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

Hongze Chang [1], **Xiaolong Yang** [1], **Kemin You** [1], **Mingwei Jiang** [1], **Feng Cai** [1], **Yan Zhang** [1], **Liang Liu** [1], **Hui Liu** [1], **Xiaodong Liu** [Corresp. 1]

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

Corresponding Author: Xiaodong Liu
Email address: xiaodong.liu@tongji.edu.cn

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.

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

5 Hongze Chang, Xiaolong Yang, Kemin You, Mingwei Jiang, Feng Cai, Yan Zhang, Liang Liu, Hui Liu,
6 Xiaodong Liu*
7
8 Department of Orthopedics, Shanghai Yangpu Hospital Affiliated to Tongji University, 450 Tengyue Road,
9 Shanghai 200090, China
10
11 * Corresponding author:
12 Xiaodong Liu (xiaodong.liu@tongji.edu.cn)
13 Department of Orthopedics, Shanghai Yangpu Hospital Affiliated to Tongji University, 450 Tengyue Road,
14 Shanghai 200090, China
15

17 Running title: Integrated analysis and machine learning methods reveal key genes and underlying mechanisms
18 underlying IDD.
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

42
43
44
45   **Abstract:**
46   Intervertebral disc degeneration (IDD), a major cause of lower back pain, has multiple contributing factors
47   including genetics, environment, age, and loading history. Bioinformatics analysis has been extensively used to
48   identify diagnostic biomarkers and therapeutic targets for IDD diagnosis and treatment. However, multiple
49   microarray dataset analysis and machine learning methods have not been integrated. In this study, we
50   downloaded the mRNA, microRNA (miRNA), long noncoding RNA (lncRNA), and circular RNA (circRNA)
51   expression profiles (GSE34095, GSE15227, GSE63492, GSE116726, GSE56081, and GSE67566) associated
52   with IDD from the GEO database. Using differential expression analysis and recursive feature elimination, we
53   extracted four optimal feature genes. We then used the support vector machine (SVM) to make a classification
54   model with the four optimal feature genes. The ROC curve was used to evaluate the model's performance, and
55   the expression profiles (GSE63492, GSE116726, GSE56081, and GSE67566) were used to construct a
56   competitive endogenous RNA (ceRNA) regulatory network and explore the underlying mechanisms of the
57   feature genes. We found that three miRNAs (hsa-miR-4728-5p, hsa-miR-5196-5p, and hsa-miR-185-5p) and
58   three circRNAs (hsa_circRNA_100723, hsa_circRNA_104471, and hsa_circRNA_100750) were important
59   regulators with more interactions than the other RNAs across the whole network. The expression level analysis
60   of the three datasets revealed that *BCAS4* and *SCRG1* were key genes involved in IDD development. Ultimately,
61   our study proposes a novel approach to determining reliable and effective targets in IDD diagnosis and treatment.
62

63   **Keywords:**
64   Intervertebral disc degeneration; Integrated analysis; Machine learning methods; Key genes
65

66   **Introduction**
67   Lower back pain, commonly caused by intervertebral disc degeneration (IDD), can be a significant
68   socioeconomic burden on patients (Vergroesen et al. 2015). IDD is characterized by the apoptosis of nucleus
69   pulposus (NP) cells, the degradation of extracellular matrix (ECM) components, and several contributing factors
70   including genetics and environment (Battie et al. 2008; Feng et al. 2016). However, the precise etiology of IDD
71   remains largely unknown. Diagnosing degenerative disc disease is difficult, and common IDD treatment and
72   management strategies primarily consist of   conservative management or surgical treatment to relieve pain,
73   without resolving the underlying tissue pathology (An et al. 2003; Zaina et al. 2016). To identify potential
74   biomarkers and specific therapeutic targets, a more detailed understanding of the molecular and cellular events
75   underlying IDD formation is needed.
76
77   The etiology of IDD is complex, but over the past several decades, it has become clear that genetic factors are
78   most dominant. Multiple candidate genes, including *thrombospondin-2*, vitamin D receptor, *COL2A1*, *ACAN*,
79   interleukins (*IL1α, IL 1β*, and *IL6*), matrix metalloproteinases (*MMP-3* and *MMP-9*), and growth/differentiation
80   factor 5, have been associated with the pathophysiological process of IDD development (Feng et al. 2016; Kalb
81   et al. 2012; Wang et al. 2018b; Yuan et al. 2018). Genome-wide association studies (GWAS) have been used to
82   help identify novel variants. Several GWAS found that the genetic polymorphisms of *PARK2* and *CHST3* were

83   relevant to IDD etiology (Song et al. 2013; Williams et al. 2013). Additionally, some bioinformatics analyses
84   based on gene expression profiles revealed that *FYN, PRKCD, YWHAB, YWHAZ, AR, Fibronectin 1, COL2A1,*
85   *β-catenin, COL6A2, IBSP, RAP1A,* and *FOXF2* genes may play key roles in IDD development (Chen et al. 2013;
86   Guo et al. 2017; Ji et al. 2015). Although a considerable number of genes associated with IDD development have
87   been found, early IDD diagnosis and precise treatment remain difficult and require further study.
88

89   Integrated analyses using combined multiple microarray data can provide a more accurate understanding of the
90   interplay across multi-level genomic features and the molecular mechanisms that cause complex diseases
91   (Momtaz et al. 2018). Machine learning is a type of artificial intelligence that can "learn" a model using past
92   data in order to predict future data. Machine learning algorithms have been used in key feature training,
93   recognition, and group classification (Huang et al. 2018). Modern researchers have unprecedented access to
94   machine learning methods that can elucidate complex molecular mechanisms and predict disease genes from
95   large biomedical datasets (Libbrecht & Noble 2015; Obermeyer & Emanuel 2016). To the best of our knowledge,
96   no investigations have used integrated analysis and machine learning methods to identify key IDD-associated
97   genes.
98

99   Accumulating evidence has indicated that noncoding RNAs, including microRNAs (miRNAs), long noncoding
100   RNAs (lncRNAs), and circular RNAs (circRNAs), are important gene expression regulators that influence
101   cellular function and disease states (Adams et al. 2017). lncRNAs and circRNAs act as competitive endogenous
102   RNAs (ceRNAs) and competitively bind miRNA response elements (MREs) to construct a regulatory network
103   involved in IDD progression. Zhao et al. (2016) used RNA sequencing to identify 1,854 lncRNAs and 2,804
104   protein-coding genes that were differentially expressed in the IDD group. Tan et al. (2018) found that
105   LncSNHG1 promoted NP cell proliferation by suppressing miR-326 expression and upregulating *CCND1*
106   expression. Recent studies have focused on the functional roles of circRNAs during IDD development and found
107   that circ-4099, circ-GRB10, circVMA21, and circ_001653 play pivotal roles during NP cell proliferation,
108   apoptosis, and extracellular matrix synthesis/degradation (Cheng et al. 2018; Cui & Zhang 2020; Guo et al. 2018;
109   Wang et al. 2018a). The diverse biological functions of ceRNAs deserve further exploration.
110

