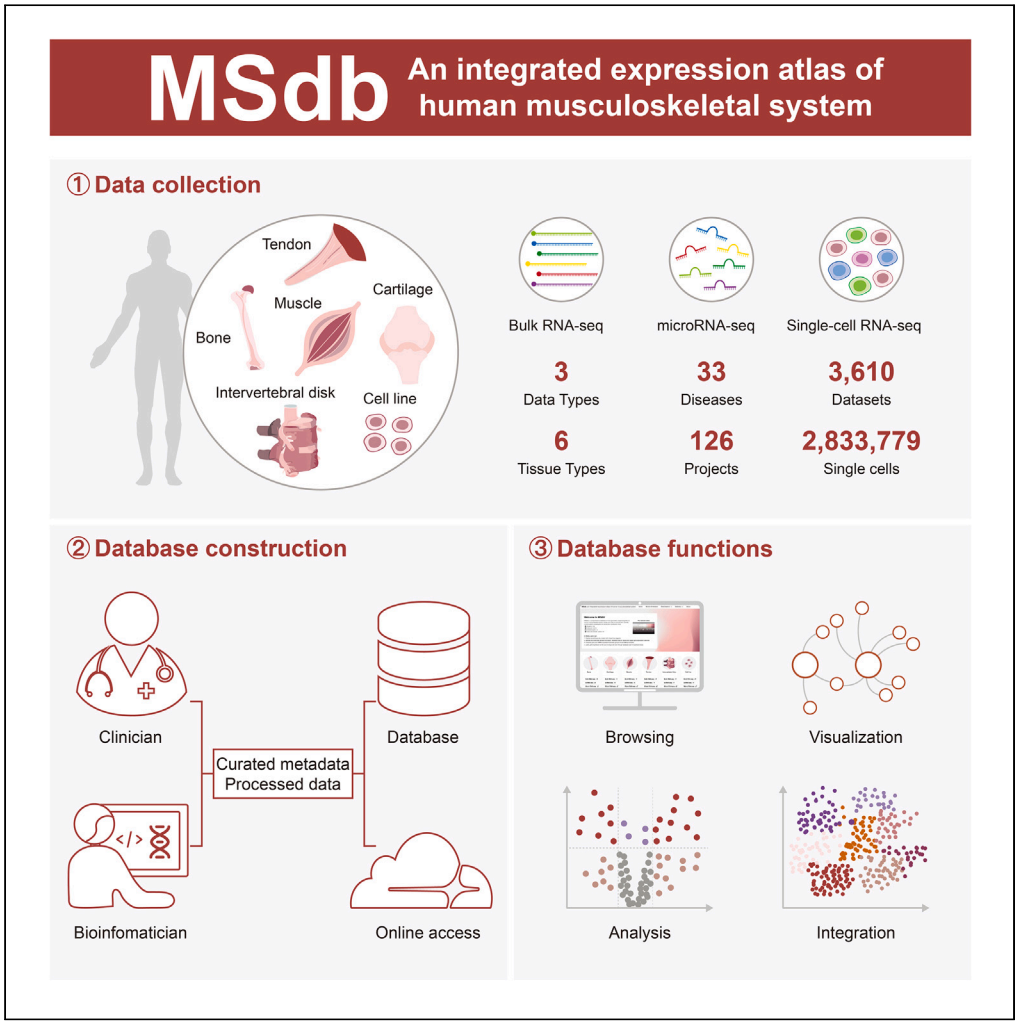
iScience

CellPress
OPEN ACCESS

Article

MSdb: An integrated expression atlas of human musculoskeletal system



Ruonan Tian, Ziwei Xue, Dengfeng Ruan, ..., Hongwei Ouyang, Wanlu Liu, Junxin Lin

hwoy@zju.edu.cn (H.O.)
wanluliu@intl.zju.edu.cn (W.L.)
linjunxin@zju.edu.cn (J.L.)

Highlights

A comprehensive database for human musculoskeletal system gene expression data

Systematically sorted metadata facilitate the reuse of public data

Various online analysis functionalities improve the data mining efficiency

Tian et al., iScience 26, 106933
June 16, 2023
https://doi.org/10.1016/j.isci.2023.106933

iScience

CellPress
OPEN ACCESS

Article

MSdb: An integrated expression atlas of human musculoskeletal system

Ruonan Tian,[1,2,8] Ziwei Xue,[1,2,8] Dengfeng Ruan,[1,3,8] Pengwei Chen,[1] Yiwen Xu,[1,3] Chao Dai,[1] Weiliang Shen,[3,4,5,6] Hongwei Ouyang,[1,3,4,5,6,*] Wanlu Liu,[1,2,3,5,7,*] and Junxin Lin[1,3,9,*]

SUMMARY

The global prevalence and burden of musculoskeletal (MSK) disorders are immense. Advancements in next-generation sequencing (NGS) have generated vast amounts of data, accelerating the research of pathological mechanisms and the development of therapeutic approaches for MSK disorders. However, scattered datasets across various repositories complicate uniform analysis and comparison. Here, we introduce MSdb, a database for visualization and integrated analysis of next-generation sequencing data from human musculoskeletal system, along with manually curated patient phenotype data. MSdb provides various types of analysis, including sample-level browsing of metadata information, gene/miRNA expression, and single-cell RNA-seq dataset. In addition, MSdb also allows integrated analysis for cross-samples and cross-omics analysis, including customized differentially expressed gene/microRNA analysis, microRNA-gene network, scRNA-seq cross-sample/disease integration, and gene regulatory network analysis. Overall, systematic categorizing, standardized processing, and freely accessible knowledge features MSdb a valuable resource for MSK research community.

INTRODUCTION

The rising burden of musculoskeletal (MSK) disorders constitutes one of the major global health challenges. In 2019, over 1600 million adults throughout the world were estimated to have a condition that would benefit from rehabilitation, with MSK disorders accounting for nearly two-thirds of these cases.[1] Due to population growth and aging, the number of people living with MSK disorders and associated functional limitations is continuously growing. With the amplification of this trend, robust prevention and treatment approaches are urgently needed.

In the past decades, scientists have been devoted to understand the mechanisms of MSK disorders and development in order to develop effective therapeutic, regenerative or rehabilitative treatments.[2] The widespread use of next-generation sequencing (NGS) technologies has led to the accumulation of a wealth of sequencing data of human physiological and pathological MSK tissues, which has sparked the identification of key regulators in MSK disorders. By utilizing gene expression data, recent studies have discovered disease subtypes of rheumatoid arthritis (RA) and osteoarthritis, advancing our knowledge of the mechanisms underlying arthritis and providing valuable information for precise diagnosis and treatment.[3–7] In addition, integrative analysis of microRNAs and gene expression data revealed gene regulatory networks (GRNs) governing MSK homeostasis and disease, as well as providing diagnostic markers for MSK disorders.[8–14] In recent years, the emergence of single-cell technology permits researchers to discover unexpected biological findings relative to bulk-level profiling. Leveraging single-cell RNA sequencing (scRNA-seq), scientists have revealed developmentally and pathologically important cell populations, and tracked the trajectories of distinct cell lineages in bone, cartilage, muscle, intervertebral disk and tendon.[15–22] Overall, the rapid emergence of NGS data in recent years offers us an unprecedented opportunity to understand molecular and cellular mechanisms of MSK disorders and development.

