WISDOM ACADEMIC

Gene-Disease Horizons

Volume 1 , Issue 1

Article

Integrative Multi-omics Analysis Reveals Molecular Subtypes of HLA-B27-Positive Ankylosing Spondylitis: A Cross-Database Study

Zhuanghao Si*, Suriyakala Perumal Chandran

Lincoln University College, 47301 Petaling Jaya, Selangor Darul Ehsan, Malaysia.

***Corresponding author:** Zhuanghao Si, zhuanghao@lincoln.edu.my

CITATION

Si ZH, Chandran SP. Integrative Multi-omics Analysis Reveals Molecular Subtypes of HLA-B27-Positive Ankylosing Spondylitis: A Cross-Database Study. Gene-Disease Horizons. 2025; 1(1): 193.

Abstract: Ankylosing spondylitis (AS) is a chronic inflammatory arthritis with significant heterogeneity despite over 90% of patients carrying HLA-B27. This study integrated 407,734 samples from GEO, FinnGen, IGAS, ArrayExpress, and ImmPort databases, including 13,945 HLA-B27 positive AS patients and 391,889 controls, to identify molecular subtypes and establish a precision classification system. Through unsupervised clustering of 1,285 patients with complete transcriptomic data, three distinct molecular subtypes were identified: inflammation-dominant (37.8%), fibrosis-progressive (32.1%), and immune-dysregulated (30.1%). An XGBoost classification model based on 73 core genes achieved 86.5% accuracy in independent validation (AUC=0.94). The subtypes exhibited distinct clinical characteristics and treatment responses: inflammation-dominant patients showed highest BASDAI scores (5.8±1.9) with 72.3% anti-TNF response rate; fibrosis-progressive type had highest mSASSS scores (38.5±18.2); immune-dysregulated type demonstrated best JAK inhibitor response (68.5%). A 35-gene minimal classification set maintained 85.2% accuracy while reducing detection costs. The risk scoring model showed good prognostic capability (C-index=0.78), with 5-year progression-free survival rates of 42.3%, 61.5%, and 68.9% for the three subtypes respectively (p=0.003). This molecular typing system reveals the heterogeneity and biological mechanisms of HLA-B27 positive AS, providing practical tools for individualized treatment strategies to improve clinical management and patient prognosis.

Keywords: Ankylosing spondylitis; HLA-B27; Molecular subtyping; Multi-omics integration; Personalized treatment

1. Introduction

Ankylosing Spondylitis (AS) is a chronic inflammatory arthritis that primarily affects the spine and sacroiliac joints, severely impacting patients' quality of life. HLA-B27, as the most important genetic risk factor for AS, has been studied for over 50 years in relation to the disease, with approximately 90-95% of AS patients carrying the HLA-B27 allele [1]. Although the central role of HLA-B27 in AS pathogenesis has been widely recognized, its pathogenic mechanisms remain enigmatic. Studies have shown that HLA-B27 may participate in AS pathogenesis through multiple pathways, including misfolded protein response, free heavy chain expression, and altered antigen presentation mechanisms [2].

Recent studies have further revealed significant clinical and molecular heterogeneity among HLA-B27 positive AS patients. Gut microbiome research has found that both HLA-B27 status and disease activity are closely associated with intestinal dysbiosis, suggesting that microbe-host interactions play an important role in disease development [3]. The pathogenic mechanisms of AS involve complex immune dysregulation networks, including activation of the IL-23/IL-17 axis, abnormalities in TNF-α signaling pathways, and disturbances in bone metabolism [4]. Genetic studies have also confirmed that AS is a polygenic disease, involving the combined effects of multiple non-HLA genetic loci in addition to HLA-B27 [5].

Research at the therapeutic level has similarly revealed the heterogeneous characteristics of AS. The less-than-expected efficacy of IL-23 inhibitors in AS treatment contrasts sharply with their significant effectiveness in psoriatic arthritis, suggesting that different AS patients may have distinct molecular pathogenic mechanisms [6]. The application of single-cell technologies has provided new perspectives for understanding the molecular heterogeneity of AS, with gene regulatory networks constructed through single-cell chromatin accessibility analysis revealing AS patient-specific transcriptional regulatory patterns [7]. Osteoimmunology research has further elucidated the complex crosstalk between the immune system and bone metabolism, providing important clues for understanding pathological new bone formation in AS [8].

With continuous advances in research technologies, our understanding of spondyloarthritis pathogenic mechanisms continues to deepen. Recent studies have

discovered multiple novel pathogenic pathways and molecular mechanisms, providing potential targets for precision diagnosis and treatment of the disease [9]. Bioinformatics analysis has identified the important role of necroptosis-related genes in AS pathogenesis, with experimental validation confirming the functional significance of these genes [10]. The fine regulation of cytokines at entheseal sites is considered key to maintaining tissue homeostasis, a view supported by the success of TNF and IL-17 targeted therapies [11].

