 2 transformed, and screened for differentially expressed genes (DEGs), differentially expressed miRNAs (DEMs), differentially expressed lncRNAs (DELs), and differentially expressed circRNAs (DECs) in IDD and normal disc tissue using the “Limma” R package (*Ritchie et al., 2015*). Since each dataset came from a different experiment and microarray platform, the way to get the data may be different. When filtering using consistent thresholds, some datasets did not yield valid differential genes. In order to obtain effective differentially expressed genes for subsequent analysis, we used different thresholds according to each dataset’s conditions. The specific DEG screening thresholds were as follows: p-value <0.05 for the GSE34095 dataset, p-value <0.01 and |log2FC|≥ 2 for the GSE15227 dataset, and p-value <0.01 and |log2 FC|>2 for the GSE116726 dataset. We obtained the GSE67566, GSE63492, and GSE56081 datasets from the same tissue samples, and provided the differential expression analysis results and thresholds as Supplemental Materials (Tables S1 and S2, Documents S1 and S2). We visualized the differential expression analysis results using a volcano plot and heatmap with hierarchical clustering (Fig. S2).

Feature gene extraction

Using VENNY 2.1 software (http://bioinfogp.cnb.csic.es/tools/venny/index.html), we extracted the differential gene intersections of the GSE34095 and GSE15227 datasets to use as IDD-related DEGs (*Li et al., 2019*; *Zhang et al., 2018*). We used the GSE15227 dataset as the training set to screen for important feature genes. Using the R caret package, random forest, and neural network methods, we constructed the model and obtained the important DEG features (*Kuhn, 2015*). We used recursive feature elimination (RFE), a machine learning method, to extract the optimum feature genes to identify the functional biomarkers involved in IDD progression (*Guyon et al., 2002*).

SVM model verification

The support vector machine (SVM) is particularly effective for binary classification in a supervised learning manner, and is better than other machine learning methods at identifying subtle patterns in complex datasets (*Aruna & Dr, 2011*). Radial basis function kernel (RBF kernel) is commonly used in nonlinear support vector machine classification since it can enable data to operate in a high-dimensional and implicit feature space (*Jiao et al., 2017*). In this study, we used the R e1071 package (https://CRAN.R-project.org/package=e1071) to build an RBF kernel SVM model to identify the optimal feature genes, and we selected the GSE15227 dataset to train the machine. We then evaluated its performance in the GSE15227 training set using the R pROC package (http://www.biomedcentral.com/1471-2105/12/77/) (*Robin et al., 2011*). The optimal feature genes were taken from the training set, and over-fitting would occur whenever ROC verification was performed. Therefore, we used two validation sets (GSE34095 and GSE56081) to further verify the model's performance.

Identifying target miRNAs of the optimal feature genes

We used miRWalk 2.0 with 12 prediction programs (MIMATid, Microt4, miRanda, mirbridge, miRDB, miRMap, miRNAMap, Pictar2, PITA, RNA22, RNAhybrid, and Targetscan) to predict the optimal feature genes' target miRNAs (*Dweep & Gretz, 2015*). miRNAs present in more than five of the 12 prediction programs were considered as target miRNAs. We selected overlapping DEMs in the GSE116726 and GSE63492 datasets as candidate DEMs. Intersecting target miRNAs and candidate DEMs were selected as optimal feature gene interaction pairs.

ceRNA network construction

We obtained the miRNA sequences interacting with optimal feature genes from the miRbase (*Kozomara & Griffiths-Jones, 2011*) and extracted mature sequences using the Perl program. The DEL sequences were downloaded from NCBI. For DELs with several transcripts, we selected the longest transcripts for subsequent analysis. We used the BEDTOOLS command (*Quinlan & Hall, 2010*) and genomic coordinates to obtain the DEC sequences, and converted the gene name to its circRNA symbol using the Perl program.

We then used the miRanda tool to analyze the combinations of miRNAs, DELs, and DECs. We set the analysis parameters for the miRNA-DELs and miRNA-DECs as sc:120, en:-20 and sc:150, en:-7, respectively, and processed the results using python script (Documents S3 and S4). The DELs and DECs with ≥5 miRNA binding sites were identified as reliable miRNA-lncRNA and miRNA-circRNA interaction pairs. The ceRNA (DELs/DECs-miRNA-optimal feature gene) regulatory network was constructed using a combination of miRNA-DEL pairs, miRNA-DEC pairs, and miRNA-optimal feature gene pairs. We visualized the network using Cytoscape 3.6.0 (http://www.cytoscape.org/).

Statistical analysis

All data were presented as the mean ± SEM. Differential expression levels were compared using the Student's t-test in GraphPad Prism 7.0 (GraphPad Software Inc., La Jolla, CA, USA). P values < 0.05 were considered statistically significant.

RESULTS

Differentially expressed genes in IDD

We obtained a total of 334 DEGs (199 up-regulated and 135 down-regulated) in the GSE34095 dataset (Fig. 1A) and 188 DEGs (141 up-regulated and 47 down-regulated) in the GSE15227 dataset (Fig. 1B). Hierarchical cluster heatmaps showed that these DEGs could distinguish between the degenerative disc samples and the control disc samples (Figs. 1C and 1D). We obtained a total of 13 overlapping DEGs (*COL3A1*, *SCRG1*, *HTRA1*, *BCAS4*, *C11orf80*, *CRNKL1*, *GREM1*, *FGFR3*, *BDKRB1*, *WDR46*, *FN1*, *LMF2*, and *GDI2*) via the intersection between the two datasets, and considered these as IDD-related DEGs (Fig. 1E).

Optimal feature gene selection

We used random forest and neural network models to evaluate the importance of the 13 overlapping DEGs (Documents S5 and S6). The results showed that most of the DEGs from the two models had minimal differences in terms of gene importance rankings (Figs. 2A and 2B). Using the RFE method, we identified four important genes (*WDR46*, *BCAS4*, *CRNKL1*, and *SCRG1*) as optimal feature genes associated with IDD (Fig. 2C).

A classification model was constructed using the RBF kernel SVM, the four genes as features, and the GSE15227 dataset to train the machine. The parameters were set as: SVM-Kernel: radial, cost:1, gamma:0.25, and epsilon:0.1. Additionally, their performance was assessed using the R pROC package. In the GSE15227 training set, the area under the ROC curve (AUC) of the SVM model was 100 percent, suggesting that the model could accurately distinguish between IDD and normal samples (Fig. 2D). We used the GSE34095 and GSE56081 datasets as validation sets to further evaluate the model's performance and to help avoid over-fitting in the training set. The SVM model in the GSE34095 validation set had an AUC of 55.6 percent, which may be correlated with the small sample size of the microarray dataset (Fig. 2E). However, the AUC for the GSE56081 validation set was 100 percent (Fig. 2F). These results indicated that the optimal feature genes (*WDR46*, *BCAS4*, *CRNKL1*, and *SCRG1*) could be used as effective and accurate IDD diagnostic biomarkers.