Although the majority of these valuable NGS datasets might be easily accessed from public databases (such as GEO and EMBL-EBI), there are a number of obstacles that prevent their use: (1) inconsistent sample information or metadata makes it difficult for researchers to perform comparative analysis, (2) the

[1]Zhejiang University-University of Edinburgh Institute, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310058, China

[2]Future Health Laboratory, Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing, Zhejiang 314100, China

[3]Dr. Li Dak Sum & Yip Yio Chin Center for Stem Cells and Regenerative Medicine, and Department of Orthopedic Surgery of the Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310058, China

[4]Department of Sports Medicine, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310058, China

[5]Key Laboratory of Tissue Engineering and Regenerative Medicine of Zhejiang Province, Hangzhou, Zhejiang 310058, China

[6]China Orthopedic Regenerative Medicine Group (CORMed), Hangzhou, Zhejiang 310058, China

[7]Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Zhejiang University, Hangzhou, Zhejiang 310058, China

[8]These authors contributed equally

[9]Lead contact

*Correspondence: hwoy@zju.edu.cn (H.O.), wanluliu@intl.zju.edu.cn (W.L.), linjunxin@zju.edu.cn (J.L.)

https://doi.org/10.1016/j.isci.2023.106933

datasets are often in raw data format, which may result in difficulties in data mining and analysis for researchers without sufficient computational powers and bioinformatics skills, (3) even if the pre-processed data are provided, the nonuniform pipelines make integrative and comparative analysis across datasets challenging.

Therefore, we propose that a carefully curated NGS database created specifically for the field of orthopedic research will greatly benefit clinical and wet lab researchers. To this end, we developed the human musculoskeletal system database (MSdb), a multi-omics human musculoskeletal NGS database with a multi-functional and user-friendly web interface. MSdb provides convenient ways for researchers to explore gene expression at bulk and single-cell level and may provide further insights into musculoskeletal development and diseases.

RESULTS

Overview of MSdb

We developed the human musculoskeletal system database (MSdb), an integrated expression atlas specifically for the human MSK system, containing 33 diseases, 126 projects, 3,398 transcriptomes, and microRNAomes at bulk level, as well as 2,833,779 transcriptomes at single-cell level (Figures 1A and S1). MSdb incorporates cross-repository metadata into controlled vocabulary and uniform format, enabling efficient retrieval of sample information (Table S1). MSdb provides multiple built-in data exploration and analysis functionalities, including gene/microRNA expression browsing, customized differentially expressed genes/miRNAs analysis, integrated microRNA-gene interaction networks, as well as integrated single-cell expression atlas and cell type-specific GRNs analysis (Figure S2 and Video S1). Furthermore, MSdb allows downloading of processed datasets and publication-quality plots, offering wet lab scientists powerful tools to browse and re-analyze the public datasets without technical barriers.

MSdb enables users to retrieve sample information via consistent and validated metadata curated by orthopedists and bioinformatics scientists (Table S1). Users can search the database using multiple parameters like project identifier, diseases, and tissue types in order to find datasets that match their interests (Figure S3A). In metadata, four types of information are available: (i) dataset and publication identifiers, (ii) patient phenotypes, (iii) sample information, and (iv) data pre-processing summary. These items enable users to evaluate the biological meaning, clinical relevance and data quality of the samples. Summary statistics of the metadata are also presented to show global patterns of the studies (Figure S4).

Transcriptome and miRNAome modules of MSdb

At bulk level, MSdb integrates information at two aspects: (i) cross-tissue integration and (ii) cross-omics integration. For cross-tissue integration, samples were initially integrated by projects to generate gene or microRNA expression atlas, and then labeled by their tissue types, diseases, cell types, and tissue positions. Users may explore the expression of genes or microRNAs by these labels (Figures S5 and S6). In Figures 1B and 1C, uniform manifold approximation and projection (UMAP) plots show the integrated expression atlas of bulk RNA-seq or microRNA-seq data in MSdb, and samples are colored by representative tissue types and diseases. Feature plots and violin plots show that *COL3A1*, a connective tissue marker, was pervasively expressed in MSK tissues as expected (Figures 1B and S5). Cerebrospinal fluid from amyotrophic lateral sclerosis (ALS) patients showed enrichment of the potential diagnostic marker miR-4649-5p (Figures 1C and S6).[23] For cross-omics integration, MSdb enables users to analyze bulk RNA-seq and microRNA-seq data side-by-side and integrate the results to predict disease-related gene expression regulatory mechanisms. MSdb's differential expression analysis module allows users to choose two groups of samples for comparison (Figure S3B). An interactive volcano plot can be used to explore differentially expressed genes or microRNAs between user-defined groups, and the expression level of a specific gene or microRNA can be queried and will be displayed in the boxplot (Figure S3B). By integrating differentially expressed genes and microRNAs between normal and pathological intervertebral disks, we constructed a microRNA-gene interaction network (Figures 1D and 1E). The network revealed previous reported (indicated by asterisk) and unanticipated disease-associated microRNAs and their potential gene targets, which can be interactively explored on our website (Figure 1E). We demonstrated that the expression level of *miR-146a-5p* was down-regulated, while its target SPP1 was up-regulated in degenerative intervertebral disks when compared to healthy control (Figures 1D and 1E). Since abnormal expression of SPP1 plays a critical role in the pathological process of intervertebral disk degeneration (IDD),[24] our analysis indicated that targeting SPP1 with *miR-146a-5p* mimic might be a therapeutic method