111   In this study, we aimed to identify the key genes and underlying mechanisms of IDD development by
112   constructing an lncRNA/circRNA-miRNA-mRNA network using multiple microarray datasets and machine
113   learning methods. Fig. S1 shows the flow chart for this study. Our results present novel biomarkers and
114   therapeutic targets that can be used for IDD diagnosis and treatment.
115

116   **Materials & Methods**

117

118   **Microarray datasets**
119   We retrieved a total of six expression profiles from the GEO database (www.ncbi.Nlm.nih.gov/geo): two mRNA
120   expression profiles (GSE34095 (Tsai et al. 2013) and GSE15227 (Gruber et al. 2009)), two miRNA expression
121   profiles (GSE63492 (Lan et al. 2016) and GSE116726 (Ji et al. 2018)), one mRNA-lncRNA expression profile
122   (GSE56081 (Lan et al. 2016; Wan et al. 2014)), and one circRNA expression profile (GSE67566 (Lan et al.
123   2016; Liu et al. 2015)). Table 1 contains the basic information for these expression profiles.

124 **Differential expression analysis**

125 The raw data were annotated, normalized, log 2 transformed, and screened for differentially expressed genes
126 (DEGs), differentially expressed miRNAs (DEMs), differentially expressed lncRNAs (DELs), and differentially
127 expressed circRNAs (DECs) in IDD and normal disc tissue using the "Limma" R package (Ritchie et al. 2015).
128 Since each dataset came from a different experiment and microarray platform, the way to get the data may be
129 different. When filtering using consistent thresholds, some datasets did not yield valid differential genes. In order
130 to obtain effective differentially expressed genes for subsequent analysis, we used different thresholds according
131 to each dataset's conditions. The specific DEG screening thresholds were as follows: p-value < 0.05 for the
132 GSE34095 dataset, p-value < 0.01 and |log2FC| ≥ 2 for the GSE15227 dataset, and p-value < 0.01 and |log2 FC|
133 > 2 for the GSE116726 dataset. We obtained the GSE67566, GSE63492, and GSE56081 datasets from the same
134 tissue samples, and provided the differential expression analysis results and thresholds as Supplemental
135 Materials (Tables S1 and S2, Documents S1and S2). We visualized the differential expression analysis results
136 using a volcano plot and heatmap with hierarchical clustering (Fig. S2).

137

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

146

147

148 **SVM model verification**
149 The support vector machine (SVM) is particularly effective for binary classification in a supervised learning
150 manner, and is better than other machine learning methods at identifying subtle patterns in complex datasets
151 (Aruna & Dr 2011). Radial basis function kernel (RBF kernel) is commonly used in nonlinear support vector
152 machine classification since it can enable data to operate in a high-dimensional and implicit feature space (Jiao
153 et al. 2017). In this study, we used the R e1071 package (https://CRAN.R-project.org/package=e1071) to build
154 an RBF kernel SVM model to identify the optimal feature genes (Dimitriadou et al. 2012), and we selected the
155 GSE15227 dataset to train the machine. We then evaluated its performance in the GSE15227 training set using
156 the R pROC package (http://www.biomedcentral.com/1471-2105/12/77/) (Robin et al. 2011). The optimal
157 feature genes were taken from the training set, and over-fitting would occur whenever ROC verification was
158 performed. Therefore, we used two validation sets (GSE34095 and GSE56081) to further verify the model's
159 performance.

160

161

162 **Identifying target miRNAs of the optimal feature genes**
163 We used miRWalk 2.0 with 12 prediction programs (MIMATid, Microt4, miRanda, mirbridge, miRDB,
164 miRMap, miRNAMap, Pictar2, PITA, RNA22, RNAhybrid, and Targetscan) to predict the optimal feature

165   genes' target miRNAs (Dweep & Gretz 2015). miRNAs present in more than five of the 12 prediction programs
166   were considered as target miRNAs. We selected overlapping DEMs in the GSE116726 and GSE63492 datasets
167   as candidate DEMs. Intersecting target miRNAs and candidate DEMs were selected as optimal feature gene
168   interaction pairs.
169
170   **ceRNA network construction**
171   We obtained the miRNA sequences interacting with optimal feature genes from the miRbase (Kozomara &
172   Griffiths-Jones 2011) and extracted mature sequences using the Perl program. The DEL sequences were
173   downloaded from NCBI. For DELs with several transcripts, we selected the longest transcripts for subsequent
174   analysis. We used the BEDTOOLS command (Quinlan & Hall 2010) and genomic coordinates to obtain the
175   DEC sequences, and converted the gene name to its circRNA symbol using the Perl program.
176
177   We then used the miRanda tool to analyze the combinations of miRNAs, DELs, and DECs. We set the analysis
178   parameters for the miRNA-DELs and miRNA-DECs as sc:120, en:-20 and sc:150, en:-7, respectively, and
179   processed the results using python script (Documents S3 and S4). The DELs and DECs with ≥5 miRNA binding
180   sites were identified as reliable miRNA-lncRNA and miRNA-circRNA interaction pairs. The ceRNA
181   (DELs/DECs-miRNA-optimal feature gene) regulatory network was constructed using a combination of
182   miRNA-DEL pairs, miRNA-DEC pairs, and miRNA-optimal feature gene pairs. We visualized the network
183   using Cytoscape 3.6.0 (http://www.cytoscape.org/).
184
185   **Statistical analysis**
186   All data were presented as the mean±SEM. Differential expression levels were compared using the Student's t-
187   test in GraphPad Prism 7.0 (GraphPad Software Inc., La Jolla, CA, USA). P values < 0.05 were considered
188   statistically significant.
189
190
191   **Results**
192   **Differentially expressed genes in IDD**
193   We obtained a total of 334 DEGs (199 up-regulated and 135 down-regulated) in the GSE34095 dataset (Fig. 1A)
194   and 188 DEGs (141 up-regulated and 47 down-regulated) in the GSE15227 dataset (Fig. 1B). Hierarchical
195   cluster heatmaps showed that these DEGs could distinguish between the degenerative disc samples and the
196   control disc samples (Fig. 1C and 1D). We obtained a total of 13 overlapping DEGs (*COL3A1*, *SCRG1*, *HTRA1*,
197   *BCAS4*, *C11orf80*, *CRNKL1*, *GREM1*, *FGFR3*, *BDKRB1*, *WDR46*, *FN1*, *LMF2*, and *GDI2*) via the intersection
198   between the two datasets, and considered these as IDD-related DEGs (Fig. 1E).
199
200   **Optimal feature gene selection**
201   We used random forest and neural network models to evaluate the importance of the 13 overlapping DEGs
202   (Documents S5 and S6). The results showed that most of the DEGs from the two models had minimal differences
203   in terms of gene importance rankings (Fig. 2A and 2B). Using the RFE method, we identified four important
204   genes (*WDR46*, *BCAS4*, *CRNKL1*, and *SCRG1*) as optimal feature genes associated with IDD (Fig. 2C).
205