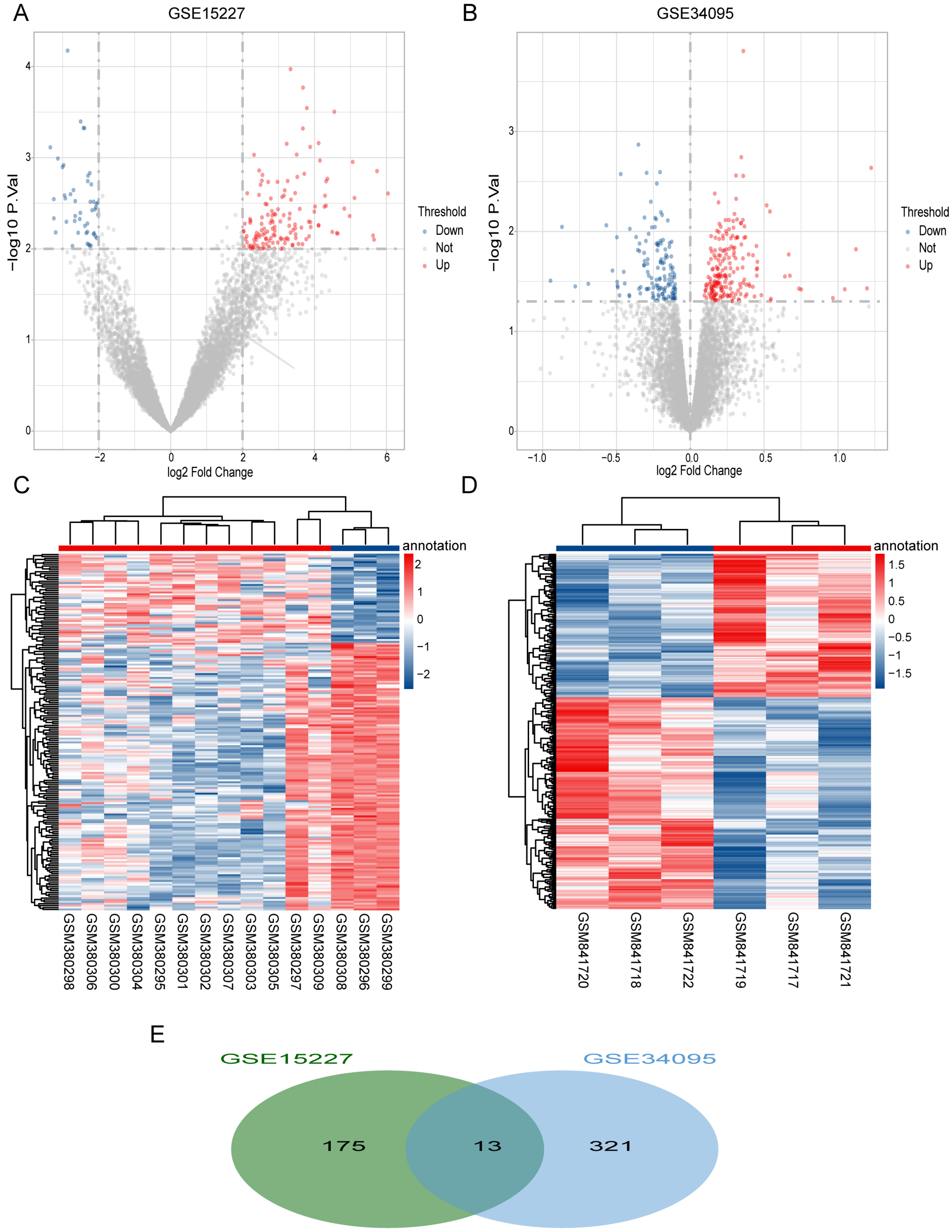

Figure 1 Differentially expressed IDD genes. (A and B) Volcano plots represent the DEGs in the degenerative disc samples and control disc samples in the GSE15227 and GSE34095 datasets, respectively. (C and D) Hierarchical cluster heatmaps of the GSE15227 and GSE34095 datasets displaying the DEGs in the degenerative disc samples and the control disc samples. Blue represents the downregulated and red represents the upregulated. (E) Venn diagram of DEGs in the GSE15227 and GSE34095 datasets. The common area represents the overlapping genes. DEG, differentially expressed genes.