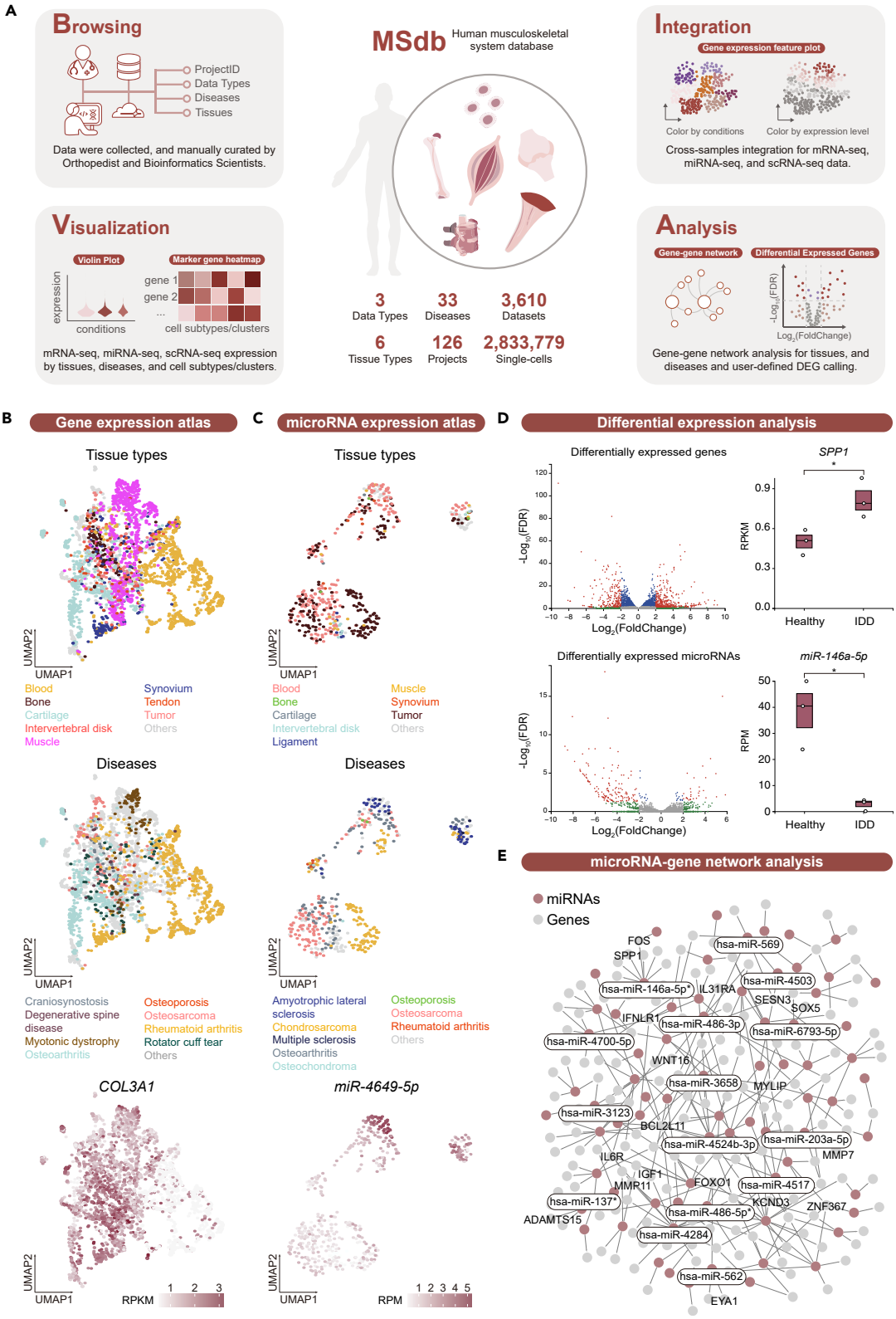

A
Browsing
ProjectID
Data Types
Diseases
Tissues
Data were collected, and manually curated by Orthopedist and Bioinformatics Scientists.
MSdb
Human musculoskeletal system database
3
Data Types
33
Diseases
3,610
Datasets
6
Tissue Types
126
Projects
2,833,779
Single-cells
Integration
Gene expression feature plot
Color by conditions
Color by expression level
Cross-samples integration for mRNA-seq, miRNA-seq, and scRNA-seq data.
Visualization
Violin Plot
Marker gene heatmap
expression
conditions
gene 1
gene 2
cell subtypes/clusters
mRNA-seq, miRNA-seq, scRNA-seq expression by tissues, diseases, and cell subtypes/clusters.
Analysis
Gene-gene network
Differential Expressed Genes
Gene-gene network analysis for tissues, and diseases and user-defined DEG calling.
B
Gene expression atlas
Tissue types
UMAP1
UMAP2
Blood
Bone
Cartilage
Intervertebral disk
Muscle
Synovium
Tendon
Tumor
Others
Diseases
Craniosynostosis
Degenerative spine disease
Myotonic dystrophy
Osteoarthritis
Osteoporosis
Osteosarcoma
Rheumatoid arthritis
Rotator cuff tear
Others
COL3A1
RPKM
1 2 3
C
microRNA expression atlas
Tissue types
Blood
Bone
Cartilage
Intervertebral disk
Ligament
Muscle
Synovium
Tumor
Others
Diseases
Amyotrophic lateral sclerosis
Chondrosarcoma
Multiple sclerosis
Osteoarthritis
Osteochondroma
Osteoporosis
Osteosarcoma
Rheumatoid arthritis
Others
miR-4649-5p
RPM
1 2 3 4 5
D
Differential expression analysis
Differentially expressed genes
-Log10(FDR)
Log2(FoldChange)
SPP1
RPKM
Healthy
IDD
Differentially expressed microRNAs
miR-146a-5p
RPM
E
microRNA-gene network analysis
miRNAs
Genes
FOS
SPP1
hsa-miR-569
hsa-miR-4503
hsa-miR-146a-5p*
IL31RA
SESN3
SOX5
IFNLR1
hsa-miR-486-3p
hsa-miR-6793-5p
hsa-miR-4700-5p
WNT16
hsa-miR-3658
MYLIP
hsa-miR-3123
BCL2L11
hsa-miR-4524b-3p
hsa-miR-203a-5p
MMP7
IL6R
IGF1
MMP11
FOXO1
hsa-miR-4517
hsa-miR-137*
KCND3
ZNF367
ADAMTS15
hsa-miR-486-5p*
hsa-miR-4284
hsa-miR-562
EYA1

Figure 1. MSdb framework and illustrative data analysis

(A) Overview of MSdb. MSdb is a comprehensive database of next-generation sequencing data on human musculoskeletal system tissues and cells, enhanced with manually curated patient phenotypes, advanced analysis, and visualization tools.

(B) UMAP plots showing gene expression atlas in MSdb. All gene expression data in MSdb were used for clustering. Samples are colored by tissue types (top), diseases (middle), and COL3A1 expression levels (bottom).

(C) UMAP plots showing microRNA expression atlas in MSdb. All microRNA expression data in MSdb were used for clustering. Samples are colored by tissue types (top), diseases (middle), and miR-4649-5p expression levels (bottom).

(D) Volcano plots and boxplots showing dysregulated genes (top) or microRNAs (bottom) between healthy (n = 3) and degenerative (n = 3) intervertebral disks. RPKM: reads per kilobase per million mapped reads; RPM: reads per million mapped reads. *: FDR <0.01.

