Gynecology and Obstetrics Clinical Medicine 3 (2023) 154–162

Contents lists available at ScienceDirect

Gynecology and Obstetrics Clinical Medicine

journal homepage: www.keaipublishing.com/en/journals/gynecology-and-obstetrics-clinical-medicine

Research Article

Differential gene expression profile evaluation between uterine leiomyoma and leiomyosarcoma using a machine learning approach

Sonal Upadhyay [a], Ravi Bhushan [a], Anima Tripathi [b], Lavina Chaubey [c], Amita Diwakar [c], Pawan K. Dubey [a,*]

[a] Centre for Genetic Disorders, Institute of Science, Banaras Hindu University, Varanasi, 221005, Uttar Pradesh, India
[b] Department of Zoology, MMV, Banaras Hindu University, Varanasi, 221005, Uttar Pradesh, India
[c] Department of Obstetrics and Gynecology, Institute of Medical Science, Banaras Hindu University, Varanasi, 221005, Uttar Pradesh, India

A B S T R A C T

Objective: The objective of this study is to differentiate between uterine leiomyomas (ULM) and uterine leiomyosarcomas (ULMS) by conducting molecular differential analysis and identifying potential prognostic biomarkers for diagnosis.
Methods: The microarray datasets (GSEID: GSE64763 and GSE185543) were retrieved from the Gene Expression Omnibus database. Data preprocessing and differential gene expressions (DEGs) analysis were performed. The DEGs were further intersected to find the common DEGs in ULM and ULMS and further validation of selected DEGs were performed. Further, a machine learning classifier was also applied in the selection of biomarkers. Protein-protein interaction network based upon STRING v 10.5, was constructed. Additionally, Gene Ontology (GO) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analyses were also performed to dissect possible functions and pathways.
Results: A total of 50 significant DEGs for ULM while 321 DEGs for ULMS have been identified with their official gene symbol. Between ULM and ULMS, a total of 14 common DEGs were identified of which 8 were up-regulated while 6 were down-regulated. The DEGs of (GSE185543) were also analyzed and the significant genes were retrieved common in both datasets for further analysis. Using a machine learning approach, 10 feature genes were identified. Using the expression profiles of these genes, a sequential minimal optimization (SMO) prediction model was built on the training set, and it accurately and reliably classified features expression in ULM and ULMS in the independent test set. Furthermore, Co- Enrichment analysis was also performed.
Conclusion: The study identified several DEGs, including ZNF365, EPYC, COL11A1, SHOX2, MMP13, TNN, GPM6A, and GATA2, through cross-validation, machine learning classifier, and Co- Enrichment analysis. These candidate disease genes may provide valuable insight into the underlying mechanisms and could be used as potential diagnostic biomarkers for ULM and ULMS. However, further validation of these genes is necessary to better understand their roles in the pathogenesis of ULM and ULMS.

1. Introduction

The uterus, derived from paramesonephric organogenesis, is an essential supportive organ for prenatal growth and development in Eutherians. Histologically, the uterus has an inner mucosal layer (endometrium) and an outer muscular layer (myometrium).[1] The myometrium is composed of highly vascularized smooth muscle cells which help in inducing contraction during childbirth.[2] Uterine leiomyomas (ULM) or uterine fibroids are benign lesions of unspecified etiology that mainly arise from myometrium.[3] These lesions composed of smooth muscles including the extracellular matrices are commonly found in the pelvic area of those women bearing their reproductive ages.[4] Uterine leiomyosarcomas (ULMS) are rare malignant tumors known for hematogenous transmission leading to recurrence at both native and distant areas of uterine smooth muscles.[5]

According to the National Institutes of Health (NIH) India, approximately 25% of Indian women were found to be suffering from ULM.[6] Among these cases, nearly 25% exhibited different symptoms, including excessive bleeding, pelvic pain, pregnancy-related complications, and menstrual cramps.[7] The incidence of ULMS cases was found to be around

* Corresponding author. Centre for Genetic Disorders, Institute of Science, Banaras Hindu University, Varanasi, 221005, India.
E-mail address: pkdubey@bhu.ac.in (P.K. Dubey).

Publishing services by Elsevier on behalf of KeAi

https://doi.org/10.1016/j.gocm.2023.08.003
Received 23 March 2023; Received in revised form 27 July 2023; Accepted 4 August 2023
Available online 17 August 2023

25%~36% in women aged 50–60 years.[8]Various predisposing factors such as obesity, stress, smoking, age, and race (with a highly prevalent in African-American women), along with hormonal imbalances like estrogen, have been associated with the occurrence of leiomyoma. Additionally, some genetic factors have also been found to be associated with the diseases.[9] It was suggested that over-expression of HMGA2 gene has been observed in both ULM and ULMS due to chromosomal aberration, although this area of research requires further investigation.[10] It has been shown the genetic changes in both the TP53 and MED12 genes of ULMS and ULM suggests some common link between these two uterine pathological conditions.[11] However, due to similar patterns of pathological appearance, it may be difficult to differentiate between these two gynecological conditions. Furthermore, at the molecular level, due to unavailability of specific marker genes makes it more difficult to diagnose the disease. Therefore, there is a pressing need for more investigation to identify biomarkers associated with both ULM and ULMS cases to aid in better diagnosis and treatment.

To date, ULM and ULMS have shown resistant to various chemotherapeutic agents, and adjuvant therapies have not been very promising in treating these tumors.[12] To gain a better understanding the pathobiology of these lesions, a comparative study was conducted, comparing them with normal myometrium. However, only a few genes associated with ULM and ULMS have been identified to be responsible for the pathogenicity of the diseases. So, to search for more candidate genes and to disclose their mechanism at the molecular level inclusive in silico approach using different bioinformatics software was applied. The purpose of the present study is to screen potent biomarkers of ULM and ULMS diseases. The present study includes the gene expression datasets (ID: GSE64763 and GSE185543) analysis to identify differential gene expressions (DEGs). A comparison study was also performed by taking different datasets (GSE185543; RNA seq dataset) and differential expression analysis was performed and significant genes retrieved were used to find the common significant genes from both datasets which would play a vital role in both ULM and ULMS cases. Machine Learning Approach (ML) was also applied in the identification of important biomarkers. The construction of protein-protein interaction (PPI) networks was performed based on the combined score. Enrichment and functional analysis of DEGs were also performed. Finally, after applying different in silico selection criteria DEGs common between ULM and ULMS were identified as potent biomarkers responsible for ULM and ULMS pathogenesis.

2. Materials and methods

2.1. Retrieval of microarray gene expression profile

Using NCBI Gene expression Omnibus (GEO) datasets (https://www.ncbi.nlm.nih.gov/geo/)12 the raw gene expression profile (ID: GSE64763)[13,14] dataset was retrieved. The sample datasets were obtained for ULMS, ULM, and normal myometrium (NM) tissue specimens. In this dataset RNA was hybridized to HG-U133A_2 Affymetrix Human Genome U133A2.0Array at GPL571 platform. The series includes a total of 79 samples with 25 ULM, 25 ULMS, and 29 NM samples. Another dataset (GSE185543; RNA seq dataset) was also retrieved as containing a total of 9 samples (6 for ULMS and 3 for ULM cases). This dataset was analyzed separately to find hub genes that are common to both datasets. Different bioinformatics tools were used for the study of differential gene expressions in ULMS, NM, and ULM samples. Series matrix and platform files were retrieved in TXT file format.