Full-size DOI: 10.7717/peerj.10120/fig-1

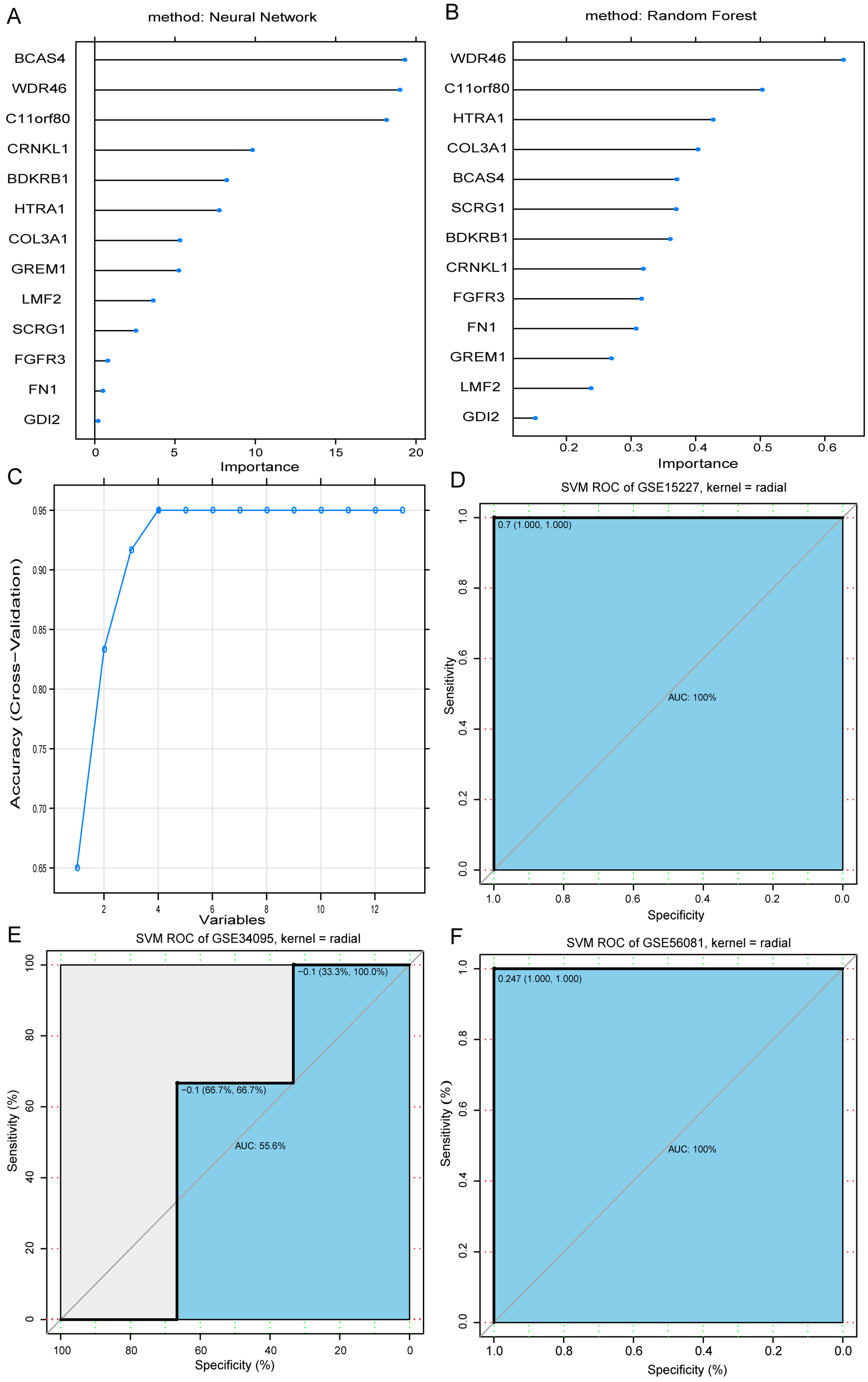

Figure 2 Selection of optimal feature genes. Ranking of the top 13 IDD-related genes using neural networks (A) and random forest (B). Extraction of the optimum feature genes from the 13 IDD-related genes was carried out using recursive feature elimination (C). Classification efficiency of the optimum feature genes in the model as evaluated using the ROC curve in the GSE15227 (D), GSE34095 (E), and GSE56081 (F) datasets, respectively.

Full-size DOI: 10.7717/peerj.10120/fig-2

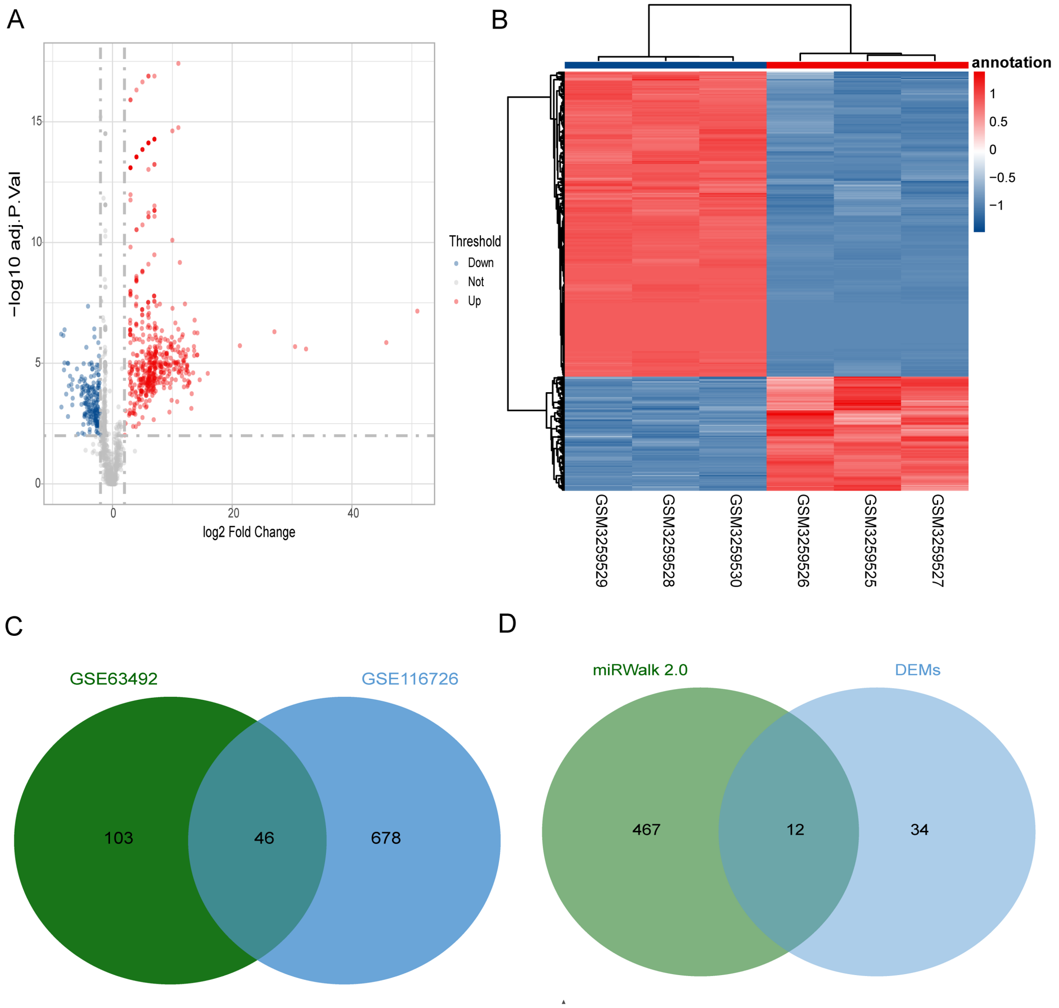

Figure 3 Identification of target miRNAs of optimal feature genes. (A) Volcano plots represent the DEMs of the degenerative disc samples and control disc samples in the GSE116726 dataset. (B) Hierarchical cluster heatmaps of the GSE116726 dataset display the DEMs to compare degenerative disc samples and control disc samples. Blue represents the degenerative samples and red represents the control samples. (C) Venn diagram of DEMs in the GSE116726 and GSE63492 datasets. The common area represents the overlapping DEMs. (D) Venn diagram of miRNAs in overlapping DEMs and target miRNAs of optimum feature genes. The common area represents the overlapping miRNAs. DEM, differentially expressed miRNA.