(E) microRNA-gene interaction network built with down-regulated microRNAs and up-regulated genes in degenerative intervertebral disks. Red dots represent the microRNAs and gray dots represent the genes. Complete and interactive plots of the network are available online on MSdb database.

to counteract disk degeneration. Collectively, MSdb offers users with integrated expression atlas and useful data analysis functionalities to understand gene function and regulation in homeostasis and diseases.

Single-cell transcriptome module of MSdb

The advent of single-cell technology has enabled researchers to uncover new biological insights compared to bulk RNA-seq. Intriguingly, MSdb contains a wealth of single-cell RNA sequencing (scRNA-seq) data and provides a suit of functionalities for users to explore gene expression at single-cell level. For each scRNA-seq dataset, we provide textual and graphical representations for sample information, quality control metrics, unsupervised cell clustering, reference-based cell subtype prediction, as well as marker genes for each cell type (Figure S7). In a single-cell profiling of synovial tissue from a female patient with RA, cell types including fibroblasts, macrophages, T cells, and monocytes et al. were identified, and the specific *CD68* expression in the predicted macrophages (cluster 2) indicated the reliability of cell clustering and automated cell type prediction (Figures S7A–S7H). For more sophisticated cell type clustering and annotation, users can adjust leiden resolution for cell clustering and change reference dataset for cell type prediction (Figure S7I). To help with manual annotation, a heatmap and a table of marker genes are displayed and available for download. (Figures S7J and S7K).

We also implemented an in-house variational autoencoder (VAE)-based deep-learning framework (scVAE) to facilitate the integrative analysis for scRNA-seq datasets from different studies. In line with a recent study, our analysis demonstrated that the scVAE method produced better integration and cell clustering results while using Harmony for data integration led to the inter-sample heterogeneity being skewed by over-integration (Figures 2 and S8), indicating the advantages of variational autoencoder-based algorithms in complex data integration.[25] Figure 2 displays the result of integrated analysis of single-cell gene expression data from healthy, osteoarthritis (OA), rheumatoid arthritis (RA), and undifferentiated arthritis (UA) patients. Using the VAE model, we were able to remove batch differences and integrate heterogeneous data from different studies (Figure 2B). The cell types could be identified by the known marker genes, whose expression patterns support that our integration method appropriately aligned gene expression for each cell types (Figures 2A and 2C). Interestingly, we observed a distinct distribution pattern of fibroblasts from OA and RA patients (Figure 2B). Differential expression analysis revealed that *IGFBP3* and *LOX* were more enriched in OA-derived fibroblast when compared to RA-derived fibroblast (Figure 2D). *IGFBP3* and *LOX* were involved in extracellular matrix remodeling, which may determine synovial fibrosis and may be associated with the clinical symptoms of pain, hyperalgesia, and stiffness in osteoarthritis.[26] It was also noted that *CD74* and *HLA-DRA* were specifically expressed in RA fibroblasts, demonstrating an inflammatory state of fibroblasts that have been reported to be a major source of pro-inflammatory cytokines and highlighted as a potential therapeutic target in RA (Figure 2D).[27,28] Moreover, we inferred GRNs using a previously published deep regenerative model for each cell types.[29] Users may interactively visualize the GRNs in our database to explore the complicated molecular interactions governing potential cell identity (Figure S9).

DISCUSSION

Musculoskeletal conditions, including RA, osteoarthritis, low back pain, fractures, amputation, and other injuries, affect people of all ages everywhere around the world.[1] In recent years, vast amount of NGS data have emerged to facilitate the study of regulatory mechanisms governing musculoskeletal disorders, which has greatly improved the disease diagnosis and patient management. Considering the value of these data, it thus will be of great benefit to researchers in the musculoskeletal system field to have all these datasets in one database for free exploration. Our MSdb database permits efficient searching of its content

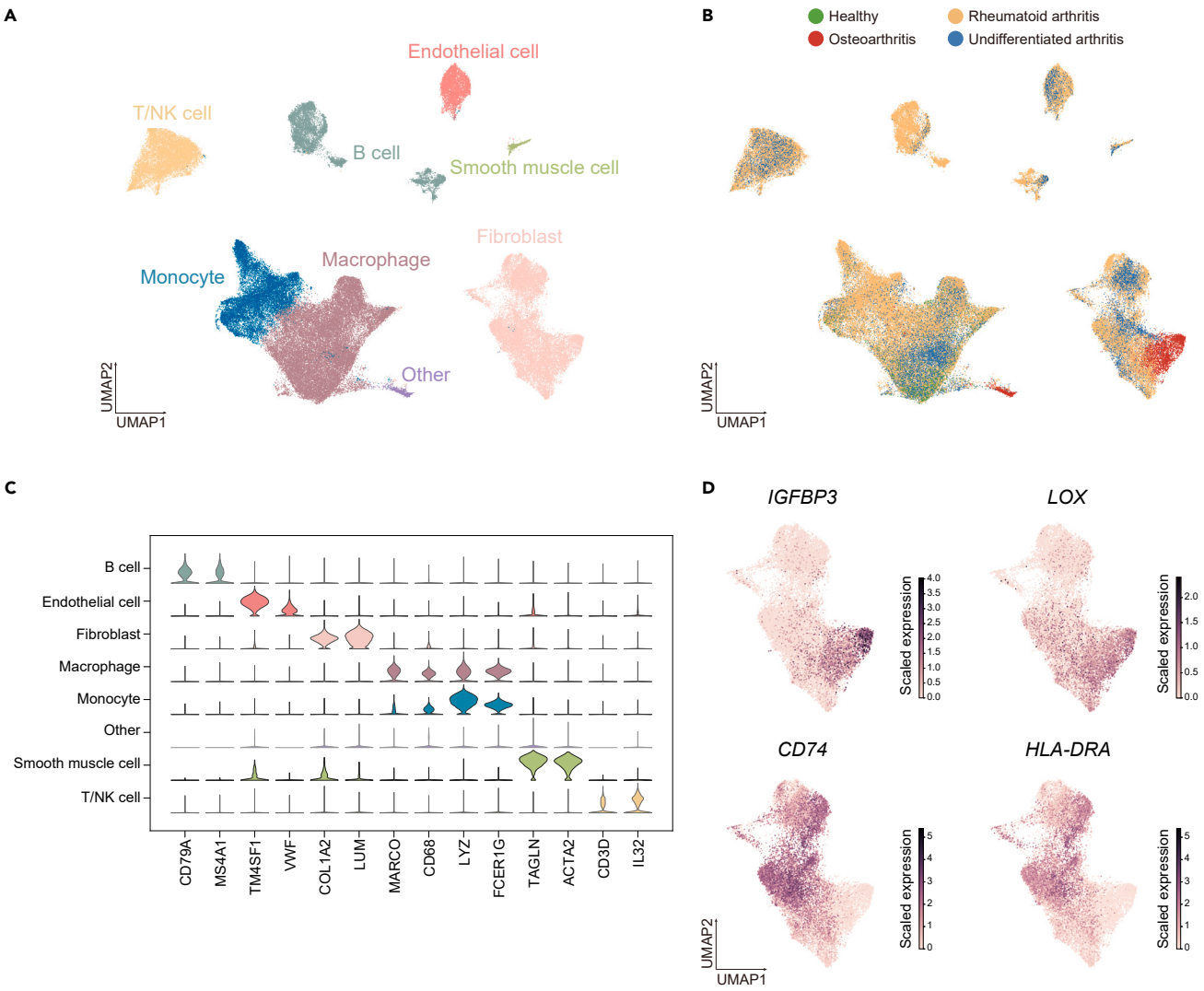


Figure 2. Integrated analysis of scRNA-seq data of synovial tissues-derived cells from OA, RA, UA and healthy patients