2.2. Preprocessing dataset and screening DEGs

The preprocessing of retrieved raw datasets involved several steps. Firstly, the values of gene expression of probes related to specific genes were averaged. Subsequently, BiGGEsTs software[15] was utilized to identify up-regulated and down-regulated genes. GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r/) tool was used for converting the probe-level symbols into gene-level symbols. It is a web-based tool where the comparison between two or more datasets was carried out in a GEO series to identify DEGs. The adjusted *p*-value across experimental conditions using Benjamini and Hochberg false discovery rate was performed to find the statistically significant genes. The selected DEGs have <0.05 adjusted *p* values and threshold log fold change (FC) values > 0.1 for up-regulated and <-0.1 for down-regulated genes.

2.3. Machine learning approach

In the study, 70% of samples were selected randomly as the training set. The attribute feature selection method was used as a filter on the training set to select feature genes, which was implemented using Waikato Environment for Knowledge Analysis (WEKA) software.[16] Three well-known supervised classification methods (Naive Bayes, Sequential Minimal Optimization, and J48 tree) were utilized to create classification models using the training set, with default parameter settings in WEKA.[17] J48, sequential minimal optimization (SMO), and Naive Bayes are three popular classifiers used in differential gene expression analysis, which is a common task in bioinformatics and computational biology. These classifiers are used to identify genes that are differentially expressed between different conditions or groups (e.g., healthy vs. diseased, treated vs. untreated).

2.3.1. J48 (C4.5 decision tree algorithm)

J48 is an implementation of the C4.5 algorithm, which is a popular decision tree-based classification algorithm. In the context of gene expression analysis, J48 constructs a decision tree using the expression levels of genes as features to classify samples into different groups or conditions. The decision tree is built based on the information gained or other criteria to split the data into subsets, and the resulting tree is then used for classification. J48 is known for its simplicity, interpretability, and ability to handle both categorical and continuous data.

2.3.2. Sequential minimal optimization

Sequential Minimal Optimization (SMO) is a support vector machine (SVM) algorithm used for classification tasks. SVM is a powerful and widely used supervised learning method that aims to find an optimal hyperplane that separates data points of different classes with the largest margin. In differential gene expression analysis, SMO can be applied to classify samples based on the expression levels of genes, identifying which genes are most informative for distinguishing between different conditions.

2.3.3. Naive Bayes classifier

The Naive Bayes classifier is based on the application of Bayes' theorem with the assumption of independence among features (genes in this context). Despite the “naive” assumption, it has shown to be surprisingly effective in many text and classification tasks, including gene expression analysis. In this approach, the probability of a sample belonging to a certain condition is estimated based on the joint probabilities of the expression levels of individual genes in that condition. It's a computationally efficient method and can handle large datasets with many features (genes) quite well.

These models were then evaluated for their performance using tenfold cross-validation. The best-performing classifier, which had the highest accuracy, was selected from the training set and further evaluated on an independent test set (as a 30% split dataset randomly), which consisted of 7 samples from patients with ULM and 7 samples from ULMS patients. Various evaluation metrics, such as precision, recall, F-measure, Matthews correlation coefficient (MCC), AUC, auPRC, true positive rate, false positive rate, and Kappa statistic were used to assess the performance of the classifier on the test set.

In differential gene expression analysis, researchers often use these classifiers to build predictive models to determine which genes are most

important for distinguishing between different groups or conditions. These methods can provide valuable insights into the molecular mechanisms underlying diseases or the effects of specific treatments, helping researchers to better understand the complex processes involved in gene regulation and disease development.

2.4. Generation of principal component analysis and heatmap plots

Using the online tool ClustVis,[18] a heatmap and Principal component analysis (PCA) plot were generated for DEGs. This tool can support up to a maximum of 2 MB of file size thus it was impossible to generate a PCA plot for the total gene expression dataset.

2.5. PPI network and sub-network construction

To predict functional interactions among proteins, an online tool STRING version 10.5[19] (https://www.string-db.org/) was employed. This online tool provides combined scores between gene pairs for protein-protein interactions. For the present study, the DEGs which were identified were uploaded to this online database and a combined score >0.4 was set as the parameter for analysis. Then Cytoscape v 3.2.1(http://www.cytoscape.org),[20] an in-silico software package was used for different network and sub-networking creation. Degree and edge betweenness criteria were employed for constructing networks. Different nodes within the network, important in protein interactions were retrieved by ranking each node related to its score. Most scale-free networks were only considered, the hub genes with degrees ≥10 were screened out. For a simplified outlook of the full network isolated nodes and unconnected pairs with single edges were removed.

2.6. DEGs functional analysis

All the nodes in the network were enriched for their functions and pathways. DAVID (Database for Annotation, Visualization, and Integrated Discovery) software[21] (https://david.abcc.ncifcrf.gov) integrates an extensive set of functional annotations of large sets of gene records. The Gene Ontology (GO) functional annotation provides refers to the biological functions description for gene and protein functions through of biological functions using standard expression terms for gene and protein functions in various databases. This project was established by the Gene Ontology Consortium is founder of this project. GO enrichment analysis involves molecular function (MF), cellular component (CC), and biological process (BP) which by using DAVID v 6.8 and STRING v 10.5 tools were performed. Depending upon the hypergeometric distribution, DAVID uses a whole set of genes based on similar or closely associated functions.

2.7. Statistical analysis

For data organization and storage, Microsoft Excel 2013 was utilized. DEG was characterized as a gene with an fold change (FC) value range >1.2 or < 0.8. Visualization by color intensity of average FC of important gene ($p < 0.05$) was also carried out, red for up-regulation, and blue for down-regulation. The statistical analysis was performed by comparing groups using *t*-test. Multiple comparisons were performed using One-way-ANOVA. The significance level was set at $p < 0.05$. Moreover, to evaluate enrichment analysis ($\alpha = 0.05$) of these DEGs in each classification Fisher's fine test was applied. Also, by searching the KEGG database, Fisher's fine testing evaluated the enrichment significance of DEGs in each signaling pathway, to find the significant signaling pathway ($a = 0.05$).

3. Results

3.1. Selection of DEGs between ULM and ULMS

Microarray data of ULM, ULMS, and control specimens were normalized (Supplementary Fig. 1) using GEO2R. The dataset (GSE64763; RNA Seq) revealed a total of 50 significant DEGs for ULM while 321 DEGs for ULMS have been identified with their official gene symbol. In ULM, out of total DEGs, 29 were up-regulated and 21 were down-regulated while in ULMS, 154 were up-regulated and 167 were down-regulated (Fig. 1A and B) (Supplementary File 1). Among total DEGs, 8 up-regulated DEGs while 6 down-regulated DEGs were found to be common between ULM and ULMS (Fig. 1C and D). The *p*-value <0.05 and | log2 FC > 0.1 were used as selection criteria. Based on the average gene expression value DEGs were selected. Further, 2 DEGs in ULMS and 1 DEGs in ULM were found to be common in Online Mendelian Inheritance in Man (OMIM) and Gene Cards. Moreover, the RNA Seq dataset (GSE185543) revealed 2970 significant DEGs where 2113 genes were found upregulated and 855 were found downregulated (Supplementary Fig. 2 and Supplementary File 2).