206 A classification model was constructed using the RBF kernel SVM, the four genes as features, and the GSE15227
207 dataset to train the machine. The parameters were set as: SVM-Kernel: radial, cost:1, gamma:0.25, and
208 epsilon:0.1. Additionally, their performance was assessed using the R pROC package. In the GSE15227 training
209 set, the area under the ROC curve (AUC) of the SVM model was 100 percent, suggesting that the model could
210 accurately distinguish between IDD and normal samples (Fig. 2D). We used the GSE34095 and GSE56081
211 datasets as validation sets to further evaluate the model's performance and to help avoid over-fitting in the
212 training set. The SVM model in the GSE34095 validation set had an AUC of 55.6 percent, which may be
213 correlated with the small sample size of the microarray dataset (Fig. 2E). However, the AUC for the GSE56081
214 validation set was 100 percent (Fig. 2F). These results indicated that the optimal feature genes (*WDR46, BCAS4*,
215 *CRNKL1*, and *SCRG1*) could be used as effective and accurate IDD diagnostic biomarkers.
216
217
218 **Identifying target miRNAs of the optimal feature genes**
219 We predicted a total of 467 target miRNA optimal feature genes using miRWalk 2.0. Compared to the control
220 disc samples, we identified 724 DEMs (527 up-regulated and 197 down-regulated) from the GSE116726 dataset
221 (Fig. 3A and 3B) and 149 DEMs from the GSE63492 dataset. Fig. 3C shows a Venn diagram of the 46 common
222 DEMs that were found. Based on the intersections between the target miRNAs and common DEMs, we further
223 analyzed 12 overlapping miRNAs which we considered to be the optimal feature genes' interaction miRNAs
224 (Fig. 3D).
225
226 **ceRNA network construction**
227 The differentially expressed lncRNAs and circRNAs between the degenerated disc samples and the control disc
228 samples were downloaded from the GEO database. We analyzed the miRNA-DEL interactions and the miRNA-
229 DEC interactions using the miRanda tool with ≥5 miRNA binding sites. The miRNA-DEL pairs, miRNA-DEC
230 pairs, and miRNA-optimal feature gene pairs were combined to build a DELs/DECs-DEMs-optimal feature gene
231 regulatory network, which included four mRNA nodes, 12 miRNA nodes, 10 lncRNA nodes, and 75 circRNA
232 nodes (Fig. 4, Table S3). Out of these, three miRNAs (hsa-miR-4728-5p, hsa-miR-5196-5p, and hsa-miR-185-
233 5p) and three circRNAs (hsa_circRNA_100723, hsa_circRNA_104471, and hsa_circRNA_100750) were key
234 regulators, based on the optimal feature genes and connective degrees of the whole network (Table 2). However,
235 the LncRNA connective degree was very low.
236
237 **Validating optimal feature genes**
238 The expression of the four optimal feature genes across the three datasets was visualized using box plots. In the
239 GSE15227 and GSE56081 datasets, *BCAS4* expression levels were significantly down-regulated (Fig.5A and
240 5C), and *SCRG1* was significantly up-regulated (Fig. 5G and 5I). *CRNKL1* was only up-regulated in the
241 GSE34095 dataset (Fig. 5E). Unfortunately, the expression levels of *WDR46* showed no significant difference
242 across the three datasets (Fig. 5J,5K and 5L). These results indicated that *BCAS4* and *SCRG1* are key genes
243 involved in IDD development.
244
245 **Discussion**
246 IDD is linked to lower back pain and spine-related diseases, and although IDD's underlying mechanisms have

247   been studied for many years, they still remain unclear. Insufficient early IDD diagnosis and treatment methods
248   affect the quality of life for patients and impose a heavy economic burden on society (Vergroesen et al. 2015).
249   NP cells play an important role in maintaining intervertebral disc homeostasis by synthesizing ECM, which
250   includes aggrecan and type II collagen (Zhang et al. 2016). Recent studies show that targeting gene therapy can
251   inhibit NP cell senescence and apoptosis, and can ultimately ameliorate IDD (Chen et al. 2018). Therefore,
252   reliable and specific gene targets are essential for IDD diagnosis and treatment.
253
254   Recent developments in bioinformatics and computational biology have led to the identification of several key
255   genetic targets related to IDD and the prediction of potential molecular mechanisms (Petryszak et al. 2014). In
256   this study, we downloaded multiple microarray datasets associated with IDD from the GEO database, including
257   two mRNA expression profiles, two miRNA expression profiles, one mRNA-lncRNA expression profile, and
258   one circRNA expression profile. A total of four optimal feature genes (*WDR46*, *BCAS4*, *CRNKL1*, and *SCRG1*)
259   were identified using machine-learning methods. The construction of a DELs/DECs-miRNA-optimal feature
260   genes network revealed that three miRNAs (hsa-miR-4728-5p, hsa-miR-5196-5p, and hsa-miR-185-5p) and
261   three circRNAs (hsa_circRNA_100723, hsa_circRNA_104471, and hsa_circRNA_100750) may be important
262   mediators for optimal feature genes. To further explore the different optimal feature genes in normal and
263   degenerative NP tissues, we also investigated their expression levels across three datasets. The results indicated
264   that *BCAS4* and *SCRG1* were key genes related to IDD.
265
266   Breast carcinoma amplified sequence 4 (*BCAS4*), a novel gene cloned from breast cancer cells, encodes a 211-
267   amino acid cytoplasmic protein with no significant homologies to any known protein (Barlund et al. 2002).
268   Previous studies have demonstrated that the specific DNA methylation of *BCAS4* acts as an epigenetic marker
269   and can be used to distinguish saliva from other body fluids. It is also widely used in forensic investigations
270   (Silva et al. 2016; Taki & Kibayashi 2015). However, *BCAS4*'s biological role in disease requires further
271   investigation. Stimulator of chondrogenesis 1 (*SCRG1*) was first found in the genes associated with, or
272   responsible for, the neurodegenerative changes observed in transmissible spongiform encephalopathies
273   (Dandoy-Dron et al. 1998). *SCRG1* transcript is found in the brain, heart, and spinal cord, and its sequence is
274   highly conserved in humans, mice, and rats. *SCRG1* has also been observed to be specifically expressed in human
275   articular cartilage, and is involved in human mesenchymal stem cell (hMSC) growth suppression and
276   differentiation during dexamethasone-dependent chondrogenesis (Ochi et al. 2006). Recent studies have shown
277   that *SCRG1* is an important regulator during hMSC self-renewal, migration, and osteogenic differentiation along
278   with its receptor *BST1* (Aomatsu et al. 2014). However, the function of *SCRG1* in IDD development has not yet
279   been explored.
280
281   circRNAs and lncRNAs may act as ceRNAs by competitively binding to miRNA and suppressing mRNA
282   expression (Adams et al. 2017; Tay et al. 2014). This ceRNA hypothesis suggests that there is a novel mechanism
283   for RNA interactions. In this study's ceRNA analysis, hsa-miR-4728-5p, hsa-miR-5196-5p, hsa-miR-185-5p,
284   hsa_circRNA_100723, hsa_circRNA_104471, and hsa_circRNA_100750 showed more interactions compared
285   to the other RNAs in the whole network. Additionally, miR-155, miR-21, and miR-133a were shown to be
286   differentially expressed in degenerative NP cells, indicating that they may be potential biomarkers for early IDD
287   diagnosis (Liu et al. 2014; Wang et al. 2011; Xu et al. 2016). The dysregulation of these miRNAs is closely