(A) UMAP plots showing clustering of 101,610 cells from healthy, osteoarthritis, rheumatoid arthritis, and undifferentiated arthritis patients. Cells are colored by cell types.

(B) UMAP plots showing clustering of 101,610 cells from healthy, osteoarthritis, rheumatoid arthritis, and undifferentiated arthritis patients. Cells are colored by diseases.

(C) Violin plots showing marker gene expression of each cell type.

(D) Scatterplot showing the expression level of genes specifically expressed in RA- or OA-derived fibroblasts. Cells are colored by the expression of the indicated genes.

containing comprehensive information for public bulk RNA-Seq, microRNA-seq, and single-cell RNA-seq datasets related to human musculoskeletal system development and disease. It provides initial analyses of the data including gene/microRNA expression levels, cell clustering, cell type annotation and marker gene identification. It also contains a multi-functional and user-friendly web interface, which provides various ways for researchers to explore gene expression at bulk and single-cell level.

Overall, MSdb is a resource created for the MSK research community and aims to fulfill the findability, accessibility, interoperability, and reusability (FAIR) principles of scholarly data.[30] The uniformity of sample information in MSdb enables metadata-based and database-scale analysis. MSdb's utility will continue to grow as public NGS datasets of human musculoskeletal system expand. We envision that it will broaden the use of human MSK datasets and will be invaluable to researchers in the MSK field.

Limitations of the study

The current study presents certain limitations that warrant discussion. Firstly, MSdb contains only the bulk RNA-seq, microRNA-seq and scRNA-Seq datasets, and has not collected other omics, such as genomics and epigenetics data. In the future, we will collect more datasets, including Bisulfite-seq, ChIP-seq, ATAC-seq and Hi-C seq to provide multiple layers of regulatory information that controls gene expression. Secondly, integrating or splitting scRNA-seq data by individual patients or subjects could facilitate the discovery of shared molecular mechanisms underlying diseases. However, this feature necessitates substantial online computing resources, which our current web server cannot accommodate. We plan to incorporate this functionality in the subsequent version of the database. Lastly, the integration of scRNA-seq data is presently confined to arthritis-related diseases, as the inclusion of scRNA-seq datasets from various tissues and diseases would introduce extraneous information, complicating the interpretation of results. Nonetheless, with the increasing of single-cell datasets, we could create more integrated maps of pertinent diseases for MSK database.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data collection and meta information curation
  - Bulk RNA-seq and microRNA-seq processing and data analysis
  - Single cell RNA-seq data processing
  - Integrated scRNA-seq data analysis
  - Website development

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106933.

ACKNOWLEDGMENTS

The authors would like to thank all researchers who generated the data sets that are collected, analyzed, and displayed in MSdb. We thank Dr. Xiao Chen, Dr. Zi Yin, Dr. Wenyan Zhou, Dr. Can Zhang, Dr. Xiaolei Zhang, Fan Jiayi, and clinicians Dr. Yan Wu (M.D.), Dr. Yejun Hu (M.D.), Dr. Kun Zhao (M.D.), Dr. Yuzi Xu (M.D.), and Dr. Geyu Gu for their helpful discussions and valuable suggestions. We would also like to thank the technical support provided by the Core Facilities, especially the ZJE server of ZJU-UoE Institute. This work was funded by National Natural Science Foundation of China (T2121004), Fundamental Research Funds for the Central Universities (226-2022-00134), and Alibaba Cloud.

AUTHOR CONTRIBUTIONS

J.L., W.L., H.O., and W.S. conceived the study and designed database. R.T., P.C., C.D., and J.L. collected and processed the data. R.T., D.R., Y.X., and J.L. curated the metadata. R.T. and Z.X. developed the database and web server. R.T., Z.X., D.R., J.L., and W.L. analyzed the data. R.T., Z.X., D.R., W.L., and J.L. wrote the manuscript. All authors contributed to the review and corrections of the manuscripts.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: December 13, 2022
Revised: April 26, 2023
Accepted: May 16, 2023
Published: May 18, 2023

REFERENCES

1. Cieza, A., Causey, K., Kamenov, K., Hanson, S.W., Chatterji, S., and Vos, T. (2021). Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study. Lancet *396*, 2006–2017. https://doi.org/10.1016/S0140-6736(20)32340-0.

2. Paskins, Z., Farmer, C.E., Manning, F., Andersson, D.A., Barlow, T., Bishop, F.L., Brown, C.A., Clark, A., Clark, E.M., Dulake, D., et al. (2022). Research priorities to reduce the impact of musculoskeletal disorders: a priority setting exercise with the child health and nutrition research initiative method. Lancet. Rheumatol. *4*, e635–e645. https://doi.org/10.1016/s2665-9913(22)00136-9.

3. Soul, J., Dunn, S.L., Anand, S., Serracino-Inglott, F., Schwartz, J.M., Boot-Handford, R.P., and Hardingham, T.E. (2018). Stratification of knee osteoarthritis: two major patient subgroups identified by genome-wide expression analysis of articular cartilage. Ann. Rheum. Dis. *77*, 423. https://doi.org/10.1136/annrheumdis-2017-212603.

4. Yuan, C., Pan, Z., Zhao, K., Li, J., Sheng, Z., Yao, X., Liu, H., Zhang, X., Yang, Y., Yu, D., et al. (2020). Classification of four distinct osteoarthritis subtypes with a knee joint tissue transcriptome atlas. Bone Res. *8*, 38. https://doi.org/10.1038/s41413-020-00109-x.

5. Coutinho de Almeida, R., Mahfouz, A., Mei, H., Houtman, E., den Hollander, W., Soul, J., Suchiman, E., Lakenberg, N., Meessen, J., Huetink, K., et al. (2021). Identification and characterization of two consistent osteoarthritis subtypes by transcriptome and clinical data integration. Rheumatology *60*, 1166–1175. https://doi.org/10.1093/rheumatology/keaa391.