3.2. Principal component and hierarchical clustering analysis of DEGs

Principal component analysis for ULM and ULMS reveals a scatter plot showing a total variance of 50.6% and 44.9% corresponding to the principal component 1 (x-axis) while 7.3% and 7.4% corresponding to principal component 2 (y-axis) respectively (Supplementary Figs. 3A and 3B). Heat-map shows a data matrix where coloring gives an overview of the numeric differences. Two separate heat maps for ULM and ULMS were constructed for respective differential expressed genes (Supplementary Figs. 4A and 4B).

3.3. Known disease genes and candidate genes to ULM and ULMS

Cross-validation of DEGs related to ULM and ULMS reveals 14 common DEGs of which 8 were up-regulated while 6 were down-regulated. However, out of these common DEGs, only 10 DEGs were found to have combined scores> 0.4 and hence included in the interaction network (Fig. 2A). Furthermore, we were interested to know the genes which have already been validated. For this, we compared our DEGs list with annotated gene list obtained from OMIM and Gene Cards (Fig. 2B) which leads to the identification of only 3 known disease genes (2 for ULMS and 1 for ULM; represented as a red and blue triangle in the network). Direct neighbors of these three known genes were considered candidate genes related to ULM and ULMS; this yields a total of 4 candidate genes which are shown in Fig. 2C. All the common as well as known candidate genes related to ULM and ULMS both were found to be significantly altered when plotted against average gene expression value. Out of 13 common as well as known candidate genes, 9 genes namely Kinesin Family Member 5C(KIF5C) with significant p-value 1.95E-10, Zinc Finger Protein 365(ZNF365) with significant *p*-value of 4.29E-08, Epiphycan precursor(EPYC) with significant *p*-value 3.39E-03, COLLAGEN, TYPE XI, ALPHA-1(COL11A1) with significant *p*-value of 5.96E-08, Short Stature Homeobox 2(SHOX2) with significant *p*-value of 9.59E-11, Matrix metalloproteinase 13(MMP13) with significant *p*-value of 1.29E-03, Tenascin N(TNN) with significant *p*-value of 5.16E-02, Ring Finger Protein(RNF128) with significant *p*-value of 1.21E-03, RAD51 Paralog B(RAD51B) with significant *p*-value of 1.35E-08 were up-regulated while 3 genes namely GATA Binding Protein 2 (GATA2) with significant *p*-value of 7.06E-13, Glycoprotein M6A (GPM6A) with significant *p*-value of 1.35E-08, Estrogen Receptor 1 (ESR1) with significant *p*-value of 9.57E-02 and PDGFRA(platelet-derived growth factor receptor alpha) with significant *p-value* of 3.85E-01 were down-regulated. However, from RNA seq dataset analysis the significant genes retrieved were matched with the above-mentioned significant genes to find some common genes from both datasets. Few significant genes were found to be common in both datasets namely COL11A1, and SHOX2 as up-regulated genes, and GATA2 as down-regulated genes.

3.4. Performance evaluation of ULM and ULMS classifier

Using the Training set, the number of features (14) selected after

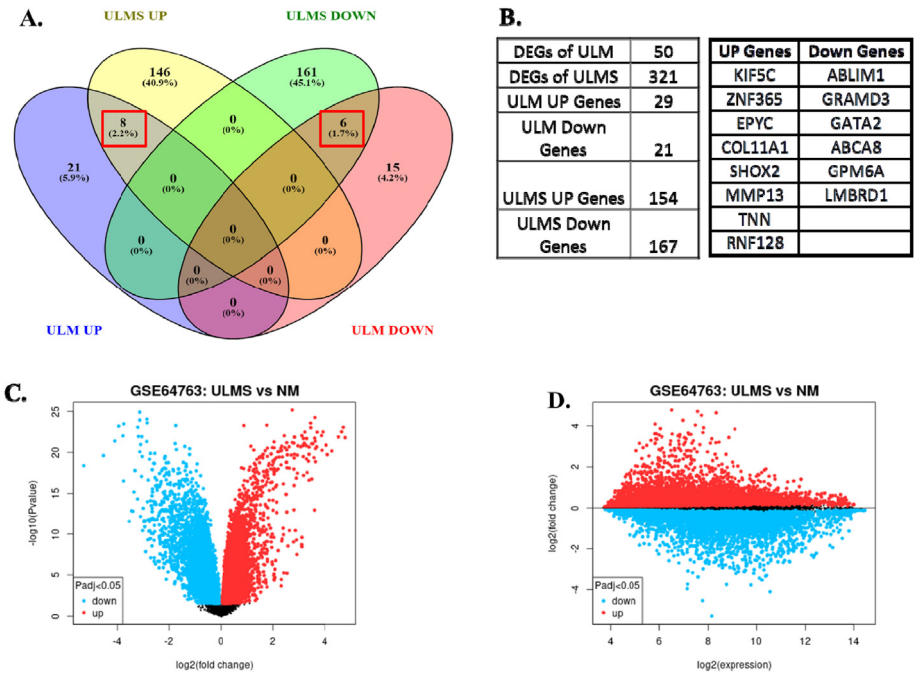

Fig. 1. Venn diagram showing common DEGs in ULM and ULMS. Total DEGs with up-regulated and down-regulated genes. The red rectangle highlights the UP and down-regulated genes common in both cases. Venny tool v 2.1.0 was used to draw the Venn diagram.

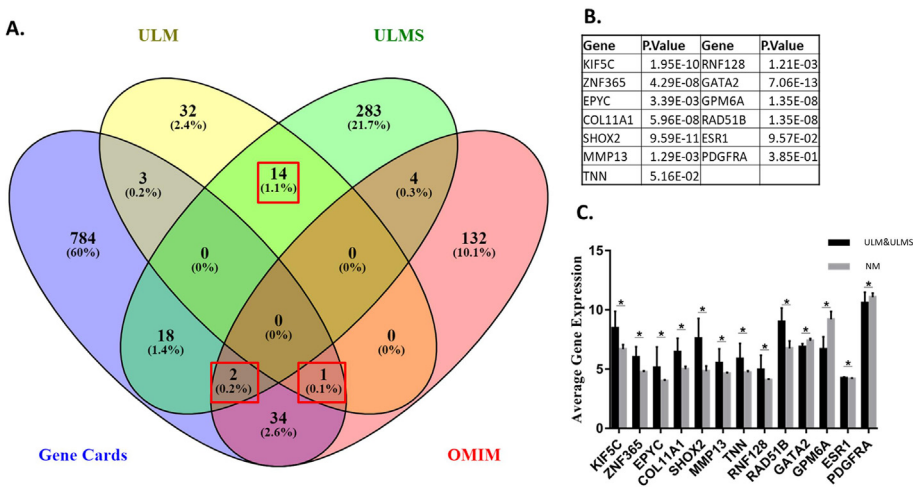

Fig. 2. Venn diagram showing the common as well as known uterine fibroids. The red rectangle highlights the common genes between ULM and ULMS as well as related candidate genes. Venny tool v 2.1.0 was used to draw the Venn diagram.
Note: ULM: uterine leiomyoma; ULMS: uterine leiomyosarcoma; NM: normal myometrium.