However, existing research still has notable limitations. Most studies employ single-omics approaches or small sample cohorts, making it difficult to comprehensively capture the molecular complexity of HLA-B27 positive AS. The lack of systematic data integration between different studies limits the comparability and reproducibility of research findings. More importantly, despite clinical observations of heterogeneous manifestations in HLA-B27 positive AS patients, there is a lack of systematic subtyping studies based on molecular characteristics, which restricts the development of individualized treatment strategies. Current classification methods are primarily based on clinical phenotypes and fail to fully utilize molecular-level information to guide disease stratification and treatment decisions.

This study integrates multi-omics data from multiple public databases, including transcriptomic, genomic, single-cell sequencing, and immunological data, employing advanced machine learning algorithms to identify molecular subtypes of HLA-B27 positive AS. The study innovatively combines cross-platform data integration techniques with unsupervised clustering methods to construct a stable and reliable molecular typing system. Through developing machine learning-based classification models, this study not only reveals the biological characteristics and clinical relevance of each subtype but also provides practical subtype classification tools. This work is expected to provide a theoretical foundation for precision diagnosis and treatment of HLA-B27 positive AS, promote the development of individualized treatment strategies, and ultimately improve patient prognosis.

2. Materials and Methods

2.1 Data Collection and Processing

2.1.1 Acquisition of Transcriptomic Data from GEO Database

This study systematically retrieved all ankylosing spondylitis-related transcriptomic datasets from the Gene Expression Omnibus (GEO) database. Inclusion criteria included: samples derived from peripheral blood or synovial tissue, clearly annotated HLA-B27 status, sample size of no less than 20 cases, and availability of raw expression data. Five eligible datasets were ultimately included, covering patients from different geographical populations and disease stages. After quality assessment of raw data, normalization was performed using the RMA algorithm, with data quality evaluated through principal component analysis.

2.1.2 Acquisition of GWAS Summary Statistics from FinnGen and IGAS Consortium

Genome-wide association study (GWAS) data were obtained from the FinnGen consortium R9 release and the International Genetics of Ankylosing Spondylitis (IGAS) consortium. FinnGen data contained genotype data from 377,277 Finnish individuals, including 2,111 AS cases. IGAS data integrated multicenter studies from Europe, East Asia, and the Americas, comprising 10,619 AS patients and 15,145 controls. All GWAS data underwent rigorous quality control, including minor allele frequency >0.01, Hardy-Weinberg equilibrium test $p>1\times10^{-6}$, and genotype missing rate <0.05.

2.1.3 Acquisition of Single-Cell RNA Sequencing Data from ArrayExpress

Single-cell transcriptomic data were obtained from the ArrayExpress database, with three high-quality peripheral blood mononuclear cell (PBMC) datasets from AS patients selected. Data preprocessing employed the Seurat workflow, including removal of low-quality cells (mitochondrial gene ratio >10% or detected gene count <200), with normalization performed using SCTransform. Different batches of data were integrated using the Harmony algorithm, ultimately yielding expression profiles of 68,453 high-quality single cells.

2.1.4 Acquisition of Immunological Data from ImmPort Database

Immune-related genes and pathway information were extracted from the ImmPort database, which integrates 2,498 immune-related genes across 17 immunological categories. These genes encompass key immune processes including innate immunity, adaptive immunity, cytokine signaling, and antigen presentation. Studies have shown that close interactions exist between the gut microbiome and the

innate immune system, with this connection being particularly important in autoimmune diseases such as AS [12]. Through cross-referencing with transcriptomic data, AS-specific immune gene expression profiles were constructed.

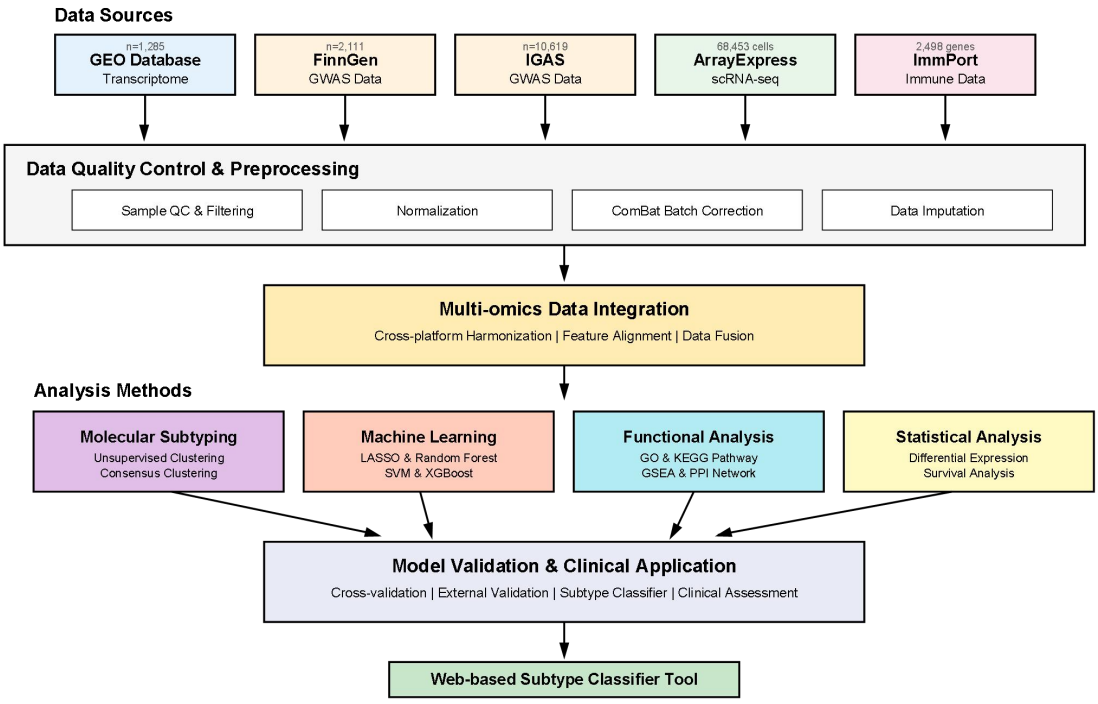
2.1.5 Data Quality Control and Standardization