6. Orange, D.E., Agius, P., DiCarlo, E.F., Robine, N., Geiger, H., Szymonifka, J., McNamara, M., Cummings, R., Andersen, K.M., Mirza, S., et al. (2018). Identification of three rheumatoid arthritis disease subtypes by machine learning integration of synovial histologic features and RNA sequencing data. Arthritis Rheumatol. *70*, 690–701. https://doi.org/10.1002/art.40428.

7. Pinal-Fernandez, I., Casal-Dominguez, M., Derfoul, A., Pak, K., Miller, F.W., Milisenda, J.C., Grau-Junyent, J.M., Selva-O'Callaghan, A., Carrion-Ribas, C., Paik, J.J., et al. (2020). Machine learning algorithms reveal unique gene expression profiles in muscle biopsies from patients with different types of myositis. Ann. Rheum. Dis. *79*, 1234–1242. https://doi.org/10.1136/annrheumdis-2019-216599.

8. Kosela-Paterczyk, H., Paziewska, A., Kulecka, M., Balabas, A., Kluska, A., Dabrowska, M., Piatkowska, M., Zeber-Lubecka, N., Ambrozkiewicz, F., Karczmarski, J., et al. (2020). Signatures of circulating microRNA in four sarcoma subtypes. J. Cancer *11*, 874–882. https://doi.org/10.7150/jca.34723.

9. Lietz, C.E., Garbutt, C., Barry, W.T., Deshpande, V., Chen, Y.L., Lozano-Calderon, S.A., Wang, Y., Lawney, B., Ebb, D., Cote, G.M., et al. (2020). MicroRNA-mRNA networks define translatable molecular outcome phenotypes in osteosarcoma. Sci. Rep. *10*, 4409. https://doi.org/10.1038/s41598-020-61236-3.

10. Geng, Y., Chen, J., Chang, C., Zhang, Y., Duan, L., Zhu, W., Mou, L., Xiong, J., and Wang, D. (2021). Systematic analysis of mRNAs and ncRNAs in BMSCs of senile osteoporosis patients. Front. Genet. *12*, 776984. https://doi.org/10.3389/fgene.2021.776984.

11. Urdinez, J., Boro, A., Mazumdar, A., Arlt, M.J., Muff, R., Botter, S.M., Bode-Lesniewska, B., Fuchs, B., Snedeker, J.G., and Gvozdenovic, A. (2020). The miR-143/145 cluster, a novel diagnostic biomarker in chondrosarcoma, acts as a tumor suppressor and directly inhibits fascin-1. J. Bone Miner. Res. *35*, 1077–1091. https://doi.org/10.1002/jbmr.3976.

12. Nicolle, R., Ayadi, M., Gomez-Brouchet, A., Armenoult, L., Banneau, G., Elarouci, N., Tallegas, M., Decouvelaere, A.V., Aubert, S., Rédini, F., et al. (2019). Integrated molecular characterization of chondrosarcoma reveals critical determinants of disease progression. Nat. Commun. *10*, 4622. https://doi.org/10.1038/s41467-019-12525-7.

13. Lorenzi, L., Chiu, H.S., Avila Cobos, F., Gross, S., Volders, P.J., Cannoodt, R., Nuytens, J., Vanderheyden, K., Anckaert, J., Lefever, S., et al. (2021). The RNA Atlas expands the catalog of human non-coding RNAs. Nat. Biotechnol. *39*, 1453–1465. https://doi.org/10.1038/s41587-021-00936-1.

14. Li, Z., Sun, Y., He, M., and Liu, J. (2021). Differentially-expressed mRNAs, microRNAs and long noncoding RNAs in intervertebral disc degeneration identified by RNA-sequencing. Bioengineered *12*, 1026–1039. https://doi.org/10.1080/21655979.2021.1899533.

15. Xi, H., Langerman, J., Sabri, S., Chien, P., Young, C.S., Younesi, S., Hicks, M., Gonzalez, K., Fujiwara, W., Marzi, J., et al. (2020). A human skeletal muscle atlas identifies the trajectories of stem and progenitor cells across development and from human pluripotent stem cells. Cell Stem Cell *27*, 181–185. https://doi.org/10.1016/j.stem.2020.06.006.

16. Alivernini, S., MacDonald, L., Elmesmari, A., Finlay, S., Tolusso, B., Gigante, M.R., Petricca, L., Di Mario, C., Bui, L., Perniola, S., et al. (2020). Distinct synovial tissue macrophage subsets regulate inflammation and remission in rheumatoid arthritis. Nat. Med. *26*, 1295–1306. https://doi.org/10.1038/s41591-020-0939-8.

17. He, J., Yan, J., Wang, J., Zhao, L., Xin, Q., Zeng, Y., Sun, Y., Zhang, H., Bai, Z., Li, Z., et al. (2021). Dissecting human embryonic skeletal stem cell ontogeny by single-cell transcriptomic and functional analyses. Cell Res. *31*, 742–757. https://doi.org/10.1038/s41422-021-00467-z.

18. Nakajima, T., Nakahata, A., Yamada, N., Yoshizawa, K., Kato, T.M., Iwasaki, M., Zhao, C., Kuroki, H., and Ikeya, M. (2021). Grafting of iPS cell-derived tenocytes promotes motor function recovery after Achilles tendon rupture. Nat. Commun. *12*, 5012. https://doi.org/10.1038/s41467-021-25328-6.

19. Ji, Q., Zheng, Y., Zhang, G., Hu, Y., Fan, X., Hou, Y., Wen, L., Li, L., Xu, Y., Wang, Y., and Tang, F. (2019). Single-cell RNA-seq analysis reveals the progression of human osteoarthritis. Ann. Rheum. Dis. *78*, 100–110. https://doi.org/10.1136/annrheumdis-2017-212863.

20. Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., et al. (2020). Construction of a human cell landscape at single-cell level. Nature *581*, 303–309. https://doi.org/10.1038/s41586-020-2157-4.

21. Gan, Y., He, J., Zhu, J., Xu, Z., Wang, Z., Yan, J., Hu, O., Bai, Z., Chen, L., Xie, Y., et al. (2021). Spatially defined single-cell transcriptional profiling characterizes diverse chondrocyte subtypes and nucleus pulposus progenitors in human intervertebral discs. Bone Res. *9*, 37. https://doi.org/10.1038/s41413-021-00163-z.

22. Zhou, Y., Yang, D., Yang, Q., Lv, X., Huang, W., Zhou, Z., Wang, Y., Zhang, Z., Yuan, T., Ding, X., et al. (2020). Single-cell RNA landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. Nat. Commun. *11*, 6322. https://doi.org/10.1038/s41467-020-20059-6.