Full-size DOI: 10.7717/peerj.10120/fig-3

Identifying target miRNAs of the optimal feature genes

We predicted a total of 467 target miRNA optimal feature genes using miRWalk 2.0. Compared to the control disc samples, we identified 724 DEMs (527 up-regulated and 197 down-regulated) from the GSE116726 dataset (Figs. 3A and 3B) and 149 DEMs from the GSE63492 dataset. Figure 3C shows a Venn diagram of the 46 common DEMs that were found. Based on the intersections between the target miRNAs and common DEMs, we further analyzed 12 overlapping miRNAs which we considered to be the optimal feature genes' interaction miRNAs (Fig. 3D).

ceRNA network construction

The differentially expressed lncRNAs and circRNAs between the degenerated disc samples and the control disc samples were downloaded from the GEO database. We analyzed the miRNA-DEL interactions and the miRNA-DEC interactions using the miRanda tool with ≥5 miRNA binding sites. The miRNA-DEL pairs, miRNA-DEC pairs, and miRNA-optimal feature gene pairs were combined to build a DELs/DECs-DEMs-optimal feature gene regulatory network, which included four mRNA nodes, 12 miRNA nodes, 10 lncRNA nodes, and 75 circRNA nodes (Fig. 4, Table S3). Out of these, three miRNAs (hsa-miR-4728-5p, hsa-miR-5196-5p, and hsa-miR-185-5p) and three circRNAs (hsa_circRNA_100723, hsa_circRNA_104471, and hsa_circRNA_100750) were key regulators, based on the optimal feature genes and connective degrees of the whole network (Table 2). However, the LncRNA connective degree was very low.

Validating optimal feature genes

The expression of the four optimal feature genes across the three datasets was visualized using box plots. In the GSE15227 and GSE56081 datasets, *BCAS4* expression levels were significantly down-regulated (Figs. 5A and 5C), and *SCRG1* was significantly up-regulated (Figs. 5G and 5I). *CRNKL1* was only up-regulated in the GSE34095 dataset (Fig. 5E). Unfortunately, the expression levels of *WDR46* showed no significant difference across the three datasets (Figs. 5J, 5K and 5L). These results indicated that *BCAS4* and *SCRG1* are key genes involved in IDD development.

DISCUSSION

IDD is linked to lower back pain and spine-related diseases, and although IDD's underlying mechanisms have been studied for many years, they still remain unclear. Insufficient early IDD diagnosis and treatment methods affect the quality of life for patients and impose a heavy economic burden on society (*Vergroesen et al., 2015*). NP cells play an important role in maintaining intervertebral disc homeostasis by synthesizing ECM, which includes aggrecan and type II collagen (*Zhang et al., 2016*). Recent studies show that targeting gene therapy can inhibit NP cell senescence and apoptosis, and can ultimately ameliorate IDD (*Chen et al., 2018*). Therefore, reliable and specific gene targets are essential for IDD diagnosis and treatment.

Recent developments in bioinformatics and computational biology have led to the identification of several key genetic targets related to IDD and the prediction of potential molecular mechanisms (*Petryszak et al., 2014*). In this study, we downloaded multiple microarray datasets associated with IDD from the GEO database, including two mRNA expression profiles, two miRNA expression profiles, one mRNA-lncRNA expression profile, and one circRNA expression profile. A total of four optimal feature genes (*WDR46*, *BCAS4*, *CRNKL1*, and *SCRG1*) were identified using machine-learning methods. The construction of a DELs/DECs-miRNA-optimal feature genes network revealed that three miRNAs (hsa-miR-4728-5p, hsa-miR-5196-5p, and hsa-miR-185-5p) and three circRNAs (hsa_circRNA_100723, hsa_circRNA_104471, and hsa_circRNA_100750) may be important mediators for optimal feature genes. To further explore the different optimal

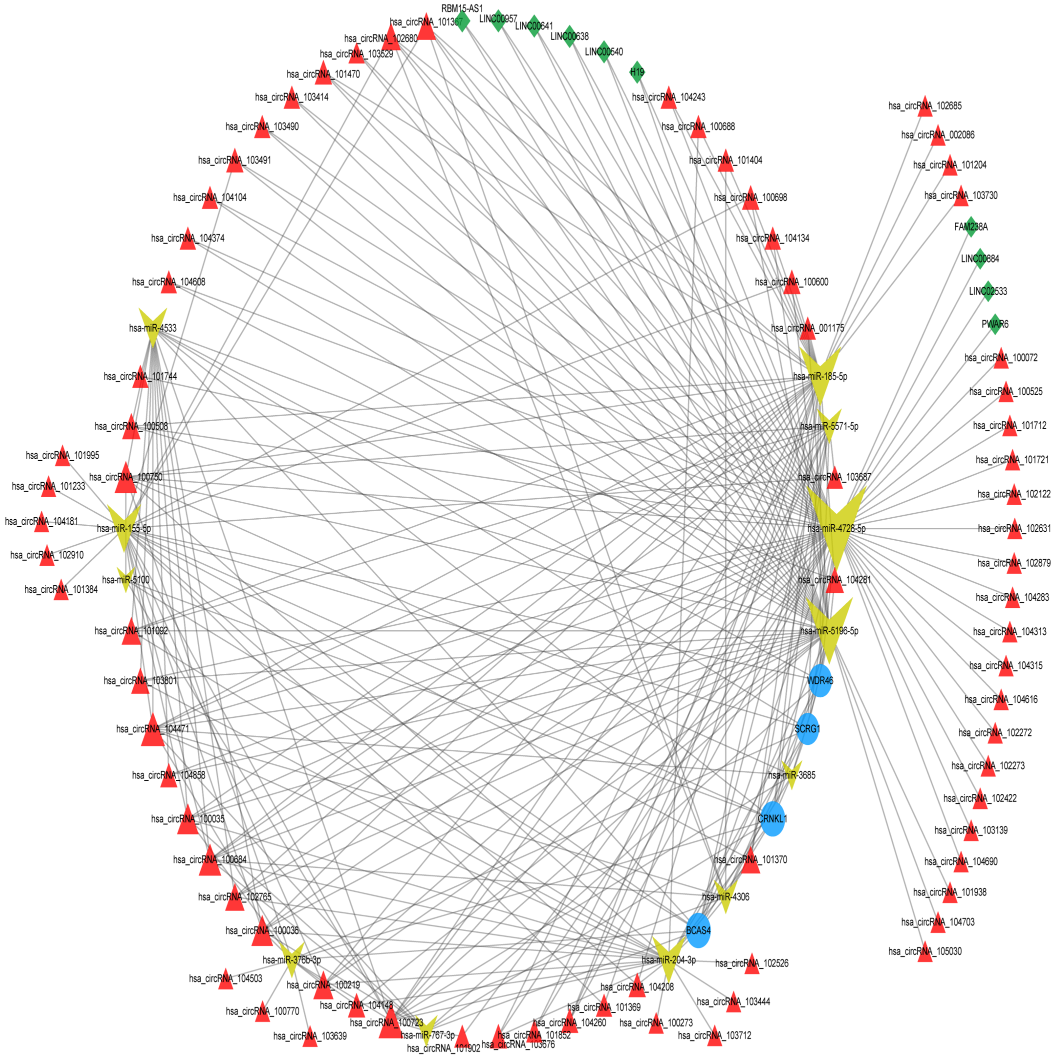

Figure 4 ceRNA network construction. ceRNA network of optimum IDD feature genes with DEMs, DELs, and DECs. The blue circles represent optimum IDD feature genes, the red triangles represent DECs, the green diamonds represent DELs, and the yellow arrows represent DEMs. ceRNA, competing endogenous RNA; DEC, differentially expressed circRNA; DEL, differentially expressed lncRNA; and DEM, differentially expressed miRNA.