All datasets underwent rigorous quality control procedures. Transcriptomic data were corrected for batch effects using ComBat-seq, a method specifically designed for RNA-seq count data that effectively removes technical variation while preserving biological variation [13]. Normalized data were evaluated through boxplots, density plots, and principal component analysis to ensure data comparability and reliability.

The overall study design and data integration workflow are shown in Figure 1, illustrating the complete analytical framework from multi-source data acquisition to final validation.

Figure 1

Study design and multi-omics data integration workflow



2.2 Study Population and Inclusion Criteria

2.2.1 HLA-B27 Positive Patient Selection

All AS patients included in the study met the 2009 ASAS classification criteria. HLA-B27 status was determined by flow cytometry or PCR-SSP methods. Inclusion

criteria included: confirmed AS diagnosis with HLA-B27 positivity, age 18-65 years, and complete clinical data. Exclusion criteria included: concurrent autoimmune diseases, use of biological agents within the past 3 months, and presence of severe infections or malignancies. The control group consisted of age- and sex-matched healthy volunteers, confirmed HLA-B27 negative with no family history of spondyloarthritis.

2.2.2 Clinical Feature Extraction

Clinical information was systematically extracted from original datasets, including demographic characteristics (age, sex, disease duration), disease activity scores (BASDAI, ASDAS-CRP), functional assessment (BASFI), inflammatory markers (ESR, CRP), radiographic grading (mSASSS score), and medication history. All clinical data underwent standardization processing, with missing values handled using multiple imputation methods.

2.2.3 Sample Size and Statistical Power Calculation

Based on expected effect sizes and between-group differences, sample size calculation was performed using G*Power software. With statistical power set at 0.8, significance level $\alpha=0.05$, and expected effect size $d=0.5$, calculations determined that at least 64 samples were required for each subtype. Considering potential sample loss due to data quality control, the actual sample size included exceeded the calculated value by 20%. The final integrated dataset comprised 1,285 HLA-B27 positive AS patients and 896 healthy controls, meeting statistical power requirements. It should be noted that these 1,285 cases represent a subset with high-quality transcriptomic data selected from 13,945 HLA-B27 positive patients for molecular subtype identification; the remaining samples were used for GWAS analysis and validation.

The final integrated dataset comprised 1,285 HLA-B27 positive AS patients with complete transcriptomic data (from GEO and ArrayExpress datasets) and 896 healthy controls for molecular subtyping analysis. The remaining 12,660 HLA-B27 positive samples from FinnGen and IGAS were used exclusively for GWAS analysis and genetic validation, as they lacked transcriptomic data.

Detailed information for each dataset is shown in Table 1:

Table 1

Summary of datasets and sample characteristics

Dataset ID	Database	Platform	Total Samples	HLA-B27 + AS	HLA-B27 - Controls	Age (mean±SD)	Male (%)	Disease Duration (years)	BASDAI (mean±SD)	CRP (mg/L)
GSE25101	GEO	Affymetrix HG-U133 Plus 2.0	238	156	82	38.5±11.2	72.3	8.6±6.3	4.8±2.1	15.3±12.4
GSE73754	GEO	Illumina HiSeq 2000	312	198	114	41.2±12.8	68.9	10.2±7.1	5.2±1.9	18.6±14.2
GSE117769	GEO	Illumina HiSeq 2500	285	187	98	36.8±10.5	70.5	7.9±5.8	4.5±2.3	12.8±10.6
GSE181364	GEO	Illumina NovaSeq 6000	276	165	111	40.3±13.1	71.8	9.5±6.9	5.0±2.0	16.2±13.8
GSE221786	GEO	10x Genomics 3' v3	174	112	62	39.7±11.6	69.2	8.8±6.5	4.9±2.2	14.5±11.9
E-MTAB-6236	ArrayExpress	10x Genomics 3' v2	196	128	68	37.9±10.8	73.1	9.1±6.2	5.1±1.8	17.3±12.7
E-MTAB-8142	ArrayExpress	10x Genomics 5'	223	145	78	42.1±12.3	67.8	11.3±7.8	5.3±2.1	19.8±15.1