23. Takahashi, I., Hama, Y., Matsushima, M., Hirotani, M., Kano, T., Hohzen, H., Yabe, I., Utsumi, J., and Sasaki, H. (2015). Identification of plasma microRNAs as a biomarker of sporadic Amyotrophic Lateral Sclerosis. Mol. Brain *8*, 67. https://doi.org/10.1186/s13041-015-0161-7.

24. Marfia, G., Navone, S.E., Di Vito, C., Tabano, S., Giammattei, L., Di Cristofori, A., Gualtierotti, R., Tremolada, C., Zavanone, M., Caroli, M., et al. (2015). Gene expression profile analysis of human mesenchymal stem cells from herniated and degenerated intervertebral discs reveals different expression of osteopontin. Stem Cell. Dev.

24, 320–328. https://doi.org/10.1089/scd.2014.0282.

25. Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., and Theis, F.J. (2022). Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods *19*, 41–50. https://doi.org/10.1038/s41592-021-01336-8.

26. Zhang, L., Xing, R., Huang, Z., Ding, L., Zhang, L., Li, M., Li, X., Wang, P., and Mao, J. (2021). Synovial fibrosis involvement in osteoarthritis. Front. Med. *8*, 684389. https://doi.org/10.3389/fmed.2021.684389.

27. Filer, A. (2013). The fibroblast as a therapeutic target in rheumatoid arthritis. Curr. Opin. Pharmacol. *13*, 413–419. https://doi.org/10.1016/j.coph.2013.02.006.

28. Zhang, F., Wei, K., Slowikowski, K., Fonseka, C.Y., Rao, D.A., Kelly, S., Goodman, S.M., Tabechian, D., Hughes, L.B., Salomon-Escoto, K., et al. (2019). Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. Nat. Immunol. *20*, 928–942. https://doi.org/10.1038/s41590-019-0378-1.

29. Shu, H., Zhou, J., Lian, Q., Li, H., Zhao, D., Zeng, J., and Ma, J. (2021). Modeling gene regulatory networks using neural network architectures. Nat. Comput. Sci. *1*, 491–501. https://doi.org/10.1038/s43588-021-00099-8.

30. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data *3*, 160018. https://doi.org/10.1038/sdata.2016.18.

31. Martin, M. (2011). Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads, *17*, p. 3. https://doi.org/10.14806/ej.17.1.200.

32. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

33. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

34. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

35. Zhang, Y., Parmigiani, G., and Johnson, W.E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom. Bioinform. *2*, lqaa078. https://doi.org/10.1093/nargab/lqaa078.

36. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.

37. Kutmon, M., Ehrhart, F., Willighagen, E.L., Evelo, C.T., and Coort, S.L. (2018). CyTargetLinker App Update: A Flexible Solution for Network Extension in Cytoscape. F1000Res *7*. https://doi.org/10.12688/f1000research.14613.2.

38. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

39. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.

40. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002.

41. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

42. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172. https://doi.org/10.1038/s41590-018-0276-y.

43. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. *41*, D991–D995. https://doi.org/10.1093/nar/gks1193.

44. Madeira, F., Pearce, M., Tivey, A.R.N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., and Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. *50*, W276–W279. https://doi.org/10.1093/nar/gkac240.

45. Hsu, S.D., Lin, F.M., Wu, W.Y., Liang, C., Huang, W.C., Chan, W.L., Tsai, W.T., Chen, G.Z., Lee, C.J., Chiu, C.M., et al. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res. *39*, D163–D169. https://doi.org/10.1093/nar/gkq1107.

46. Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233. https://doi.org/10.1038/s41598-019-41695-z.

47. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0.

48. Büttner, M., Miao, Z., Wolf, F.A., Teichmann, S.A., and Theis, F.J. (2019). A test metric for assessing single-cell RNA-seq batch correction. Nat. Methods *16*, 43–49. https://doi.org/10.1038/s41592-018-0254-1.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
R (v4.2.0)	The R Foundation for Statistical Computing	https://www.r-project.org/
Python3 (v3.9.12)	Python Software Foundation	https://www.python.org/
FastQC (v0.11.9)	N/A	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Cutadapt (v3.7)	Martin et al., 2011[31]	https://github.com/marcelm/cutadapt
STAR (v2.7.9a)	Dobin et al., 2013[32]	https://github.com/alexdobin/STAR
featureCounts (v2.0.3)	Liao et al., 2014[33]	https://subread.sourceforge.net/
Samtools (v1.14)	Li et al., 2009[34]	https://github.com/samtools/samtools
ComBat-seq	Zhang et al., 2020[35]	https://github.com/zhangyuqing/ComBat-seq
Scanpy (v1.9.1)	Wolf et al., 2018[36]	https://github.com/scverse/scanpy
CyTargetLinker (v4.1.0)	Kutmon et al., 2018[37]	https://cytargetlinker.github.io/
Cytoscape (v3.9.1)	Shannon et al., 2003[38]	https://cytoscape.org/
Cell Ranger (v7.0.0)	Zheng et al., 2017[39]	https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome
Drop-seq_tools	Macosko et al., 2015[40]	https://github.com/broadinstitute/Drop-seq
Seurat (v4.1.1)	Hao et al., 2021[41]	https://satijalab.org/seurat/
SingleR (v1.10.0)	Aran et al., 2019[42]	https://github.com/dviraran/SingleR
scVAE	This paper	https://github.com/wanluliuLab/MSdb
Other		
MSdb database	This paper	https://www.msdb.org.cn/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for data and code resources should be directed to and will be fulfilled by the lead contact, Junxin Lin (linjunxin@zju.edu.cn).

Materials availability

This study did not generate any new unique reagents.

Data and code availability

All raw data are available on GEO (https://www.ncbi.nlm.nih.gov/geo/) and EMBL-EBI (https://www.ebi.ac.uk/) repositories. All curated sample information and processed bulk RNA-seq and microRNA-seq data can be downloaded from the MSdb database (https://www.msdb.org.cn). Single-cell RNA-seq matrices are available upon reasonable request. Original code for scVAE can be found at https://github.com/wanluliuLab/MSdb. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data collection and meta information curation

Bulk RNA-seq, microRNAs-seq and single-cell RNA-seq data of human musculoskeletal system were originated from NCBI Gene Expression Omnibus (GEO) and EMBL's European Bioinformatics Institute (EMBL-EBI).[43,44] We manually curated both GEO and EMBL-EBI- derived sample information to provide a coherent and standardized metadata. The resulting collection offers the following information for each data set. 'SampleName' contains the sample's identification code in GEO (e.g. GSM2112324) or EMBL-EBI (e.g.