Full-size DOI: 10.7717/peerj.10120/fig-4

feature genes in normal and degenerative NP tissues, we also investigated their expression levels across three datasets. The results indicated that *BCAS4* and *SCRG1* were key genes related to IDD.

Breast carcinoma amplified sequence 4 (*BCAS4*), a novel gene cloned from breast cancer cells, encodes a 211-amino acid cytoplasmic protein with no significant homologies to any known protein (*Barlund et al., 2002*). Previous studies have demonstrated that the specific DNA methylation of *BCAS4* acts as an epigenetic marker and can be used to distinguish saliva from other body fluids. It is also widely used in forensic investigations (*Silva et al., 2016*; *Taki & Kibayashi, 2015*). However, *BCAS4*'s biological role in disease requires

Table 2 The top 3 miRNAs and circRNAs related to optimal feature genes in the network.

Type	Name	Number of directed edges
miRNA	hsa-miR-4728-5p	58
miRNA	hsa-miR-5196-5p	41
miRNA	hsa-miR-185-5p	34
circRNA	hsa_circRNA_100723	12
circRNA	hsa_circRNA_104471	11
circRNA	hsa_circRNA_100750	9

further investigation. Stimulator of chondrogenesis 1 (*SCRG1*) was first found in the genes associated with, or responsible for, the neurodegenerative changes observed in transmissible spongiform encephalopathies (*Dandoy-Dron et al., 1998*). *SCRG1* transcript is found in the brain, heart, and spinal cord, and its sequence is highly conserved in humans, mice, and rats. *SCRG1* has also been observed to be specifically expressed in human articular cartilage, and is involved in human mesenchymal stem cell (hMSC) growth suppression and differentiation during dexamethasone-dependent chondrogenesis (*Ochi, Derfoul & Tuan, 2006*). Recent studies have shown that *SCRG1* is an important regulator during hMSC self-renewal, migration, and osteogenic differentiation along with its receptor *BST1* (*Aomatsu et al., 2014*). However, the function of *SCRG1* in IDD development has not yet been explored.

circRNAs and lncRNAs may act as ceRNAs by competitively binding to miRNA and suppressing mRNA expression (*Adams et al., 2017*; *Tay, Rinn & Pandolfi, 2014*). This ceRNA hypothesis suggests that there is a novel mechanism for RNA interactions. In this study's ceRNA analysis, hsa-miR-4728-5p, hsa-miR-5196-5p, hsa-miR-185-5p, hsa_circRNA_100723, hsa_circRNA_104471, and hsa_circRNA_100750 showed more interactions compared to the other RNAs in the whole network. Additionally, miR-155, miR-21, and miR-133a were shown to be differentially expressed in degenerative NP cells, indicating that they may be potential biomarkers for early IDD diagnosis (*Liu et al., 2014*; *Wang et al., 2011*; *Xu et al., 2016*). The dysregulation of these miRNAs is closely associated with NP cell apoptosis which affects IDD progression. miR-185-5p has been reported as a critical regulator, but miR-4728-5p and miR-5196-5p have rarely been reported. Chang et al. (2017) reported that miR-185-5p induced by *Runx2* could directly target *Dlx2* to inhibit amylogenesis and osteogenesis, providing a new treatment option for cleidocranial dysplasia. Multiple lncRNAs bind to miR-185-5p in order to modulate the progression of different types of human cancer, including prostate cancer (*Tian et al., 2018*), colorectal cancer (*Zhu et al., 2018*), and glioblastoma (*Wang et al., 2018b*). However, the regulatory effects of miR-185-5p on IDD require further investigation.

CircRNAs, a new type of ncRNAs formed by special loop splicing, are thought to be potential diagnostic biomarkers and therapeutic targets because they are more stable and conserved than other RNAs (*Lee et al., 2019*). However, very few studies have focused on the role of circRNAs in IDD development. *Guo et al. (2018)* found that circ-GRB10 was downregulated in IDD cells, and circ-GRB10 overexpression inhibited NP cell apoptosis by