288   associated with NP cell apoptosis which affects IDD progression. miR-185-5p has been reported as a critical
289   regulator, but miR-4728-5p and miR-5196-5p have rarely been reported. Chang et al. (2017) reported that miR-
290   185-5p induced by *Runx2* could directly target *Dlx2* to inhibit amylogenesis and osteogenesis, providing a new
291   treatment option for cleidocranial dysplasia. Multiple lncRNAs bind to miR-185-5p in order to modulate the
292   progression of different types of human cancer, including prostate cancer (Tian et al. 2018), colorectal cancer
293   (Zhu et al. 2018), and glioblastoma (Wang et al. 2018b). However, the regulatory effects of miR-185-5p on IDD
294   require further investigation.
295
296   CircRNAs, a new type of ncRNAs formed by special loop splicing, are thought to be potential diagnostic
297   biomarkers and therapeutic targets because they are more stable and conserved than other RNAs (Lee et al.
298   2019). However, very few studies have focused on the role of circRNAs in IDD development. Guo et al. (2018)
299   found that circ-GRB10 was downregulated in IDD cells, and circ-GRB10 overexpression inhibited NP cell
300   apoptosis by sequestering miR-328-5p and upregulating target genes involved in cell proliferation through the
301   ErbB pathway. Cheng et al. (2018) reported that circVMA21 acted as a sponge for miR-200c and regulated the
302   activity and function of NP cells by targeting miRNA-200c and *XIAP*, providing a new IDD intervention and
303   treatment strategy. Wang et al. (2018a) found that circRNA_4099 could act as a sponge for miR-616-5p and
304   eliminate *Sox9* inhibition, increasing ECM secretion. Similarly, Cui & Zhang (2020) reported that circ_001653
305   silencing may bind to miR-486-3p in order to inhibit *CEMIP* expression, thus attenuating NP cell apoptosis and
306   ECM degradation. In this study, we used integrated analysis to identify that hsa_circRNA_100723,
307   hsa_circRNA_104471, and hsa_circRNA_100750, which have not been previously reported, are more likely to
308   be important molecules involved in IDD regulation, and require further investigation.
309
310   **Conclusion**
311   In this study, we used integrated bioinformatics analysis and machine learning methods to identify *BCAS4* and
312   *SCRG1* as key genes associated with IDD development. Additionally, after constructing the ceRNA network,
313   we found three miRNAs and three circRNAs that may act as important regulators during IDD development by
314   targeting key genes. This novel study may provide new insights into IDD pathogenesis and therapy. Further
315   experiments should be conducted to verify this study's results.
316
317
318   **Acknowledgements**
319   Not applicable.
320
321   **Funding**
322   This work was supported by Shanghai Natural Science Foundation(NO.20ZR1452400).
323
324   **Disclosures**
325   The authors have no financial conflicts of interest.
326
327   **Availability of data**
328   The datasets used and/or analyzed in the present study are available from the public Gene Expression Omnibus

329   (GEO) database.

330

331

332

333

334

335

336   **References**

337

338   Adams BD, Parsons C, Walker L, Zhang WC, and Slack FJ. 2017. Targeting noncoding RNAs in disease. *J Clin Invest*

339   127:761-771. 10.1172/jci84424

340   An H, Boden SD, Kang J, Sandhu HS, Abdu W, and Weinstein J. 2003. Summary statement: emerging techniques for

341   treatment of degenerative lumbar disc disease. *Spine (Phila Pa 1976)* 28:S24-25.

342   10.1097/01.brs.0000076894.33269.19

343   Aomatsu E, Takahashi N, Sawada S, Okubo N, Hasegawa T, Taira M, Miura H, Ishisaki A, and Chosa N. 2014. Novel

344   SCRG1/BST1 axis regulates self-renewal, migration, and osteogenic differentiation potential in

345   mesenchymal stem cells. *Sci Rep* 4:3652. 10.1038/srep03652

346   Aruna S, and Dr S. 2011. A Novel SVM based CSSFFS Feature Selection Algorithm for Detecting Breast Cancer.

347   *International Journal of Computer Applications* 31:14-20.

348   Barlund M, Monni O, Weaver JD, Kauraniemi P, Sauter G, Heiskanen M, Kallioniemi OP, and Kallioniemi A. 2002.

349   Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and

350   fusion in breast cancer. *Genes Chromosomes Cancer* 35:311-317. 10.1002/gcc.10121

351   Battie MC, Videman T, Levalahti E, Gill K, and Kaprio J. 2008. Genetic and environmental effects on disc

352   degeneration by phenotype and spinal level: a multivariate twin study. *Spine (Phila Pa 1976)* 33:2801-2808.

353   10.1097/BRS.0b013e31818043b7

354   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,

355   Wang XY, and Wu YS. 2018. Sirt6 overexpression suppresses senescence and apoptosis of nucleus pulposus

356   cells by inducing autophagy in a model of intervertebral disc degeneration. *Cell Death Dis* 9:56.