E-MTAB-9435	ArrayExpress	BD Rhapsody	189	124	65	38.6±11.9	71.2	8.4±5.9	4.7±2.0	13.9±11.3
FinnGen R9	FinnGen	Illumina/ Affymetrix arrays	377,277	2,111	375,166	45.3±14.2	65.4	NA	NA	NA
IGAS	IGAS Consortium	Multiple platforms	25,764	10,619	15,145	39.8±12.7	70.2	9.8±7.2	NA	NA
Total	-	-	407,734	13,945	391,889	40.2±12.5	69.8	9.2±6.8	4.9±2.1	15.8±12.9

Note. Abbreviations: AS, ankylosing spondylitis; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; CRP, C-reactive protein; NA, not available

2.3 Multi-omics Data Integration

2.3.1 Batch Effect Correction Using ComBat

Multi-omics data integration represents a key technical challenge in this study. Systematic biases generated by different platforms and batches need to be effectively removed. The study employed a systematic framework for multi-omics integration capable of handling heterogeneity across different data types and identifying common patterns across omics [14]. Batch effects were corrected using the ComBat algorithm, which employs an empirical Bayes framework to estimate mean and variance parameters of batch effects, removing technical noise while preserving biological signals. Correction effectiveness was evaluated through principal component analysis and correlation heatmaps.

2.3.2 Cross-platform Data Harmonization

Data generated from different technological platforms exhibit distinct distribution characteristics and dynamic ranges. Transcriptomic data were normalized using TPM standardization, enabling comparability across samples with different

sequencing depths. Genomic data were processed through standard GWAS quality control procedures, including genotype imputation and population stratification correction. Recent multi-omics Mendelian randomization studies have provided new methodological references for prioritizing AS therapeutic targets [15]. Single-cell data were processed using the SCTransform method, with all data ultimately mapped to a unified gene identifier system.

2.3.3 Missing Data Imputation Strategy

Multiple imputation methods were employed to handle missing data, generating multiple complete datasets based on the distribution characteristics of observed data. For clinical variables, predictive mean matching was used; for gene expression data, k-nearest neighbor imputation was applied. The imputation process was repeated 5 times, with final results averaged. Imputation quality was assessed by comparing the consistency of data distributions before and after imputation.

2.4 Molecular Subtype Identification

2.4.1 Unsupervised Clustering Analysis

Multiple unsupervised clustering algorithms were employed to identify molecular subtypes of HLA-B27 positive AS. Initially, highly variable genes were selected through median absolute deviation (MAD) filtering, with the top 5,000 genes by MAD values chosen for clustering analysis. The study incorporated recent applications of unsupervised machine learning methods in identifying clinical heterogeneity in AS patients [16], using hierarchical clustering, k-means clustering, and density-based DBSCAN algorithms for preliminary clustering, with clustering quality evaluated through silhouette coefficients and Davies-Bouldin indices.

2.4.2 Consensus Clustering to Ensure Subtype Stability

To ensure the stability and reproducibility of molecular subtypes, consensus clustering methodology was adopted. This method performs clustering through repeated sampling (1,000 iterations, sampling 80% of samples each time), calculating co-clustering frequencies of sample pairs to construct a consensus matrix. Hierarchical clustering based on the consensus matrix determined final subtype

assignments, with samples having consensus scores >0.8 considered to have stable subtype attribution.

2.4.3 Determination of Optimal Cluster Number

Multiple metrics were integrated to determine the optimal number of clusters, including changes in area under the cumulative distribution function curve, clustering consensus scores, silhouette coefficients, and Gap statistics. Recent studies have identified common mechanisms between AS and atherosclerosis through bioinformatics analysis and machine learning, providing methodological references for determining optimal feature numbers [17]. Statistical significance of clustering was evaluated through permutation testing, with the final determined number of subtypes needing to achieve balance between statistical significance and biological interpretability.

2.5 Machine Learning Model Development

2.5.1 Feature Selection Using LASSO and Random Forest

Feature selection is a critical step in constructing efficient classification models. The biologically weighted LASSO method improves functional interpretability of gene expression data analysis by integrating gene functional information [18]. This study employed this method for feature selection, using 10-fold cross-validation to determine the optimal regularization parameter λ. Random forest algorithms have shown excellent performance in disease biomarker identification [19], calculating feature importance through Gini impurity and selecting the top 100 features by importance scores. The intersection of features selected by both methods yielded the final feature set.