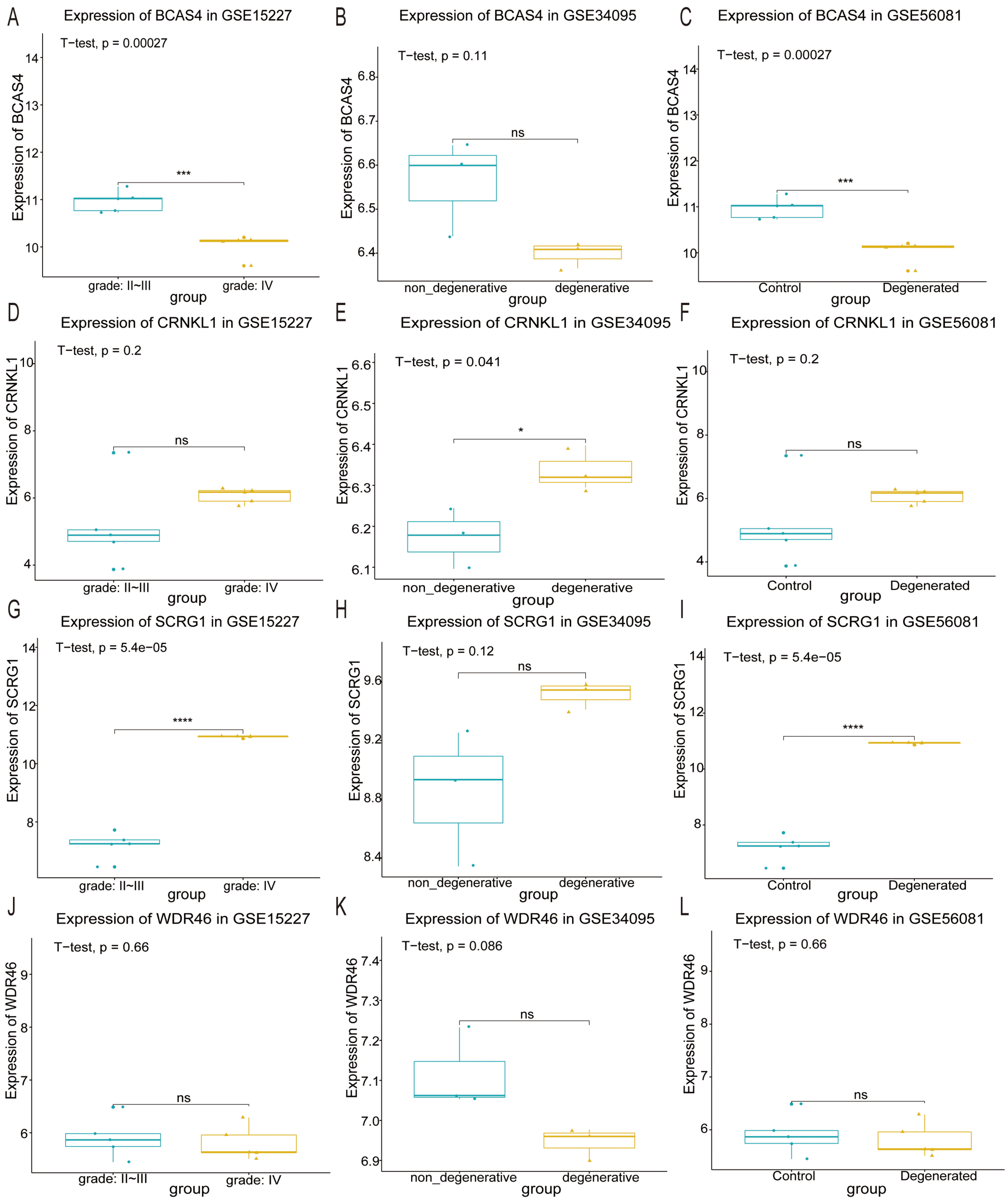

Figure 5 Validation of optimal feature genes. The expression levels of the four optimum feature genes in the GSE15227, GSE34095, and GSE56081 datasets, respectively. (A, B and C) The expression level of BCSA4. (D, E and F) The expression level of CRNKL1. (G, H and I) The expression level of SCRG1. (J, K and L) The expression level of WDR46. * represents *P* value < 0.05, *** represents *P* value < 0.001, **** represents *P* value < 0.0001, and NS represents not significant.

Full-size DOI: 10.7717/peerj.10120/fig-5

sequestering miR-328-5p and upregulating target genes involved in cell proliferation through the ErbB pathway. *Cheng et al. (2018)* reported that circVMA21 acted as a sponge for miR-200c and regulated the activity and function of NP cells by targeting miRNA-200c and *XIAP*, providing a new IDD intervention and treatment strategy.

Wang et al. (2018a) found that circRNA_4099 could act as a sponge for miR-616-5p and eliminate *Sox9* inhibition, increasing ECM secretion. Similarly, Cui & Zhang (2020) reported that circ_001653 silencing may bind to miR-486-3p in order to inhibit *CEMIP* expression, thus attenuating NP cell apoptosis and ECM degradation. In this study, we used integrated analysis to identify that hsa_circRNA_100723, hsa_circRNA_104471, and hsa_circRNA_100750, which have not been previously reported, are more likely to be important molecules involved in IDD regulation, and require further investigation.

CONCLUSION

In this study, we used integrated bioinformatics analysis and machine learning methods to identify *BCAS4* and *SCRG1* as key genes associated with IDD development. Additionally, after constructing the ceRNA network, we found three miRNAs and three circRNAs that may act as important regulators during IDD development by targeting key genes. This novel study may provide new insights into IDD pathogenesis and therapy. Further experiments should be conducted to verify this study's results.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Shanghai Natural Science Foundation (NO.20ZR1452400). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Shanghai Natural Science Foundation: 20ZR1452400.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Hongze Chang performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Xiaolong Yang, Kemin You and Mingwei Jiang performed the experiments, prepared figures and/or tables, and approved the final draft.
- Feng Cai, Yan Zhang, Liang Liu and Hui Liu analyzed the data, prepared figures and/or tables, and approved the final draft.
- Xiaodong Liu conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Raw data is available in the Supplemental Files. The datasets are available from Gene Expression Omnibus (GEO): GSE34095, GSE15227, GSE63492, GSE116726, GSE56081 and GSE67566.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10120#supplemental-information.

REFERENCES

Adams BD, Parsons C, Walker L, Zhang WC, Slack FJ. 2017. Targeting noncoding RNAs in disease. *Journal of Clinical Investigation* **127**:761–771 DOI 10.1172/jci84424.

An H, Boden SD, Kang J, Sandhu HS, Abdu W, Weinstein J. 2003. Summary statement: emerging techniques for treatment of degenerative lumbar disc disease. *Spine* **28**:S24–S25 DOI 10.1097/01.brs.0000076894.33269.19.

Aomatsu E, Takahashi N, Sawada S, Okubo N, Hasegawa T, Taira M, Miura H, Ishisaki A, Chosa N. 2014. Novel SCRG1/BST1 axis regulates self-renewal, migration, and osteogenic differentiation potential in mesenchymal stem cells. *Scientific Reports* **4**:3652 DOI 10.1038/srep03652.

Aruna S, Dr S. 2011. A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer. *International Journal of Computer Applications* **31**:14–20 DOI 10.5120/3795-5226.

Barlund M, Monni O, Weaver JD, Kauraniemi P, Sauter G, Heiskanen M, Kallioniemi OP, Kallioniemi A. 2002. Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. *Genes Chromosomes Cancer* **35**:311–317 DOI 10.1002/gcc.10121.

Battie MC, Videman T, Levalahti E, Gill K, Kaprio J. 2008. Genetic and environmental effects on disc degeneration by phenotype and spinal level: a multivariate twin study. *Spine* **33**:2801–2808 DOI 10.1097/BRS.0b013e31818043b7.

Chen J, Xie JJ, Jin MY, Gu YT, Wu CC, Guo WJ, Yan YZ, Zhang ZJ, Wang JL, Zhang XL, Lin Y, Sun JL, Zhu GH, Wang XY, Wu YS. 2018. Sirt6 overexpression suppresses senescence and apoptosis of nucleus pulposus cells by inducing autophagy in a model of intervertebral disc degeneration. *Cell Death & Disease* **9**:Article 56 DOI 10.1038/s41419-017-0085-5.

Chen K, Wu D, Zhu X, Ni H, Wei X, Mao N, Xie Y, Niu Y, Li M. 2013. Gene expression profile analysis of human intervertebral disc degeneration. *Genetics and Molecular Biology* **36**:448–454 DOI 10.1590/s1415-47572013000300021.

Cheng X, Zhang L, Zhang K, Zhang G, Hu Y, Sun X, Zhao C, Li H, Li YM, Zhao J. 2018. Circular RNA VMA21 protects against intervertebral disc degeneration through targeting miR-200c and X linked inhibitor-of-apoptosis protein. *Annals of the Rheumatic Diseases* **77**:770–779 DOI 10.1136/annrheumdis-2017-212056.