357   10.1038/s41419-017-0085-5

358   Chen K, Wu D, Zhu X, Ni H, Wei X, Mao N, Xie Y, Niu Y, and Li M. 2013. Gene expression profile analysis of

359   human intervertebral disc degeneration. *Genet Mol Biol* 36:448-454. 10.1590/s1415-47572013000300021

360   Cheng X, Zhang L, Zhang K, Zhang G, Hu Y, Sun X, Zhao C, Li H, Li YM, and Zhao J. 2018. Circular RNA VMA21

361   protects against intervertebral disc degeneration through targeting miR-200c and X linked inhibitor-of-

362   apoptosis protein. *Ann Rheum Dis* 77:770-779. 10.1136/annrheumdis-2017-212056

363   Cui S, and Zhang L. 2020. circ_001653 Silencing Promotes the Proliferation and ECM Synthesis of NPCs in IDD by

364   Downregulating miR-486-3p-Mediated CEMIP. *Mol Ther Nucleic Acids* 20:385-399.

365   10.1016/j.omtn.2020.01.026

366   Dandoy-Dron F, Guillo F, Benboudjema L, Deslys JP, Lasmezas C, Dormont D, Tovey MG, and Dron M. 1998. Gene

367   expression in scrapie. Cloning of a new scrapie-responsive gene and the identification of increased levels of

368   seven other mRNA transcripts. *J Biol Chem* 273:7691-7697. 10.1074/jbc.273.13.7691

369   Dimitriadou E, Hornik K, Leisch F, Meyer D, and Weingessel A. 2012. Misc functions of the Department of Statistics

370   (e1071), TU Wien.

371   Dweep H, and Gretz N. 2015. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat Methods*
372   12:697. 10.1038/nmeth.3485

373   Feng Y, Egan B, and Wang J. 2016. Genetic Factors in Intervertebral Disc Degeneration. *Genes Dis* 3:178-185.
374   10.1016/j.gendis.2016.04.005

375   Gruber HE, Hoelscher G, Loeffler B, Chow Y, Ingram JA, Halligan W, and Hanley EN, Jr. 2009. Prostaglandin E1
376   and misoprostol increase epidermal growth factor production in 3D-cultured human annulus cells. *Spine J*
377   9:760-766. 10.1016/j.spinee.2009.04.024

378   Guo W, Zhang B, Li Y, Duan HQ, Sun C, Xu YQ, and Feng SQ. 2017. Gene expression profile identifies potential
379   biomarkers for human intervertebral disc degeneration. *Mol Med Rep* 16:8665-8672.
380   10.3892/mmr.2017.7741

381   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, and
382   Li YJ. 2018. Circular RNA GRB10 as a competitive endogenous RNA regulating nucleus pulposus cells
383   death in degenerative intervertebral disk. *Cell Death Dis* 9:319. 10.1038/s41419-017-0232-z

384   Guyon I, Weston J, Barnhill S, and Vapnik V. 2002. Gene Selection for Cancer Classification using Support Vector
385   Machines. *Machine Learning* 46:389-422.

386   Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, and Xu W. 2018. Applications of Support Vector Machine (SVM)
387   Learning in Cancer Genomics. *Cancer Genomics Proteomics* 15:41-51. 10.21873/cgp.20063

388   Ji ML, Jiang H, Zhang XJ, Shi PL, Li C, Wu H, Wu XT, Wang YT, Wang C, and Lu J. 2018. Preclinical development
389   of a microRNA-based therapy for intervertebral disc degeneration. *Nat Commun* 9:5051. 10.1038/s41467-
390   018-07360-1

391   Ji SC, Han N, Liu Y, Li G, Sun Z, and Li Z. 2015. Identification of genes associated with disc degeneration using
392   bioinformatics. *Biotech Histochem* 90:353-360. 10.3109/10520295.2015.1007481

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

395   Kalb S, Martirosyan NL, Kalani MY, Broc GG, and Theodore N. 2012. Genetics of the degenerated intervertebral
396   disc. *World Neurosurg* 77:491-501. 10.1016/j.wneu.2011.07.014

397   Kozomara A, and Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data.
398   *Nucleic Acids Res* 39:D152-157. 10.1093/nar/gkq1027

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

400   Lan PH, Liu ZH, Pei YJ, Wu ZG, Yu Y, Yang YF, Liu X, Che L, Ma CJ, Xie YK, Hu QJ, Wan ZY, and Wang HQ.
401   2016. Landscape of RNAs in human lumbar disc degeneration. *Oncotarget* 7:63166-63176.
402   10.18632/oncotarget.11334

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

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

409   Libbrecht MW, and Noble WS. 2015. Machine learning applications in genetics and genomics. *Nat Rev Genet* 16:321-
410   332. 10.1038/nrg3920

411   Liu H, Huang X, Liu X, Xiao S, Zhang Y, Xiang T, Shen X, Wang G, and Sheng B. 2014. miR-21 promotes human
412   nucleus pulposus cell proliferation through PTEN/AKT signaling. *Int J Mol Sci* 15:4007-4018.
413   10.3390/ijms15034007

414   Liu X, Che L, Xie YK, Hu QJ, Ma CJ, Pei YJ, Wu ZG, Liu ZH, Fan LY, and Wang HQ. 2015. Noncoding RNAs in
415   human intervertebral disc degeneration: An integrated microarray study. *Genom Data* 5:80-81.
416   10.1016/j.gdata.2015.05.027

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

419   Obermeyer Z, and Emanuel EJ. 2016. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine.
420   *New England Journal of Medicine* 375:1216-1219. 10.1056/NEJMp1606181

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

424   Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych
425   N, McMurry J, Marioni JC, Malone J, Megy K, Rustici G, Tang AY, Taubert J, Williams E, Mannion O,
426   Parkinson HE, and Brazma A. 2014. Expression Atlas update--a database of gene and transcript expression
427   from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res* 42:D926-932.
428   10.1093/nar/gkt1270

429   Quinlan AR, and Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*
430   26:841-842. 10.1093/bioinformatics/btq033

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

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

435   Silva D, Antunes J, Balamurugan K, Duncan G, Alho CS, and McCord B. 2016. Developmental validation studies of
436   epigenetic DNA methylation markers for the detection of blood, semen and saliva samples. *Forensic Sci Int*
437   *Genet* 23:55-63. 10.1016/j.fsigen.2016.01.017

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

444   Taki T, and Kibayashi K. 2015. Characterization of cellular and extracellular DNA in saliva. *Leg Med (Tokyo)* 17:471-
445   474. 10.1016/j.legalmed.2015.10.003

446   Tay Y, Rinn J, and Pandolfi PP. 2014. The multilayered complexity of ceRNA crosstalk and competition. *Nature*
447   505:344-352. 10.1038/nature12986

448   Tian C, Deng Y, Jin Y, Shi S, and Bi H. 2018. Long non-coding RNA RNCR3 promotes prostate cancer progression
449   through targeting miR-185-5p. *Am J Transl Res* 10:1562-1570.

450   Tsai TT, Lai PL, Liao JC, Fu TS, Niu CC, Chen LH, Lee MS, Chen WJ, Fang HC, Ho NY, and Pang JH. 2013.
451   Increased periostin gene expression in degenerative intervertebral disc cells. *Spine J* 13:289-298.