cross-validation including the expression values of these selected features were applied to the filter using attribute selection containing (Info gain attribute eval and ranked as options) which selected 10 features (GPMA6, GRAMD3, COL11A1, ZNF365, SHOX2, LMBRD1, EPYC, ABCA8, MMP13, and TNN) among 14 features which were applied to three supervised ML algorithms to generate classifiers — treeJ48, NB and SMO, depending upon the training set. We first executed tenfold cross-validation to classify ULM and ULMS samples. All classifiers were conducted well with a precision of 69.4% for tree J48, 77.7% for NB, and 80.55% for SMO (Table 1). After conducting a thorough assessment of various metrics, it was found that the SMO classifier outperformed the others. The SMO classifier was built using the entire training set and was then tested on an independent set where it achieved 78.57% accuracy, and its rate performance was also evaluated using different criteria, such as precision, recall, F-measure, MCC, auPRC, true positive rate, false positive rate, and Kappa statistic. The precision of the SMO classifier was 79%, and its F-measure was 0.792. The MCC value was 0.57, and the auPRC area was 0.72. The true positive rate was 0.78, while the false positive rate was 0.21, and the Kappa statistic was 0.57. These results confirmed that the SMO classifier was highly accurate and that the 10-gene feature is an important biomarker for both ULM and ULMS.

3.5. *The protein-protein interaction network*

For the protein-protein interaction network, all DEGs with a combined score >0.4 (283 gene pairs out of 371 DEGs) were used which yielded one main network having 266 nodes and 883 edges (Supplementary Fig. 5) while a separate network of DEGs with combined score >0.9 was extracted separately. A total of 110 DEGs with a combined score >0.9 were included in the network (red node for upregulated and

Table 1
All three classifiers' accuracy and rate performance on the training set.

Classifier	Precision	F-Measure	MCC	auPRC	TP Rate	FP Rate	Kappa statistics
NB (Naive Bayes)	0.78	0.77	0.559	0.76	0.77	0.22	0.55
J48	0.7	0.69	0.39	0.64	0.69	0.30	0.38
SMO	0.81	0.80	0.62	0.75	0.80	0.19	0.61

NB: Naive Bayes, SMO: sequential minimal optimization, MCC: Matthews correlation coefficient, auPRC: area under the precision-recall curve, TP: true positive, FP: false positive.

blue node for down-regulated) (Fig. 3).

3.6. Functional enrichment analysis

Gene ontology enrichment analysis for DEGs of ULM and ULMS was performed and significantly enriched functions, processes, and cellular components (p-value <0.05) were listed in Supplementary File 3. Major significant (p-value <0.05) processes enriched for ULM were regulation of cell death, regulation of apoptosis, cell-cell adhesion, and cell morphogenesis (Fig. 4A) while extracellular matrix organization, response to steroid hormone stimulus, regulation of cell proliferation, blood-vessel morphogenesis, cell motility, and cell cycle phase were significant processes (p-value <0.05) for ULMS (Fig. 4B).

Co-enrichment analysis of common and known candidate genes related both to ULM and ULMS led to the identification of several important processes. A separate biological processes network was created for those genes. UP-regulated genes like KIF5C, ZNF365, EPYC, COL11A1, SHOX2, MMP13, TNN, and RNF128, were found to be involved in the regulation of cell proliferation, cell adhesion, and response to estrogen stimulus. Major processes regulated by down-regulated genes (GATA2, GPM6A) were found their involvement in the regulation of transcription, cell morphogenesis, cell differentiation, cell projection, and extracellular matrix organization (Fig. 5).

KEGG pathway enrichment analysis for DEGs of ULM and ULMS revealed a total of 8 significantly enriched pathways (p-value<0.05). Small cell lung cancer, cell cycle, vascular smooth muscle contraction, focal adhesion, cell adhesion molecules (CAMs), and extracellular matrix (ECM) receptor interaction were pathways identified for ULM while ECM receptor interaction and focal adhesion are significant pathways associated with ULMS (Fig. 6).

Moreover, after high throughput sequencing dataset analysis (GSE) it was found that COL11A1, SHOX2, and GATA2 were found to be significant genes that were common in both analyzed DEGs retrieved from two different datasets. Also, these genes were found to be common in ULM and ULMS cases.

Therefore, based on cross-validation, ML classifier analysis, Co-enrichment analysis, and RNA sequencing dataset analysis, the following DEGs: ZNF365, EPYC, COL11A1, SHOX2, MMP13, TNN, GATA2, and GPM6A, were identified as potential candidate disease genes associated with both ULM and ULMS.