Cui S, Zhang L. 2020. circ_001653 silencing promotes the proliferation and ECM synthesis of NPCs in IDD by downregulating miR-486-3p-mediated CEMIP. *Molecular Therapy - Nucleic Acids* **20**:385–399 DOI 10.1016/j.omtn.2020.01.026.

Dandoy-Dron F, Guillo F, Benboudjema L, Deslys JP, Lasmezas C, Dormont D, Tovey MG, Dron M. 1998. Gene expression in scrapie. Cloning of a new scrapie-responsive gene and the identification of increased levels of seven other mRNA transcripts. *Journal of Biological Chemistry* **273**:7691–7697 DOI 10.1074/jbc.273.13.7691.

Dweep H, Gretz N. 2015. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nature Methods* **12**:697 DOI 10.1038/nmeth.3485.

Feng Y, Egan B, Wang J. 2016. Genetic factors in intervertebral disc degeneration. *Genes and Diseases* **3**:178–185 DOI 10.1016/j.gendis.2016.04.005.

Gruber HE, Hoelscher G, Loeffler B, Chow Y, Ingram JA, Halligan W, Hanley Jr EN. 2009. Prostaglandin E1 and misoprostol increase epidermal growth factor production in 3D-cultured human annulus cells. *Spine Journal* **9**:760–766 DOI 10.1016/j.spinee.2009.04.024.

Guo W, Zhang B, Li Y, Duan HQ, Sun C, Xu YQ, Feng SQ. 2017. Gene expression profile identifies potential biomarkers for human intervertebral disc degeneration. *Molecular Medicine Reports* **16**:8665–8672 DOI 10.3892/mmr.2017.7741.

Guo W, Zhang B, Mu K, Feng SQ, Dong ZY, Ning GZ, Li HR, Liu S, Zhao L, Li Y, Yu BB, Duan HQ, Sun C, Li YJ. 2018. Circular RNA GRB10 as a competitive endogenous RNA regulating nucleus pulposus cells death in degenerative intervertebral disk. *Cell Death & Disease* **9**:Article 319 DOI 10.1038/s41419-017-0232-z.

Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene Selection for cancer classification using support vector machines. *Machine Learning* **46**:389–422 DOI 10.1023/A:1012487302797.

Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. 2018. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics* **15**:41–51 DOI 10.21873/cgp.20063.

Ji ML, Jiang H, Zhang XJ, Shi PL, Li C, Wu H, Wu XT, Wang YT, Wang C, Lu J. 2018. Preclinical development of a microRNA-based therapy for intervertebral disc degeneration. *Nature Communications* **9**:Article 5051 DOI 10.1038/s41467-018-07360-1.

Ji SC, Han N, Liu Y, Li G, Sun Z, Li Z. 2015. Identification of genes associated with disc degeneration using bioinformatics. *Biotechnic and Histochemistry* **90**:353–360 DOI 10.3109/10520295.2015.1007481.

Jiao P, Cai F, Feng Y, Wang W. 2017. Link predication based on matrix factorization by fusion of multi class organizations of the network. *Scientific Reports* **7**:8937–8937 DOI 10.1038/s41598-017-09081-9.

Kalb S, Martirosyan NL, Kalani MY, Broc GG, Theodore N. 2012. Genetics of the degenerated intervertebral disc. *World Neurosurgery* **77**:491–501 DOI 10.1016/j.wneu.2011.07.014.

Kozomara A, Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research* **39**:D152–D157 DOI 10.1093/nar/gkq1027.

Kuhn M. 2015. Caret: classification and regression training. *Astrophysics Source Code Library* **129**:291–295.

Lan PH, Liu ZH, Pei YJ, Wu ZG, Yu Y, Yang YF, Liu X, Che L, Ma CJ, Xie YK, Hu QJ, Wan ZY, Wang HQ. 2016. Landscape of RNAs in human lumbar disc degeneration. *Oncotarget* **7**:63166–63176 DOI 10.18632/oncotarget.11334.

Lee ECS, Elhassan SAM, Lim GPL, Kok WH, Tan SW, Leong EN, Tan SH, Chan EWL, Bhattamisra SK, Rajendran R, Candasamy M. 2019. The roles of circular RNAs in human development and diseases. *Biomedicine and Pharmacotherapy* **111**:198–208 DOI 10.1016/j.biopha.2018.12.052.

Li CY, Zhang WW, Xiang JL, Wang XH, Wang JL, Li J. 2019. Integrated analysis highlights multiple long non-coding RNAs and their potential roles in the progression of human esophageal squamous cell carcinoma. *Oncology Reports* **42**:2583–2599 DOI 10.3892/or.2019.7377.

Libbrecht MW, Noble WS. 2015. Machine learning applications in genetics and genomics. *Nature Reviews Genetics* **16**:321–332 DOI 10.1038/nrg3920.

Liu X, Che L, Xie YK, Hu QJ, Ma CJ, Pei YJ, Wu ZG, Liu ZH, Fan LY, Wang HQ. 2015. Noncoding RNAs in human intervertebral disc degeneration: an integrated microarray study. *Genome Data* **5**:80–81 DOI 10.1016/j.gdata.2015.05.027.

Liu H, Huang X, Liu X, Xiao S, Zhang Y, Xiang T, Shen X, Wang G, Sheng B. 2014. miR-21 promotes human nucleus pulposus cell proliferation through PTEN/AKT signaling. *International Journal of Molecular Sciences* **15**:4007–4018 DOI 10.3390/ijms15034007.

Momtaz R, Ghanem NM, El-Makky NM, Ismail MA. 2018. Integrated analysis of SNP, CNV and gene expression data in genetic association studies. *Clinical Genetics* **93**:557–566 DOI 10.1111/cge.13092.

Obermeyer Z, Emanuel EJ. 2016. Predicting the future —big data, machine learning, and clinical medicine. *New England Journal of Medicine* **375**:1216–1219 DOI 10.1056/NEJMp1606181.

Ochi K, Derfoul A, Tuan RS. 2006. A predominantly articular cartilage-associated gene, SCRG1, is induced by glucocorticoid and stimulates chondrogenesis in vitro. *Osteoarthritis and Cartilage* **14**:30–38 DOI 10.1016/j.joca.2005.07.015.

Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych N, McMurry J, Marioni JC, Malone J, Megy K, Rustici G, Tang AY, Taubert J, Williams E, Mannion O, Parkinson HE, Brazma A. 2014. Expression Atlas update–a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Research* **42**:D926–D932 DOI 10.1093/nar/gkt1270.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842 DOI 10.1093/bioinformatics/btq033.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43(7)**:e47 DOI 10.1093/nar/gkv007.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**:77 DOI 10.1186/1471-2105-12-77.

Silva D, Antunes J, Balamurugan K, Duncan G, Alho CS, McCord B. 2016. Developmental validation studies of epigenetic DNA methylation markers for the detection of blood, semen and saliva samples. *Forensic Science International: Genetics* **23**:55–63 DOI 10.1016/j.fsigen.2016.01.017.