452   10.1016/j.spinee.2013.01.040

453   Vergroesen PP, Kingma I, Emanuel KS, Hoogendoorn RJ, Welting TJ, van Royen BJ, van Dieen JH, and Smit TH.
454   2015. Mechanics and biology in intervertebral disc degeneration: a vicious circle. *Osteoarthritis Cartilage*
455   23:1057-1070. 10.1016/j.joca.2015.03.028

456   Wan ZY, Song F, Sun Z, Chen YF, Zhang WL, Samartzis D, Ma CJ, Che L, Liu X, Ali MA, Wang HQ, and Luo ZJ.
457   2014. Aberrantly expressed long noncoding RNAs in human intervertebral disc degeneration: a microarray
458   related study. *Arthritis Res Ther* 16:465. 10.1186/s13075-014-0465-5

459   Wang H, He P, Pan H, Long J, Wang J, Li Z, Liu H, Jiang W, and Zheng Z. 2018a. Circular RNA circ-4099 is induced
460   by TNF-α and regulates ECM synthesis by blocking miR-616-5p inhibition of Sox9 in intervertebral disc
461   degeneration. *Exp Mol Med* 50:27. 10.1038/s12276-018-0056-7

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

465   Wang Z, Li Y, Wang Y, Wang X, Zhang J, and Tian J. 2018b. Association between GDF5 single nucleotide
466   polymorphism   rs143383   and   lumbar   disc   degeneration.   *Exp   Ther   Med*   16:1900-1904.
467   10.3892/etm.2018.6382

468   Williams FM, Bansal AT, van Meurs JB, Bell JT, Meulenbelt I, Suri P, Rivadeneira F, Sambrook PN, Hofman A,
469   Bierma-Zeinstra S, Menni C, Kloppenburg M, Slagboom PE, Hunter DJ, MacGregor AJ, Uitterlinden AG,
470   and Spector TD. 2013. Novel genetic variants associated with lumbar disc degeneration in northern
471   Europeans: a meta-analysis of 4600 subjects. *Ann Rheum Dis* 72:1141-1148. 10.1136/annrheumdis-2012-
472   201551

473   Xu YQ, Zhang ZH, Zheng YF, and Feng SQ. 2016. Dysregulated miR-133a Mediates Loss of Type II Collagen by
474   Directly Targeting Matrix Metalloproteinase 9 (MMP9) in Human Intervertebral Disc Degeneration. *Spine*
475   *(Phila Pa 1976)* 41:E717-724. 10.1097/brs.0000000000001375

476   Yuan B, Ji W, Fan B, Zhang B, Zhao Y, and Li J. 2018. Association analysis between thrombospondin-2 gene
477   polymorphisms and intervertebral disc degeneration in a Chinese Han population. *Medicine (Baltimore)*
478   97:e9586. 10.1097/md.0000000000009586

479   Zaina F, Tomkins-Lane C, Carragee E, and Negrini S. 2016. Surgical versus non-surgical treatment for lumbar spinal
480   stenosis. *Cochrane Database Syst Rev*:Cd010264. 10.1002/14651858.CD010264.pub2

481   Zhang F, Zhao X, Shen H, and Zhang C. 2016. Molecular mechanisms of cell death in intervertebral disc degeneration
482   (Review). *Int J Mol Med* 37:1439-1448. 10.3892/ijmm.2016.2573

483   Zhang Y, Zhang R, Ding X, and Ai K. 2018. EFNB2 acts as the target of miR-557 to facilitate cell proliferation,
484   migration and invasion in pancreatic ductal adenocarcinoma by bioinformatics analysis and verification. *Am*
485   *J Transl Res* 10:3514-3528.

486   Zhu Y, Qiao L, Zhou Y, Ma N, Wang C, and Zhou J. 2018. Long non-coding RNA FOXD2-AS1 contributes to
487   colorectal cancer proliferation through its interaction with microRNA-185-5p. *Cancer Sci* 109:2235-2242.
488   10.1111/cas.13632

489   **Legends of Figures**

490

491   **Figure 1: Differentially expressed IDD genes**

492   (A and B) Volcano plots represent the DEGs in the degenerative disc samples and control disc samples in the

493   GSE15227 and GSE34095 datasets, respectively. (C and D) Hierarchical cluster heatmaps of the GSE15227 and
494   GSE34095 datasets displaying the DEGs in the degenerative disc samples and the control disc samples. Blue
495   represents the downregulated and red represents the upregulated. (E) Venn diagram of DEGs in the GSE15227
496   and GSE34095 datasets. The common area represents the overlapping genes. DEG, differentially expressed
497   genes.
498
499   **Figure 2: Selection of optimal feature genes**
500   Ranking of the top 13 IDD-related genes using neural networks (A) and random forest (B). Extraction of the
501   optimum feature genes from the 13 IDD-related genes was carried out using recursive feature elimination (C).
502   Classification efficiency of the optimum feature genes in the model as evaluated using the ROC curve in the
503   GSE15227 (D), GSE34095 (E), and GSE56081 (F) datasets, respectively.
504
505
506   **Figure3: Identification of target miRNAs of the optimal feature genes**
507   (A)Volcano plots represent the DEMs of the degenerative disc samples and control disc samples in the
508   GSE116726 dataset. (B) Hierarchical cluster heatmaps of the GSE116726 dataset display the DEMs to compare
509   degenerative disc samples and control disc samples. Blue represents the degenerative samples and red represents
510   the control samples. (C) Venn diagram of DEMs in the GSE116726 and GSE63492 datasets. The common area
511   represents the overlapping DEMs. (D)Venn diagram of miRNAs in overlapping DEMs and target miRNAs of
512   optimum feature genes. The common area represents the overlapping miRNAs. DEM, differentially expressed
513   miRNA.
514
515   **Figure 4: ceRNA network construction**
516   ceRNA network of optimum IDD feature genes with DEMs, DELs, and DECs. The blue circles represent
517   optimum IDD feature genes, the red triangles represent DECs, the green diamonds represent DELs, and the
518   yellow arrows represent DEMs. ceRNA, competing endogenous RNA; DEC, differentially expressed circRNA;
519   DEL, differentially expressed lncRNA; and DEM, differentially expressed miRNA.
520
521   **Figure 5: Validation of optimal feature genes**
522   The expression levels of the four optimum feature genes in the GSE15227, GSE34095, and GSE56081 datasets,
523   respectively. (A, B and C) The expression level of BCSA4. (D, E and F) The expression level of CRNKL1. (G,
524   H and I) The expression level of SCRG1. (J, K and L) The expression level of WDR46. * represents P value <
525   0.05, *** represents P value < 0.001, **** represents P value < 0.0001, and NS represents not significant.
526
527   **Table 1. Basic information of the expression profiles**
528   **Table 2. The top three miRNAs and circRNAs related to optimal feature genes in the network**
529
530
531   **Supplemental Files**
532   **Figure S1 : Flow chart presentation of the study**
533   IDD, intervertebral disc degeneration; DEG, differential expression gene; RFE, recursive feature elimination;

534   SVM, support vector machine; DEM, differential expression miRNA; DEL, differential expression lncRNA;

535   DEC, differential expression circRNA; ceRNA, competing endogenous RNA.

536   **Figure S2: Differential expression analysis of the GSE56081, GSE63492, and GSE67566 datasets.**

537   Volcano plots represent the DEGs (A) and DELs (C) of the degenerative disc samples and control disc samples,

538   respectively, in the GSE56081 dataset. Hierarchical cluster heatmaps represent the DEGs (B) and DELs (D) of

539   the degenerative disc samples and control disc samples, respectively, in the GSE56081 dataset. Hierarchical

540   cluster heatmaps represent the DEMs in the GSE63492 dataset (E) and the DECs (F) in the GSE67566 dataset.