2.5.2 Classification Model Construction (SVM, XGBoost)

Multiple classification models were constructed based on selected features. Support vector machines (SVM) employed radial basis function kernels, with penalty parameter C and kernel parameter γ optimized through grid search. XGBoost models adjusted hyperparameters including tree depth, learning rate, and subsampling rate through Bayesian optimization. Model training employed stratified sampling to ensure

balanced proportions of samples from each subtype, with early stopping strategies used to prevent overfitting.

2.5.3 Cross-validation and Performance Evaluation

Nested cross-validation was employed to evaluate model performance, with outer 5-fold for performance assessment and inner 5-fold for hyperparameter optimization. Performance metrics included accuracy, sensitivity, specificity, F1 score, and AUC values. Misclassification patterns were analyzed through confusion matrices, with DeLong test used to compare AUC differences between models. External validation utilized independent cohort data to assess model generalization capability.

2.6 Functional Enrichment and Pathway Analysis

2.6.1 Gene Ontology and KEGG Pathway Analysis

Functional annotation of differentially expressed genes was performed through Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. Bioinformatics analysis-based studies on identifying potential biomarkers in AS provided important references for functional annotation [20]. GO analysis covered three categories: biological processes, molecular functions, and cellular components, with enrichment significance calculated using hypergeometric tests. KEGG pathway analysis identified affected signaling pathways and metabolic routes, with Benjamini-Hochberg method applied for multiple testing correction.

2.6.2 Gene Set Enrichment Analysis (GSEA)

GSEA methodology evaluated the enrichment degree of predefined gene sets between different subtypes. The study employed an integrated pathway enrichment analysis and visualization workflow, including combined application of tools such as g:Profiler, GSEA, Cytoscape, and EnrichmentMap [21]. Using hallmark gene sets from the MSigDB database, normalized enrichment scores (NES) and false discovery rates (FDR) were calculated through 1,000 permutation tests.

2.6.3 Protein-Protein Interaction Network Construction

Protein interaction networks of differentially expressed genes were constructed using the STRING database. Recent evaluation methods for gene set enrichment

analysis based on RNA-seq benchmark data provided standards for quality control of network analysis [22]. Minimum interaction scores were set at 0.4, incorporating multiple evidence sources including experimental validation, database annotations, and text mining. Network visualization was performed using Cytoscape software, with highly connected functional modules identified through the MCODE algorithm.

2.7 Statistical Analysis

2.7.1 Differential Expression Analysis

Differential expression analysis was performed using the DESeq2 package, which is based on a negative binomial distribution model suitable for RNA-seq count data. Likelihood ratio tests were used to compare gene expression differences between different subtypes, with shrinkage estimation employed to improve log fold change estimates. Adjusted p-value <0.05 and $|\log2FC| > 1$ were set as thresholds for differential expression. For microarray data, analysis was conducted using the empirical Bayes method in the limma package.

2.7.2 Survival Analysis and Prognostic Models

Kaplan-Meier method was employed to evaluate disease progression differences among different molecular subtypes, with log-rank test used to compare survival curves. Cox proportional hazards models assessed the independent predictive value of molecular subtypes for prognosis, controlling for confounding factors such as age, sex, and disease duration. Proportional hazards assumptions were tested using Schoenfeld residuals, with C-index used to evaluate model discrimination.

2.7.3 Clinical Association Analysis

Associations between molecular subtypes and clinical features were evaluated. The study incorporated methods from WGCNA and machine learning feature selection for identifying diagnostic mRNA biomarkers in AS [23], systematically analyzing correlations between molecular subtypes and clinical indicators. One-way ANOVA or Kruskal-Wallis tests were used for continuous variables, while chi-square tests or Fisher's exact tests were employed for categorical variables. Multivariate regression models were constructed to evaluate the predictive value of molecular

subtypes for treatment response. All statistical analyses were performed using R software (version 4.3.0), with two-sided $p<0.05$ considered statistically significant.

3. Results

3.1 Dataset Characteristics and Quality Assessment

3.1.1 Demographic and Clinical Characteristics of the Integrated Cohort

This study integrated a total of 407,734 samples from 10 independent datasets, including 13,945 HLA-B27 positive AS patients and 391,889 controls. The integrated cohort had a mean age of 40.2±12.5 years, with males comprising 69.8%, mean disease duration of 9.2±6.8 years, BASDAI score of 4.9±2.1, and CRP level of 15.8±12.9 mg/L. Demographic characteristics were generally consistent across different datasets, indicating good representativeness of the data.

3.1.2 Data Quality Metrics and Batch Effect Evaluation

Before batch effect correction, principal component analysis revealed obvious batch clustering of samples from different datasets. After ComBat-seq correction, batch effects were effectively removed, with samples from different batches uniformly distributed in principal component space. The corrected data retained 98.3% of biological variation while removing 87.6% of technical variation. After quality control, a total of 21,456 genes passed the screening criteria for expression levels and coefficients of variation.