4. Discussion

Uterine Leiomyoma, or uterine fibroid (ULM), is a benign lesion that

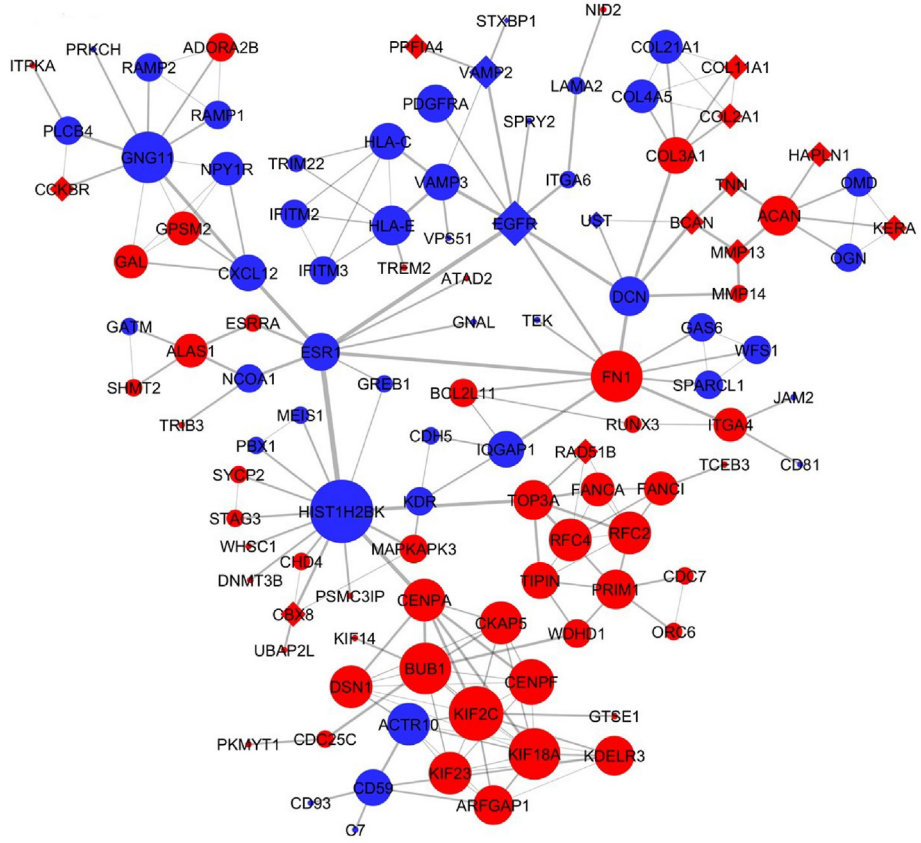

Fig. 3. Protein-protein interaction (PPI) of differentially expressed genes. Red circle and red diamond up-regulated genes, and blue circle and blue diamond down-regulated genes. The correlation between genes' Thickness of lines (edges) is proportional to the combined score.

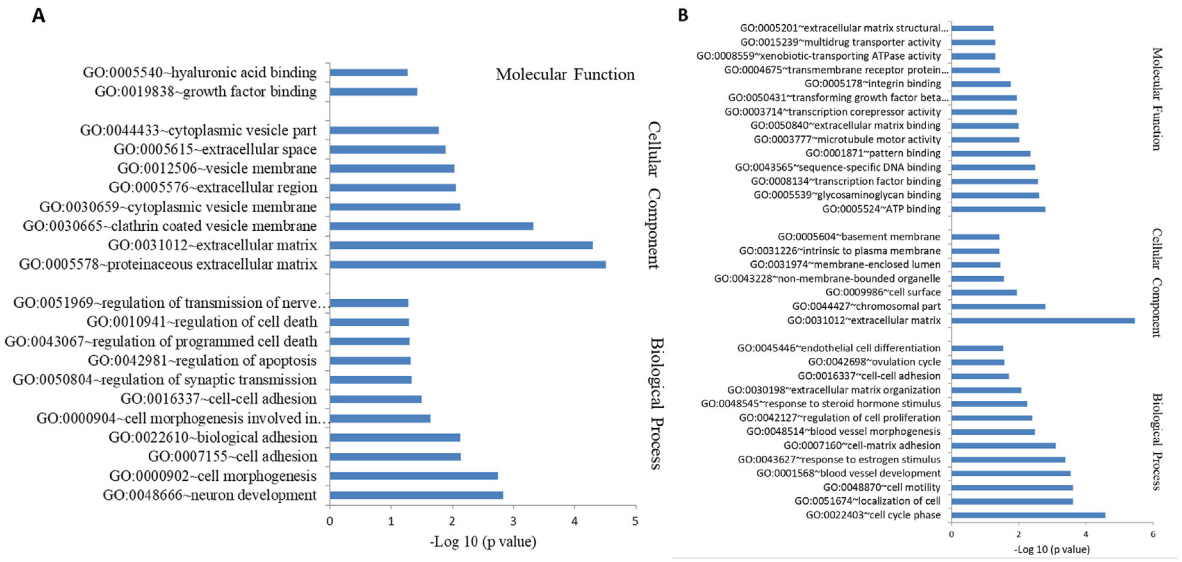

Fig. 4. Gene Ontology analysis for ULM (A) and ULMS(B) related DEGs in PPI network in uterine fibroids. Note: Bar graph showing significant processes, functions, and cellular components enriched in ULM for up-regulated genes. DAVID v 6.7 was used for annotation.

arises commonly in the muscular areas of the uterine wall.[22] Uterine leiomyosarcoma (ULMS) is a smooth muscle malignancy that arises in the smooth muscle areas of the uterus.[23]

The occurrence of ULMS is relatively rare, with approximately one to five women among every 1000 women diagnosed with fibroids found to have ULMS as well. The prevalence of both conditions is on the rise, and effective treatments for these diseases have not yet been discovered.[24] To address this medical challenge and explore more efficient treatment methods, in the present study we have used bioinformatic tools to find out molecular mechanism of ULM and ULMS pathogenesis. In this present in silico analysis, the highly efficient screening of gene expressions dataset was performed which revealed a total of 371 DEGs (50 ULM and 321 ULMS genes). Based on the GO cluster, the main biological processes of DEGs involve cell adhesion, cell motility, cell differentiation, and localization of cells in ULMS while cell adhesion, apoptosis, neuronal development, and cell morphogenesis in ULM. Major DEGs that formed

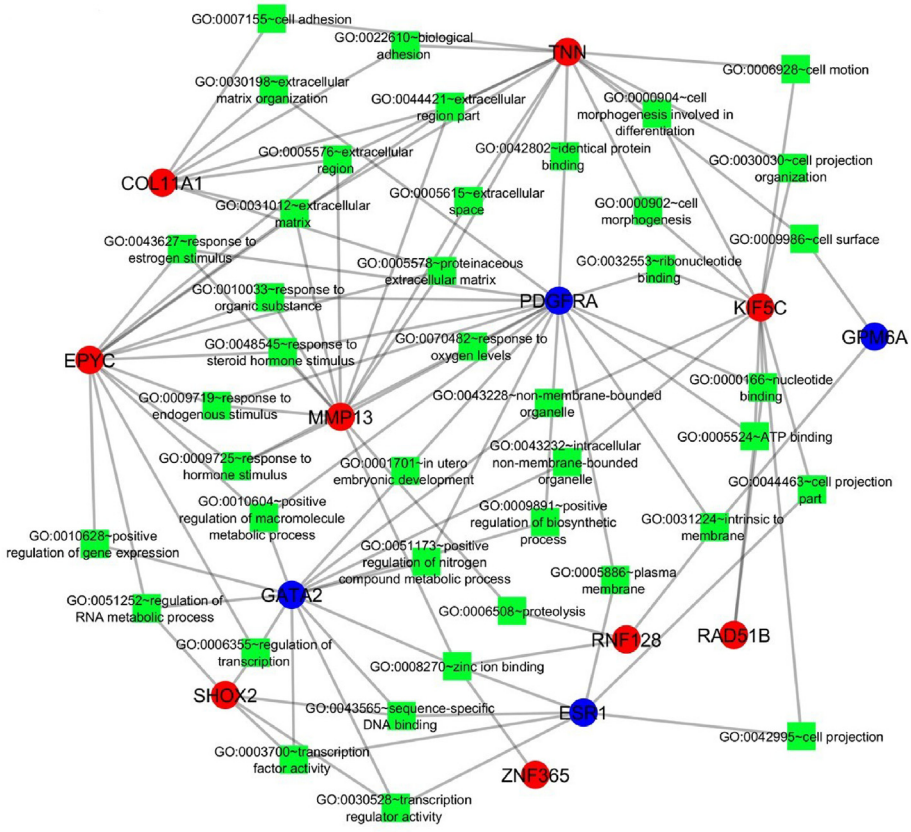

Fig. 5. Functional analysis of common as well as known UF genes. The functional analysis uncovers many significant processes being regulated by the candidate and known UF-related genes. Note: Red circle: up-regulated genes; Blue circle: up-regulated genes; Light green rectangle: biological processes.