541   Blue represents the downregulated and red represents the upregulated.

542

543   **Table S1: All differentially expressed circRNAs in the GSE67566 dataset.**

544   **Table S2: All differentially expressed miRNAs in the GSE63492 dataset.**

545   **Table S3: The nodes of optimal feature genes, miRNAs, lncRNAs, and circRNAs in the ceRNA network.**

546

547   **Document S1: Differentially expressed mRNAs in the GSE56081 dataset.**

548   **Document S2: Differentially expressed lncRNAs in the GSE56081 dataset.**

549   **Document S3: The target miRNAs of DECs analyzed by miRanda tool.**

550   **Document S4: The target miRNAs of DELs analyzed by miRanda tool.**

551   **Document S5: Importance rankings of the 13 overlapping DEGs evaluated by random forest.**

552   **Document S6: Importance rankings of the 13 overlapping DEGs evaluated by neural network.**

Basic information of the expression profiles

1 **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

The top three miRNAs and circRNAs related to optimal feature genes in the network

1   **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

Figure 1

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.
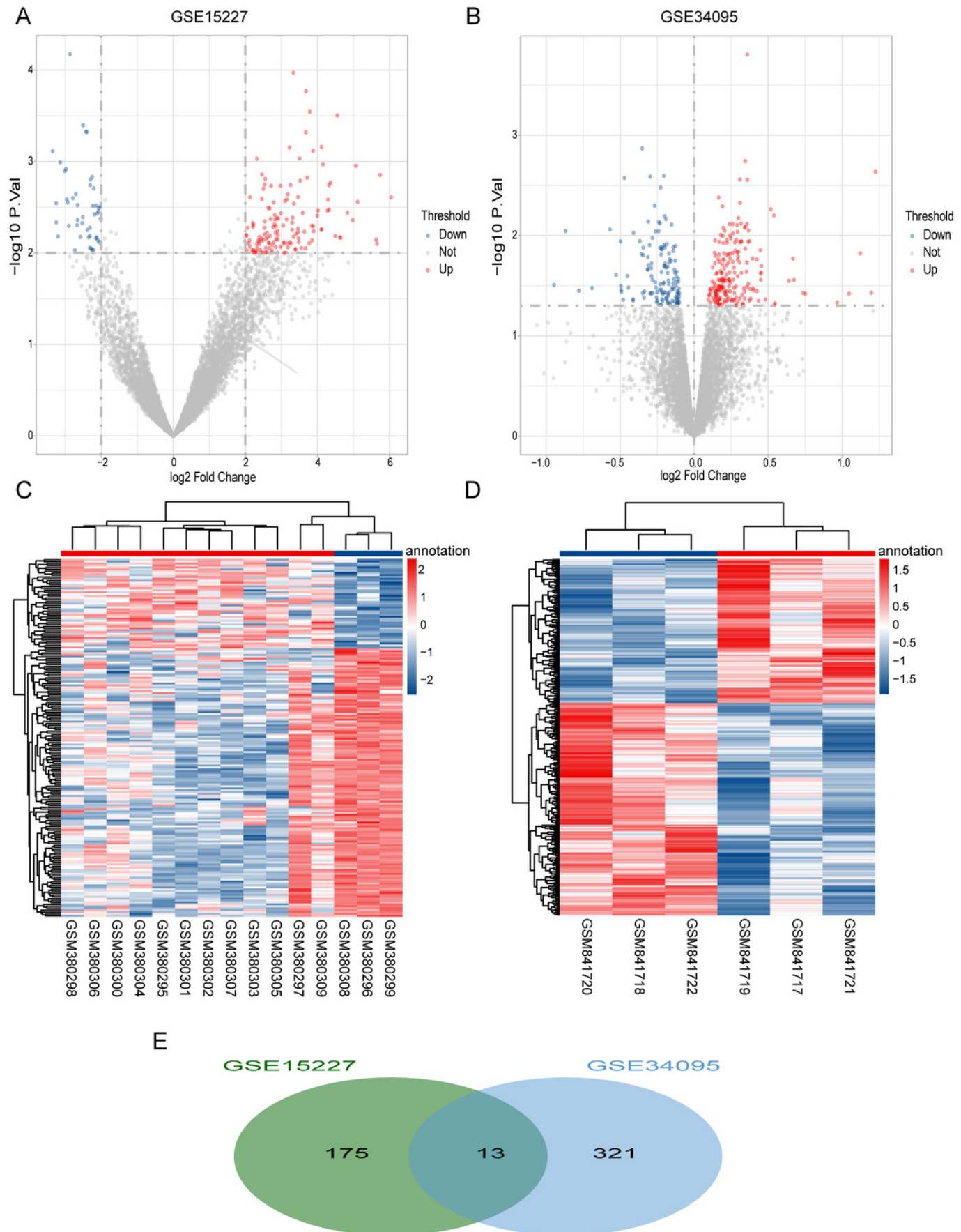
Figure 2