ERS1034560). 'ProjectID' contains the sample's project identification code in GEO (e.g. GSE80072) or EMBL-EBI (e.g. E-MTAB-4304). 'Publication_DOI' contains the digital object identifier of the original publication. 'Category' indicates the MSK tissues that the data sets are related to. 'AssayType' indicates which sequencing types were implemented on the samples. 'LibraryLayout' refers to pair-end sequencing or single-end sequencing. 'Disease' contains the information about the diseases or health status. 'SourceTissue_type' indicates tissue sources of the biological materials used for sequencing. 'Sourcetissue_condition' indicates whether the tissues are pathological or normal. 'SourceTissue_position' refers to a more specific anatomical location of the 'SourceTissue_type'. 'SourceTissue_celltype' indicates whether whole tissue or a specific cell type in the tissue was used for sequencing. 'OtherInfo' contains other information that can help users to evaluate the biological or clinical relevance of the data, such as whether the patients were response to treatment. 'Age', 'AgeGroup' and 'Gender' of the patients are also presented, if available. To assist evaluating the quality of the data sets, we also provide the following information related to data quality assessment along with metadata: sequencing library preparation kit ('LibraryPrepKit'), average spot length ('AvgSpotLen), sequencing instrument ('Instrument'), total reads ('TotalReads'), uniquely mapped reads ('UniquelyMappedReads'), percentage of uniquely mapped reads ('UniquelyMappedReads%'), percentage of multiple mapped loci ('MultipleLoci%').

Bulk RNA-seq and microRNA-seq processing and data analysis

Quality control (QC) of raw sequencing reads for each project was performed by FastQC (v0.11.9, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Cutadapt (v3.7) was used to find and remove adapter sequences, primers, low quality sequence and other types of unmated sequences.[31] For bulk RNA-seq, the trimmed reads were mapped to the reference index built on human genome assembly GRCh38 (hg38, http://ensembl.org/Homo_sapiens/) using STAR (v2.7.9a) and counts were summarized to the genomic protein coding genes by featureCounts (v2.0.3).[32,33] For microRNA-seq , trimmed reads were mapped to the reference index built on miRbase hairpins and Samtools (v1.14) was used to report alignment summary statistics and calculate the microRNA counts.[34] Batch effect between different studies was estimated and adjusted by ComBat-seq.[35] Then, all bulk RNA-seq gene counts and microRNA-seq counts were merged, respectively. To show sample correlation, Scanpy (v1.9.1) was used to reduce dimension and generate uniform manifold approximation and projection (UMAP).[36] To perform differential expression analysis, a t-test was applied to the normalized RPKM or RPM data and a false discovery rate (FDR) adjusted p value was canulated using Benjamini–Hochberg method.

The MSdb's online differential expression analysis tool (https://www.msdb.org.cn/browse/) was used to obtain differentially expressed mRNAs and miRNAs (FDR <0.01 and fold change >2). Down-regulated microRNAs and up-regulated genes in degenerative intervertebral disks were used for further analysis. CyTargetLinker (v4.1.0) was used to predict and construct the miRNA-gene interaction network with miRTarBase *Homo sapiens* release 8.0 linksets was used as a reference. [37,45] Cytoscape software (v3.9.1) was used for miRNA-mRNA regulatory network visualization.[38]

Single cell RNA-seq data processing

The genome reference used in scRNA-seq analysis is GRCh38. For droplet-based scRNA-seq, the raw data were processed using Cell Ranger (v7.0.0) or Drop-seq_tools with standard pipeline and default parameters to obtain gene expression matrix.[39,40] For full-length scRNA-seq, the data were mapped using STAR (v2.7.9a) and quantified using featureCounts (v2.0.3). To perform downstream analysis, the gene expression matrix containing UMI counts was read into an AnnData object by Scanpy (v1.9.1) in Python3 (v3.9.12).[36] Cells with unique gene counts less than 200 or genes that are detected in less than 3 cells were removed. To perform unsupervised cell clustering analysis, the UMI counts were normalized to counts per million (CPM) with *scanpy.pp.normalize_total* function, followed by log-transformation and principal component analysis (PCA) using *scanpy.pp.log1p* and *scanpy.tl.pca* functions. The neighborhood graph was calculated based on the PCA results using *scanpy.pp.neighbors* function and the Leiden algorithm was used to perform unsupervised cell clustering (*scanpy.tl.leiden*).[46] Marker genes for each cell cluster were identified by *sc.tl.rank_genes_groups* function. To perform reference-based cell subtype prediction, the filtered count matrix was loaded into Seurat (v4.1.1) in R (v4.2.0).[41] The annotation of cell subtype was performed by SingleR (v1.10.0) R-package using different references, including the BlueprintEncodeData and the HumanPrimaryCellAtlasData.[42] Marker genes for each annotated cell types were identified by *FindAllMarkers* in Seurat package. UMAP was used for the data visualization of unsupervised cell clustering, cell type prediction and marker gene expression.

Integrated scRNA-seq data analysis

We built a single-cell atlas of the synovium containing 101,610 cells from 3 studies and 34 samples. We have implemented a probabilistic model based on a variational autoencoder to integrate single-cell RNA-seq data sets and remove batch effects, accepting raw count matrix as input. The variational distribution adopts the log-normal distribution with scalar mean and variance output from the encoder, regularized by the Kullback–Leibler divergence. The decoder takes categorical encoding of the sample name to reflect biological variance and remove batch effects. The count data is modelled by the zero-inflated binomial distribution. The dimension of the latent embedding of the variational autoencoder was chosen to be 10. The top 3,000 highly variable genes were selected using Scanpy (v1.9.1) for the model to learn the latent embeddings.[36] The model was trained on NVIDIA GeForce RTX 3090 addressing 24 GB RAM. Cell annotations for the integrated data sets were based on unsupervised clustering result and prior knowledge of marker gene expression of major cell types including B cells, T/NK cells, macrophages, monocytes, fibroblasts, endothelial cells and smooth muscle cells. The final representation of the data set was projected to 2-dimensional space using the UMAP algorithm.

To assess the effectiveness of scVAE for batch correction, we compared the data integration results of scVAE with Harmony.[47] The k-nearest-neighbor batch-effect test (kBET) metric was employed to determine the degree to which both methods eliminate batch effects while maintaining biological variance.[48]

Website development

We developed a user-friendly web interface with advanced functions to present our uniformly curated metadata and NGS data. The front-end interface was developed with HTML5 and CSS3 languages, based on the BootStrap (v5.2.1) toolkit. All font-end tables were built through DataTables (v1.12.1), a plug-in for the jQuery Javascript library. All data visualizations were developed by D3.js (v7), a JavaScript library for document manipulation. All back-end data including the bulk RNA-seq and microRNA-seq gene count matrix, UMAP coordinate information, the scRNA-seq clusters, cell subtype annotation, marker genes for different clusters were maintained into PostgreSQL database management system (v14.5). The MSdb database is deployed with a Nginx web server (v1.18.0) on an Ubuntu Linux (v20.04.5 LTS) operating system.