Song YQ, Karasugi T, Cheung KM, Chiba K, Ho DW, Miyake A, Kao PY, Sze KL, Yee A, Takahashi A, Kawaguchi Y, Mikami Y, Matsumoto M, Togawa D, Kanayama M, Shi D, Dai J, Jiang Q, Wu C, Tian W, Wang N, Leong JC, Luk KD, Yip SP, Cherny SS, Wang J, Mundlos S, Kelempisioti A, Eskola PJ, Männikkö M, Mäkelä P, Karppinen J, Järvelin MR, O'Reilly PF, Kubo M, Kimura T, Kubo T, Toyama Y, Mizuta H, Cheah KS, Tsunoda T, Sham PC, Ikegawa S, Chan D. 2013. Lumbar disc degeneration is linked to a carbohydrate sulfotransferase 3 variant. *Journal of Clinical Investigation* **123**:4909–4917 DOI 10.1172/jci69277.

Taki T, Kibayashi K. 2015. Characterization of cellular and extracellular DNA in saliva. *Legal Medicine* **17**:471–474 DOI 10.1016/j.legalmed.2015.10.003.

Tan H, Zhao L, Song R, Liu Y, Wang L. 2018. The long noncoding RNA SNHG1 promotes nucleus pulposus cell proliferation through regulating miR-326 and CCND1. *American Journal of Physiology-Cell Physiology* **315**:C21–C27 DOI 10.1152/ajpcell.00220.2017.

Tay Y, Rinn J, Pandolfi P. 2014. The multilayered complexity of ceRNA crosstalk and competition. *Nature* **505**:344–352 DOI 10.1038/nature12986.

Tian C, Deng Y, Jin Y, Shi S, Bi H. 2018. Long non-coding RNA RNCR3 promotes prostate cancer progression through targeting miR-185-5p. *American Journal of Translational Research* **10**:1562–1570.

Tsai TT, Lai PL, Liao JC, Fu TS, Niu CC, Chen LH, Lee MS, Chen WJ, Fang HC, Ho NY, Pang JH. 2013. Increased periostin gene expression in degenerative intervertebral disc cells. *Spine Journal* **13**:289–298 DOI 10.1016/j.spinee.2013.01.040.

Vergroesen PP, Kingma I, Emanuel KS, Hoogendoorn RJ, Welting TJ, Van Royen BJ, Van Dieen JH, Smit TH. 2015. Mechanics and biology in intervertebral disc degeneration: a vicious circle. *Osteoarthritis Cartilage* **23**:1057–1070 DOI 10.1016/j.joca.2015.03.028.

Wan ZY, Song F, Sun Z, Chen YF, Zhang WL, Samartzis D, Ma CJ, Che L, Liu X, Ali MA, Wang HQ, Luo ZJ. 2014. Aberrantly expressed long noncoding RNAs in human intervertebral disc degeneration: a microarray related study. *Arthritis Research & Therapy* **16**:Article 465 DOI 10.1186/s13075-014-0465-5.

Wang H, He P, Pan H, Long J, Wang J, Li Z, Liu H, Jiang W, Zheng Z. 2018a. Circular RNA circ-4099 is induced by TNF- α and regulates ECM synthesis by blocking miR-616-5p inhibition of Sox9 in intervertebral disc degeneration. *Experimental & Molecular Medicine* **50**:1–14 DOI 10.1038/s12276-018-0056-7.

Wang Z, Li Y, Wang Y, Wang X, Zhang J, Tian J. 2018b. Association between GDF5 single nucleotide polymorphism rs143383 and lumbar disc degeneration. *Experimental and Therapeutic Medicine* **16**:1900–1904 DOI 10.3892/etm.2018.6382.

Wang HQ, Yu XD, Liu ZH, Cheng X, Samartzis D, Jia LT, Wu SX, Huang J, Chen J, Luo ZJ. 2011. Deregulated miR-155 promotes Fas-mediated apoptosis in human intervertebral disc degeneration by targeting FADD and caspase-3. *Journal of Pathology* **225**:232–242 DOI 10.1002/path.2931.

Williams FM, Bansal AT, Meurs JBvan, Bell JT, Meulenbelt I, Suri P, Rivadeneira F, Sambrook PN, Hofman A, Bierma-Zeinstra S, Menni C, Kloppenburg M, Slagboom PE, Hunter DJ, MacGregor AJ, Uitterlinden AG, Spector TD. 2013. Novel genetic variants associated with lumbar disc degeneration in northern Europeans: a meta-analysis of 4600 subjects. *Annals of the Rheumatic Diseases* **72**:1141–1148 DOI 10.1136/annrheumdis-2012-201551.

Xu YQ, Zhang ZH, Zheng YF, Feng SQ. 2016. Dysregulated miR-133a Mediates Loss of Type II Collagen by Directly Targeting Matrix Metalloproteinase 9 (MMP9) in Human Intervertebral Disc Degeneration. *Spine* **41**:E717-E724 DOI 10.1097/brs.0000000000001375.

Yuan B, Ji W, Fan B, Zhang B, Zhao Y, Li J. 2018. Association analysis between thrombospondin-2 gene polymorphisms and intervertebral disc degeneration in a Chinese Han population. *Medicine* **97**:e9586 DOI 10.1097/md.0000000000009586.

Zaina F, Tomkins-Lane C, Carragee E, Negrini S. 2016. Surgical versus non-surgical treatment for lumbar spinal stenosis. *Cochrane Database of Systematic Reviews* **2016(1)**:CD010264 DOI 10.1002/14651858.CD010264.pub2.

Zhang F, Zhao X, Shen H, Zhang C. 2016. Molecular mechanisms of cell death in intervertebral disc degeneration (Review). *International Journal of Molecular Medicine* **37**:1439–1448 DOI 10.3892/ijmm.2016.2573.

Zhang Y, Zhang R, Ding X, Ai K. 2018. EFNB2 acts as the target of miR-557 to facilitate cell proliferation, migration and invasion in pancreatic ductal adenocarcinoma by bioinformatics analysis and verification. *American Journal of Translational Research* **10**:3514–3528.

Zhao B, Lu M, Wang D, Li H, He X. 2016. Genome-wide identification of long noncoding RNAs in human intervertebral disc degeneration by RNA sequencing. *BioMed Research International* **2016**:3684875 DOI 10.1155/2016/3684875.

Zhu Y, Qiao L, Zhou Y, Ma N, Wang C, Zhou J. 2018. Long non-coding RNA FOXD2-AS1 contributes to colorectal cancer proliferation through its interaction with microRNA-185-5p. *Cancer Science* **109**:2235–2242 DOI 10.1111/cas.13632.