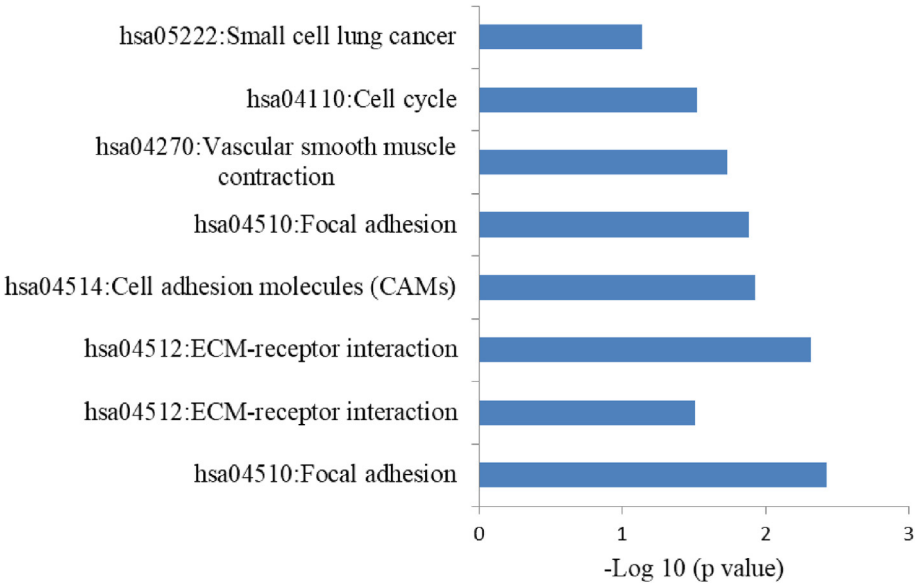

Fig. 6. KEGG pathway analysis for differentially expressed genes in uterine fibroids. Pathway enrichment for DEGs leads to the identification of 6 significant pathways for ULMS- while 2 significant pathways for ULM-related DEGs. DAVID version 6.7 was used for annotation.

the hub nodes were eight up-regulated genes (KIF5C, ZNF365, EPYC, COL11A1, SHOX2, MMP13, TNN, and RNF128) and six down-regulated genes (ABLIM1, GRAMD3, GATA3, ABCA8, GPM6A, LMBRD1).

After the cross-validating the data between the ULM and ULMS, we identified 14 genes that were common between the two groups. To determine the most important and significant selected features, the machine learning attribute selection method was applied to the training data which led to the 10 final features, wherein 6 are upregulated (COL11A1, ZNF365, SHOX2, EPYC, MMP13, and TNN) and 4 are downregulated (GPMA6, GRAMD3, LMBRD1, ABCA8). The SMO classifier based on the gene expression values of these genes (as features) present in the training data set can classify ULM and ULMS samples with a precision of 78.75% with a precision of 79% in the independent test set.

Some of these genes selected as feature genes, as well as co-enrichment genes, have been reported to be connected with both ULM and ULMS cases and their related pathogenesis. ZNF365 gene was found to be involved in the maintenance of a stable genome, repairing damaged DNA. Moreover, ZNF365 also promotes recovery of stalled replication fork to provide genomic stability which was detected in both hereditary and sporadic cancer types. Variations in the ZNF365 gene may increase the risk of having breast cancer by affecting the dense tissue proportion in the breast.[25] According to Zhang Y et al.,[26] ZNF365 loss can lead to a delay in mitosis progression and this also results in cell-cycle exit due to stress in the replication process which leads to an increase in aneuploidy, centrosome reduplication, and disruption of cytokinesis process. However, the precise mechanism of ZNF365 in the context of ULM and ULMS has not yet been identified. Nonetheless, given its speculated close relationship with to DNA repairing and genome stability, ZNF365 emerges as a potential target for the treatment of ULM and ULMS.

The EPYC genes encodes for proteoglycan that plays a crucial role in regulating fibrillogenesis. EPYC gene was found to be involved in breast, uterine, and colorectal cancer. Januchowski R et al.[27] has also reported the upregulation of found EPYC gene to be upregulated in both two cell lines (A2780DR1, A2780DR2) that were DOX resistant. However, to date no studies have been reported about involvement of TNN gene encodes proteins in cell migration. In tumors, it stimulates the angiogenesis of endothelial cells. It was also found to be one of the biomarkers for breast cancer.[28,29] According to Peterson LE et al.,[30] the TNN gene is involved in cell-matrix adhesion in lung adenocarcinoma. Liu B et al.[31] investigated that cancer genes like TNN were found to be involved in extracellular matrix interactions like pathways. Although TNN gene was not identified in ULM and ULMS-like cases, its role in cell-matrix adhesion suggests it may serve as a promising target for both conditions. MMP13 encodes proteins produced from stromal fibroblasts that are involved in the degradation of different ECM components and induces angiogenesis by increasing protein levels of VEGF and VEGFR2.[32] Halder SK et al.[33] Reported high expression of MMP13 in ULM pathogenesis, but not in ULMS. In addition, Courtoy GE et al.[34] also showed that MMP13 encoded proteins were involved in apoptosis, and cell proliferation in myoma. Given its role in ULM and potential impact on cell behavior, MMP13 presents itself as a potential therapeutic target for both ULM and ULMS.

GPMA6 gene encodes a protein involved in neuronal differentiation and development. These encoded proteins help in neuronal stem cell migration.[35] GPMA6 gene was found to be a novel target gene involved in proliferation, promoting tumor survival and development in thyroid carcinomas.[36] These features suggest that GPMA6 could be a novel candidate with potential relevance for both ULM and ULMS. Its role in neuronal development and tumor development makes it an intriguing gene to investigate further for its potential impact on these uterine conditions.

According to Liu X et al.,[37] the COL11A1 gene was identified as a marker for uterine fibroid via gene expression analysis, with its encoded proteins involved in focal adhesion and extracellular matrix receptor interactions. These characteristics suggest that COL11A1 may serve as a biomarker for leiomyoma cases. However, its role in ULM has not been reported. Further investigation is warranted to explore its potential diagnostic and prognostic relevance in ULMS.

The SHOX2 gene has been implicated in the development and progression of uterine leiomyoma and leiomyosarcoma. A study by Teixeira MR et al.[38] demonstrated that SHOX2 is overexpressed in uterine leiomyosarcoma, and this upregulation is associated with poor patient prognosis. Additionally, a study by Ak A et al.[39] found that SHOX2 is upregulated in uterine leiomyoma and may play a role in its pathogenesis. The precise mechanism by which SHOX2 contributes to the development and progression of these tumors is still not fully understood, but these findings suggest that SHOX2 may serve as a potential diagnostic and therapeutic target in the management of uterine leiomyoma and leiomyosarcoma. Further research is needed to clarify the precise role of SHOX2 in these diseases and to identify potential targeted therapies.