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.
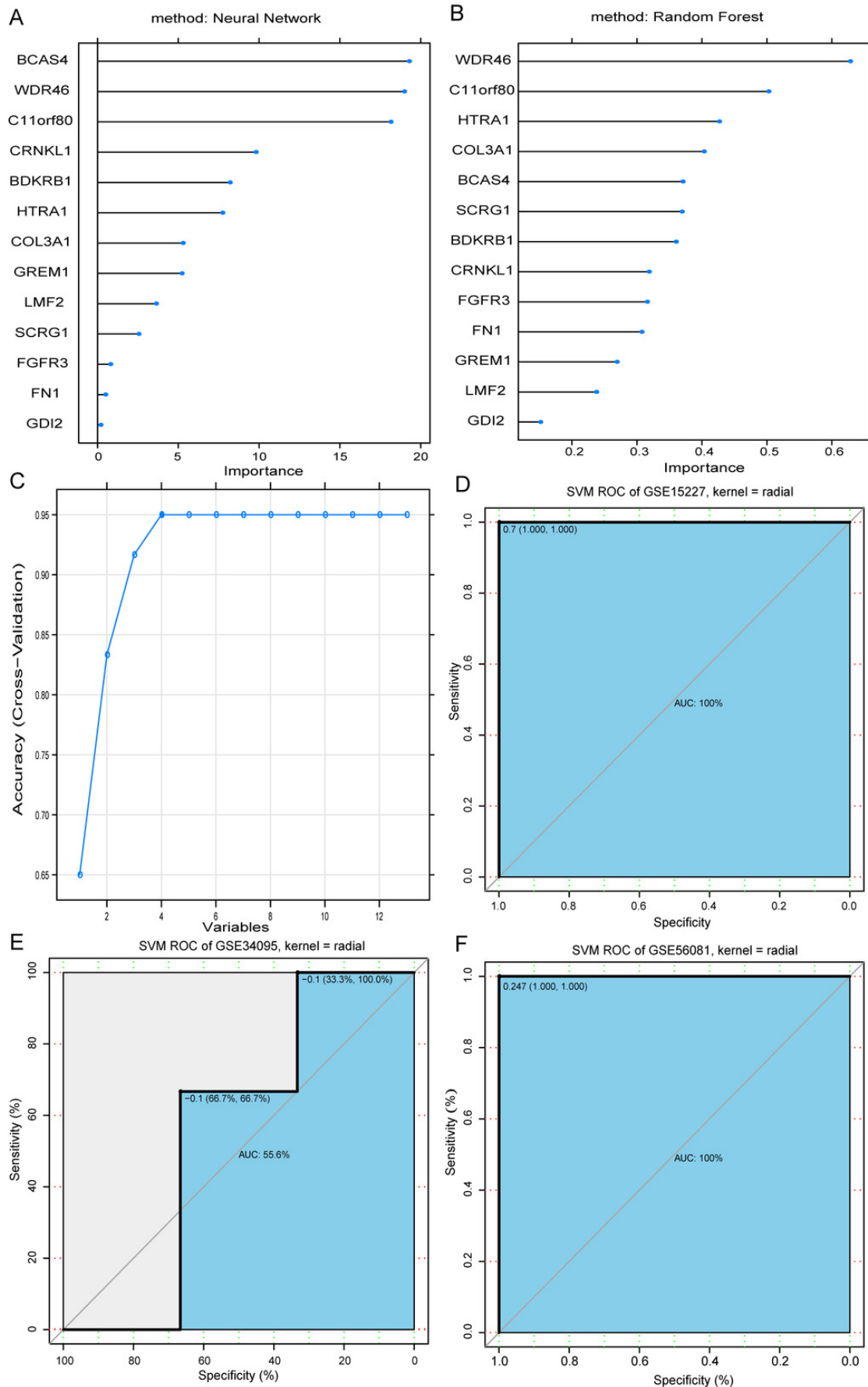
Figure 3

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.

A

B

C

D
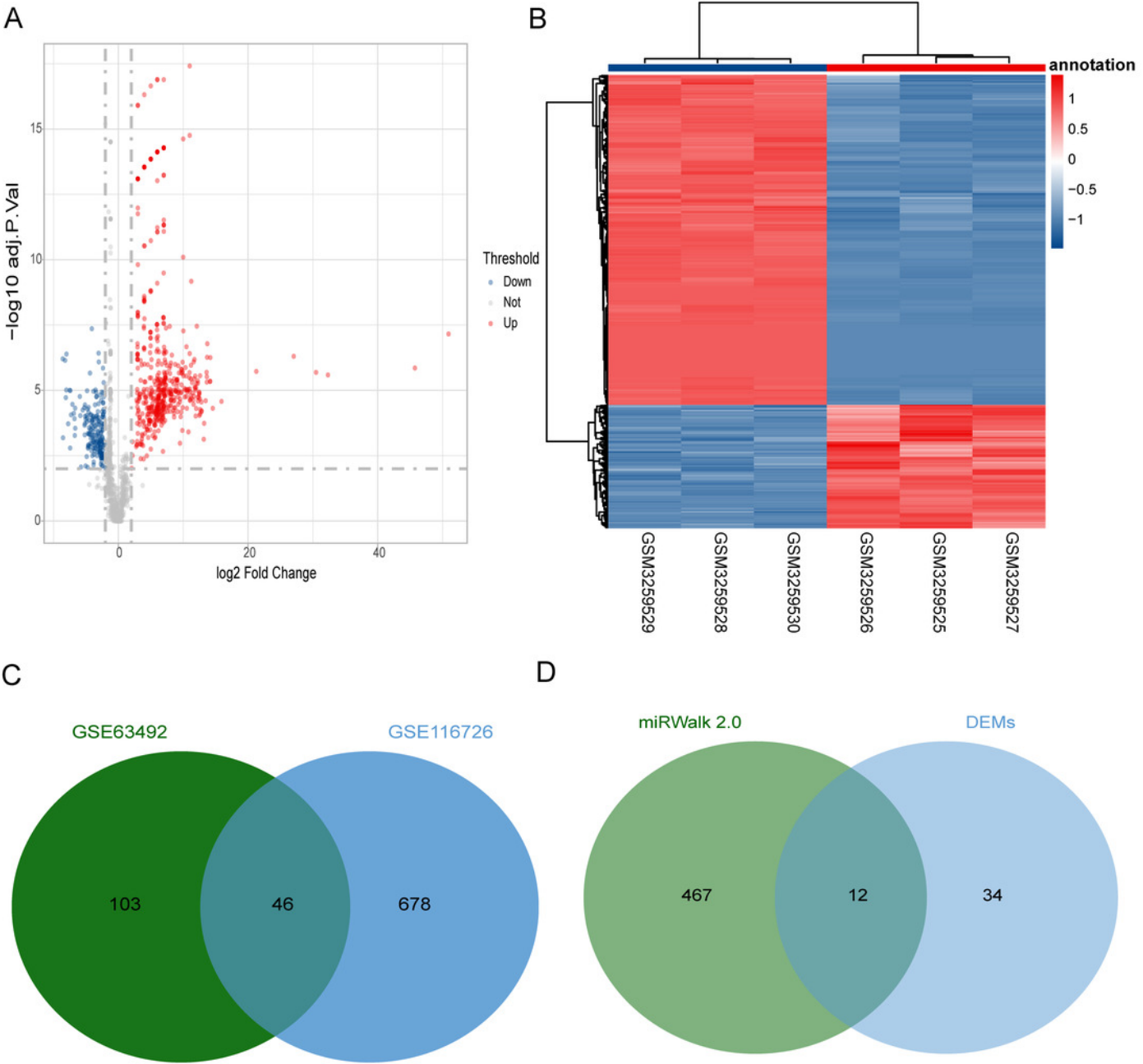
Figure 4

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.
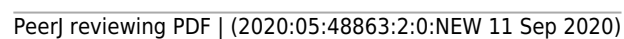
Figure 5

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.