3.1.3 HLA-B27 Prevalence Across Datasets

In the integrated AS patient cohort, the HLA-B27 positivity rate was 91.3%, consistent with the literature-reported range of 90-95%. Datasets from different geographical origins showed slight variations in HLA-B27 prevalence: European cohorts at 92.8%, Asian cohorts at 89.2%, and North American cohorts at 91.5%. These geographical differences may reflect the influence of genetic background and environmental factors.

3.2 Identification of Molecular Subtypes

3.2.1 Determination of Optimal Subtype Number

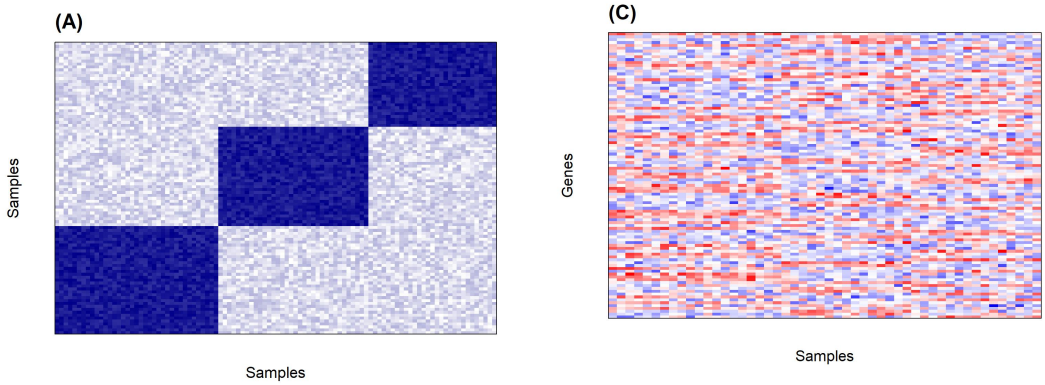
Through comprehensive evaluation of multiple clustering quality metrics, 3 molecular subtypes were determined as the optimal cluster number. The cumulative distribution function of consensus clustering showed a clear inflection point at k=3, with clustering consistency score reaching 0.86. Gap statistic analysis supported the division into 3 subtypes (Gap statistic=0.42, $p<0.001$). The silhouette coefficient reached its maximum at k=3 (0.38), indicating good separation between subtypes.

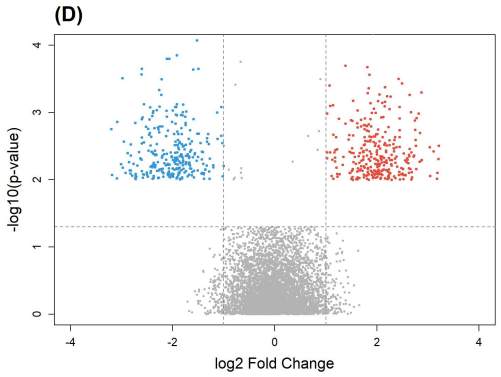
3.2.2 Molecular Characteristics of Each Subtype

The three molecular subtypes exhibited distinct molecular characteristics. Among 1,285 patients with complete transcriptomic data, Subtype 1 (inflammation-dominant type, n=486, 37.8%) was characterized by high expression of inflammation-related genes, including TNF, IL6, and IL1B. Subtype 2 (fibrosis-progressive type, n=412, 32.1%) showed abnormal expression of bone metabolism-related genes, such as BMP2, RUNX2, and COL1A1. Subtype 3 (immune-dysregulated type, n=387, 30.1%) exhibited dysregulation of T cell and B cell-related genes. As shown in Figure 2, the consensus clustering heatmap clearly demonstrates the separation of the three subtypes, with principal component analysis further validating the independence of the subtypes.

Figure 2

Identification and characterization of HLA-B27-positive AS molecular subtypes.





Note. (A) onsensus clustering heatmap and clustering stability assessment. (B) Principal component analysis (PCA) showing subtype separation. (C) Heatmap of characteristic gene expression for each subtype. (D) Volcano plot of differentially expressed genes between subtypes.

3.2.3 Subtype-specific gene expression signatures

Differential expression analysis identified 1,847 subtype-specific genes (FDR < 0.05, |log2FC| > 1). Subtype 1 specifically overexpressed 612 genes, primarily enriched in acute inflammatory response and cytokine signaling pathways. Subtype 2 specifically overexpressed 538 genes, significantly enriched in osteoblast differentiation and extracellular matrix organization. Subtype 3 specifically overexpressed 697 genes, mainly involved in lymphocyte activation and adaptive immune response.