In both uterine leiomyomas and leiomyosarcomas, we observed changes in gene expression related to ECM, cell cycle, and cell-cell contact inhibition mechanisms, which are responsible for uncontrolled cell

multiplication and metastatic behavior in cancers. At the histopathological level, the loss of cell-cell contact inhibition contributes to unrestricted cellular proliferation, leading to tumor formation and promoting metastasis.[40,41]

Leiomyosarcomas exhibited a significant enrichment of pathways related to the cell cycle, cell-cell contact inhibition, and DNA replication.[42] Our study's findings align with earlier research, which also reported commonly enriched of EC pathways in both ULMS and ULM.[43,44] This study provides further validation of previous research and sheds light on the underlying molecular mechanisms involved in the transformation of ULM to ULMS, offering potential insights into identifying diagnostic markers and therapeutic targets.

In this study, a comprehensive analysis was conducted using microarray datasets of NM, ULM, and ULMS tissues, alongside an RNA sequencing dataset containing ULMS and ULM samples. This integrated approach allowed for the identification of common genes between the datasets, providing a more comprehensive understanding of the synergistic effects of differential gene expression on various biological processes and pathways at the molecular level.

Yet, the chosen genes are ensured to be significant features for ULM and ULMS both by the integrated analysis of multi-omics data and the machine learning technology. To better understand their roles in the pathogenesis of ULM and ULMS, more research (in vitro and in vivo) is required.

In summary, the present study utilized a multi-omics analysis to identify a group of feature genes that are significantly dysregulated and associated with ULM and ULMS both. Using the expression profiles of these 10 feature genes, a prediction model based on the SMO classifier was built on the training set. The model accurately and reliably classified the gene expression patterns in ULM and ULMS in an independent test set, demonstrating the robustness and potential clinical utility of these genes as diagnostic markers. Co-Enrichment analysis was also performed to provide further insights into the biological processes and pathways associated with the identified feature genes, shedding light on their potential functional roles in ULM and ULMS development.

ZNF365, EPYC, COL11A1, SHOX2, MMP13, TNN, GATA2, and GPM6A were identified as candidate disease genes for both ULM and ULMS through cross-validation, machine learning classification, and Co-Enrichment analysis. These genes are potentially valuable targets for future research and therapeutic development in managing ULM and ULMS.

Furthermore, the alterations in gene expression associated with extracellular matrix (ECM), cell cycle regulation, and cell-cell contact inhibition mechanisms in both uterine leiomyomas and leiomyosarcomas provide important insights into the underlying mechanisms contributing to uncontrolled cell proliferation and the potential for metastatic behavior frequently observed in cancer.

In conclusion, this study offers valuable insights into the molecular mechanisms of ULM and ULMS through an integrated analysis of multi-omics data and machine learning. The identified feature genes hold promising diagnostic and therapeutic potential for future treatments of ULM and ULMS. These findings advance our understanding of these uterine conditions and contribute to the development of personalized and targeted approaches for improved patient outcomes.

Authors contribution

SU is responsible for methodology, investigation, writing - review & editing. RB is responsible for conceptualization, data curation, writing-review & editing. LC, AD, AT are responsible for visualization, investigation, validation. PKD is responsible for review & editing.

Disclosure statement

No potential conflict of interest was reported by the authors.

Ethics approval

Each patient signed a written informed consent form before surgery. This study was approved by the Research Ethics Committee of the Department.

Funding

This study was supported by the Science and Engineering Research Board (SERB), Ministry of Science and Technology, Government of India (Grant No. CRG/2019/002237) and IoE-Faculty Incentive Grant of Banaras Hindu University.

Acknowledgement

The authors are highly thankful to all study participants and acknowledge their valuable help in this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gocm.2023.08.003.

References

1. Spencer TE, Hayashi K, Hu J, et al. Comparative developmental biology of the mammalian uterus. *Curr Top Dev Biol.* 2005;68:85–122. https://doi.org/10.1016/S0070-2153(05)68004-0.
2. Commandeur AE, Styer AK, Teixeira JM. Epidemiological and genetic clues for molecular mechanisms involved in uterine leiomyoma development and growth. *Hum Reprod.* 2015;21(15):593–615. https://doi.org/10.1093/humupd/dmv030.
3. Bowden W, Skorupski J, Kovanci E, et al. Detection of novel copy number variants in uterine leiomyomas using high-resolution SNP arrays. *Mol Hum Reprod.* 2009;15(9):563–568. https://doi.org/10.1093/molehr/gap050.
4. Laughlin S, Schroeder J, Baird D. New directions in the epidemiology of uterine fibroids. *Semin Reprod Med.* 2010;28(10):204–217. https://doi.org/10.1055/s-0030-1251477.
5. Leiomyosarcoma Singh Z. A rare soft tissue cancer arising from multiple organs. *J Cancer Res and Pract.* 2018;5:1–8. https://doi.org/10.1016/j.jcrpr.2017.10.002.
6. Marsh EE, Al-Hendy A, Kappus D, et al. Burden, prevalence, and treatment of uterine fibroids: a Survey of U.S. Women. *J Womens Health (Larchmt).* 2018;27:1359–1367. https://doi.org/10.1089/jwh.2018.7076.
7. Ciavattini A, Di Giuseppe J, Stortoni P, et al. Uterine fibroids: pathogenesis and interactions with endometrium and endomyometrial junction. *Obstet Gynecol Int.* 2013;2013:1–11. https://doi.org/10.1155/2013/173184.
8. Levy AD, Manning MA, Al-Refaie WB, et al. Soft-tissue sarcomas of the abdomen and pelvis: radiologic-pathologic features, Part 1—common Sarcomas: from the radiologic pathology archives. *Radiographics.* 2017;37:462–483. https://doi.org/10.1148/rg.2017160157.
9. Rafnar T, Gunnarsson B, Stefansson OA, et al. Variants associating with uterine leiomyoma highlight genetic background shared by various cancers and hormone-relatedtraits. *Nat Commun.* 2018;9:3636. https://doi.org/10.1038/s41467-018-05428-6.
10. Mehine M, Kaasinen E, Heinonen HR, et al. Integrated data analysis reveals uterine leiomyoma subtypes with distinct driver pathways and biomarkers. *Proc Natl Acad Sci USA.* 2016;113:1315–1320. https://doi.org/10.1073/pnas.1518752113.
11. Sparić R, Andjić M, Babović I, et al. Molecular insights in uterine leiomyosarcoma: a systematic review. *Int J Mol Sci.* 2022;27;23(17):9728. https://doi.org/10.3390/ijms23179728.
12. Hensley ML, Maki R, Venkatraman E, et al. Gemcitabine and docetaxel in patients with unresectable leiomyosarcoma: results of a phase II trial. *J Clin Oncol.* 2002;20:2824–2831. https://doi.org/10.1200/JCO.2002.11.050.
13. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.* 2007;35:D760–D765. https://doi.org/10.1093/nar/gkl887 480.
14. Barlin JN, Zhou QC, Leitao MM, et al. Molecular subtypes of uterine leiomyosarcoma and correlation with clinical outcome. *Neoplasia.* 2015;17:183–189. https://doi.org/10.1016/j.neo.2014.12.007.
15. Bhushan R, Rani A, Ali A, et al. Bioinformatics enrichment analysis of genes and pathways related to maternal type 1 diabetes associated with adverse fetal outcomes. *J Diabet Complicat.* 2020;34(5):107556. https://doi.org/10.1016/j.jdiacomp.2020.107556.
16. Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. *SIGKDD Explor Newsl.* 2009;11(1):10–18.
17. Naorem LD, Muthaiyan M, Venkatesan A. Integrated network analysis and machine learning approach for the identification of key genes of triple-negative breast cancer. *J Cell Biochem.* 2019;120(4):6154–6167. https://doi.org/10.1002/jcb.27903.