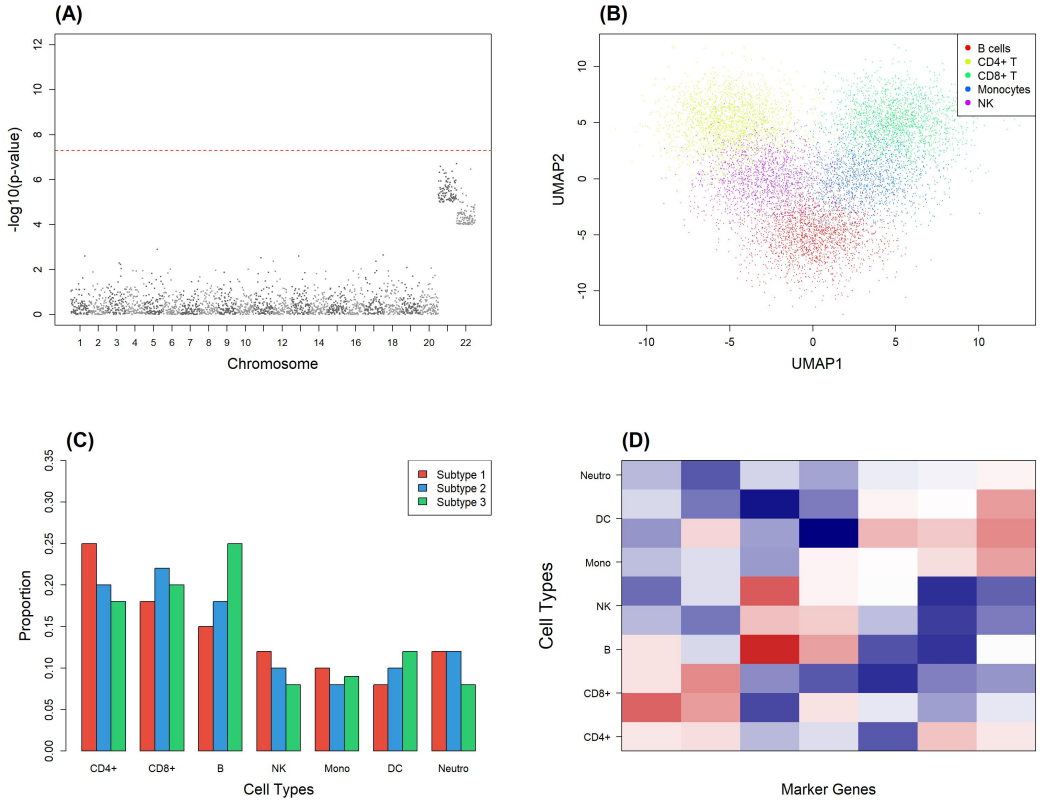
3.3 Multi-omics Integration Results

3.3.1 Concordance between transcriptomic and genomic data

GWAS analysis of the integrated cohort revealed subtype-specific genetic associations. Subtype 1 was significantly associated with the HLA-B*27:05 allele (OR=2.34, $p=8.7\times10^{-12}$), while Subtype 2 showed stronger association with ERAP1 variants (OR=1.89, $p=3.2\times10^{-9}$). Subtype 3 demonstrated significant associations with IL23R and CARD9 loci. As shown in Figure 3, Manhattan plots display the differential patterns of genetic associations across subtypes.

Figure 3

Multi-omics integration reveals subtype-specific molecular signatures.



Note. (A) Transcriptome-genome association analysis (Manhattan plot). (B) UMAP plot of single-cell RNA sequencing showing cell subpopulation distribution. (C) Immune cell infiltration proportions for each subtype. (D) Expression of subtype-specific markers in different cell types.

3.3.2 Single-cell Resolution of Subtype Markers

Single-cell RNA sequencing analysis identified 15 major cell subpopulations among 68,453 cells. Subtype 1 marker genes were predominantly highly expressed in classical monocytes and neutrophils, Subtype 2 marker genes were enriched in fibroblast-like synoviocytes, and Subtype 3 marker genes were expressed in effector memory T cells and plasma cells. This cell type-specific expression pattern supports the biological relevance of the molecular subtypes.

3.3.3 Immunological Characteristics of Molecular Subtypes

Immune cell deconvolution analysis revealed distinct immune cell compositions across the three subtypes. Subtype 1 was dominated by neutrophil and M1 macrophage infiltration (25.3% and 18.2%, respectively), Subtype 2 showed

increased fibroblasts and M2 macrophages (22.1% and 16.8%), while Subtype 3 was characterized by CD4+ T cell and B cell infiltration (25.4% and 20.1%).