18. Metsalu T, Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Res*. 2015;43: W566–W570. https://doi.org/10.1093/nar/gkv468.
19. Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*. 2012;41:D808–D815. https://doi.org/10.1093/nar/gks1094.
20. Kohl M, Wiese S, Warscheid B. Cytoscape: software for visualization and analysis of biological networks. *Methods Mol Biol*. 2011;696:291–303. https://doi.org/10.1007/978-1-60761-987-1_18.
21. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4:44–57. https://doi.org/10.1038/nprot.2008.
22. Mukhopadhaya N, De Silva C, Manyonda IT. Conventional myomectomy. *Best Pract Res Clin Obstet Gynaecol*. 2008;22(4):677–705. https://doi.org/10.1016/j.bpobgyn.2008.01.012.
23. Kaur P, Kaur A, Singla A, Kaur K. Uterine leiomyosarcoma: a case report. *J Mid life Health*. 2014;5:200. https://doi.org/10.4103/0976-7800.145175.
24. Leibsohn S, d'Ablaing G, Mishell DR, et al. Leiomyosarcoma in a series of hysterectomies performed for presumed uterine leiomyomas. *Am J Obstet Gynecol*. 1990;162(4):968–976. https://doi.org/10.1016/0002-9378(90)91298-Q.
25. Lindström S, Vachon CM, Li J, et al. Common variants in ZNF365 are associated with both mammographic density and breast cancer risk. *Nat Genet*. 2011;43(3):185–187. https://doi.org/10.1038/ng.760.
26. Zhang Y, Shin SJ, Liu D, et al. ZNF365 promotes stability of fragile sites and telomeres. *Cancer Discov*. 2013;3(7):798–811. https://doi.org/10.1158/2159-8290.CD-12-0536.
27. Januchowski R, Zawierucha P, Ruciński M, et al. Extracellular matrix proteins expression profiling in chemoresistant variants of the A2780 ovarian cancer cell line. *BioMed Res Int*. 2014;2014:365867. https://doi.org/10.1155/2014/365867.
28. Schmidt B, Liebenberg V, Dietrich D, et al. SHOX2 DNA methylation is a biomarker for the diagnosis of lung cancer based on bronchial aspirates. *BMC Cancer*. 2010;10: 600. https://doi.org/10.1186/1471-2407-10-600.
29. Fox SB, Generali DG, Harris AL. Breast tumour angiogenesis. *Breast Cancer Res*. 2007; 9(6):216. https://doi.org/10.1186/bcr1796.
30. Peterson LE, Kovyrshina T. Progression inference for somatic mutations in cancer. *Heliyon*. 2017;3:e00277. https://doi.org/10.1016/j.heliyon.2017.e00277.
31. Liu B, Hu FF, Zhang Q, et al. Genomic landscape and mutational impacts of recurrently mutated genes in cancers. *Mol Genet Genomic Med*. 2018;6(6):910–923. https://doi.org/10.1002/mgg3.458.
32. Iizuka S, Ishimaru N, Kudo Y. Matrix metalloproteinases: the gene expression signatures of head and neck cancer progression. *Cancers*. 2014;6(1):396–415. https://doi.org/10.3390/cancers6010396.
33. Halder SK, Osteen KG, Al-Hendy A. Vitamin D3 inhibits expression and activities of matrix metalloproteinase-2 and -9 in human uterine fibroid cells. *Hum Reprod*. 2013; 28(9):2407–2416. https://doi.org/10.1093/humrep/det265.
34. Courtoy GE, Donnez J, Ambroise J, et al. Gene expression changes in uterine myomas in response to ulipristal acetate treatment. *Reproductive BioMedicine*. 2018;37(2): 224–550. https://doi.org/10.1016/j.rbmo.2018.04.050.
35. Hoelting L, Scheinhardt B, Bondarenko O, et al. A 3-dimensional human embryonic stem cell (hESC)-derived model to detect developmental neurotoxicity of nanoparticles. *Arch Toxicol*. 2013;87(4):721–733. https://doi.org/10.1007/s00204-012-0984-2.
36. Liu YC, Yeh CT, Lin KH. Molecular functions of thyroid hormone signaling in regulation of cancer progression and anti-apoptosis. *Int J Mol Sci*. 2019;20(20):4986. https://doi.org/10.3390/ijms20204986.
37. Liu X, Liu Y, Zhao J, et al. Screening of potential biomarkers in uterine leiomyomas disease via gene expression profiling analysis. *Mol Med Rep*. 2018;17(5):6985–6996. https://doi.org/10.3892/mmr.2018.8756.
38. Teixeira MR, Silva J, Gomes M, et al. SHOX2 overexpression in uterine leiomyosarcoma is associated with a poor clinical outcome. *Int J Oncol*. 2018;52(4): 1252–1264. https://doi.org/10.3892/ijo.2018.4292.
39. Ak A, Cetin B, Agacayak E, et al. Expression profiling of genes related to leiomyoma pathogenesis in leiomyoma and myometrium. *J Obstet Gynaecol Res*. 2021;47(7): 2321–2331. https://doi.org/10.1111/jog.14726.
40. Seluanov A, Hine C, Azpurua J, et al. Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat. *Proc Natl Acad Sci USA*. 2009;106(46): 19352–19357. https://doi.org/10.1073/pnas.0905252106.
41. Mendonsa AM, Na TY, Gumbiner BM. E-cadherin in contact inhibition and cancer. *Oncogene*. 2018;37(35):4769–4780. https://doi.org/10.1038/s41388-018-0304-2.
42. West J, Newton PK. Cellular interactions constrain tumor growth. *Proc Natl Acad Sci USA*. 2019;116(6):1918–1923. https://doi.org/10.1073/pnas.1804150116.
43. McClatchey AI, Yap AS. Contact inhibition (of proliferation) redux. *Curr Opin Cell Biol*. 2012;24(5):685–694. https://doi.org/10.1016/j.ceb.2012.06.009.
44. Pavel M, Renna M, Park SJ, et al. Contact inhibition controls cell survival and proliferation via YAP/TAZ-autophagy axis. *Nat Commun*. 2018;9(1):2961. https://doi.org/10.1038/s41467-018-05388-x.