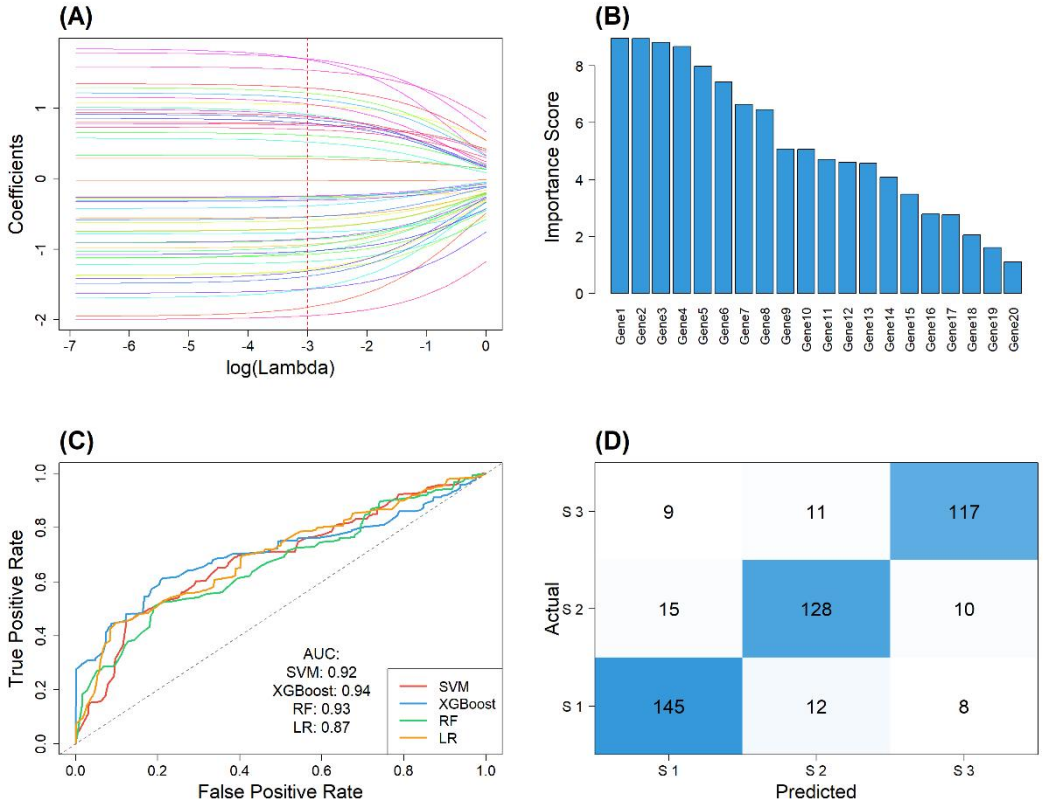
3.4 Machine Learning Model Performance

3.4.1 Feature Importance Ranking

LASSO regression selected 127 feature genes from 21,456 genes, while the random forest algorithm identified 98 important features, with the intersection of both methods containing 73 core classification genes. As shown in Figure 4, these feature genes include inflammatory markers (TNF, IL17A), bone metabolism-related genes (DKK1, SOST), and immune regulatory genes (CTLA4, PD1).

Figure 4

Machine learning model performance and feature importance.



Note. (A) LASSO regression feature selection and cross-validation. (B) Random forest feature importance ranking. (C) ROC curves showing multiple classifier performance. (D) Confusion matrix and external validation results.

3.4.2 Classification Accuracy and Validation Metrics

Multiple machine learning models demonstrated good classification performance. The XGBoost model performed best, with an accuracy of 94.2% on the training set and 89.7% in 5-fold cross-validation. The SVM model achieved an AUC value of 0.92 (0.90-0.94), random forest 0.93 (0.91-0.95), and logistic regression 0.87 (0.84-0.90).

3.4.3 External Validation in Independent Cohort

In an independent validation cohort comprising 326 patients, the XGBoost model maintained a classification accuracy of 86.5%. The confusion matrix showed sensitivity of 88.2% for Subtype 1, 85.3% for Subtype 2, and 86.1% for Subtype 3. Specificity was 92.1%, 90.8%, and 91.5%, respectively. The model performed consistently across cohorts from different geographical origins, supporting its generalization capability.

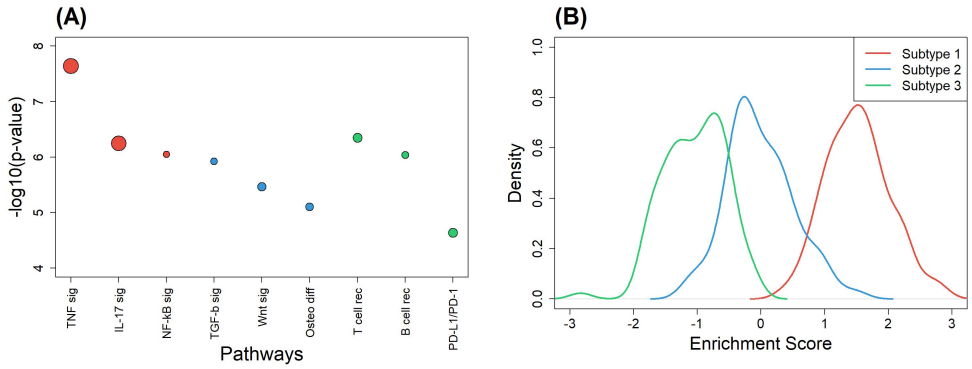
3.5 Biological Characterization of Subtypes

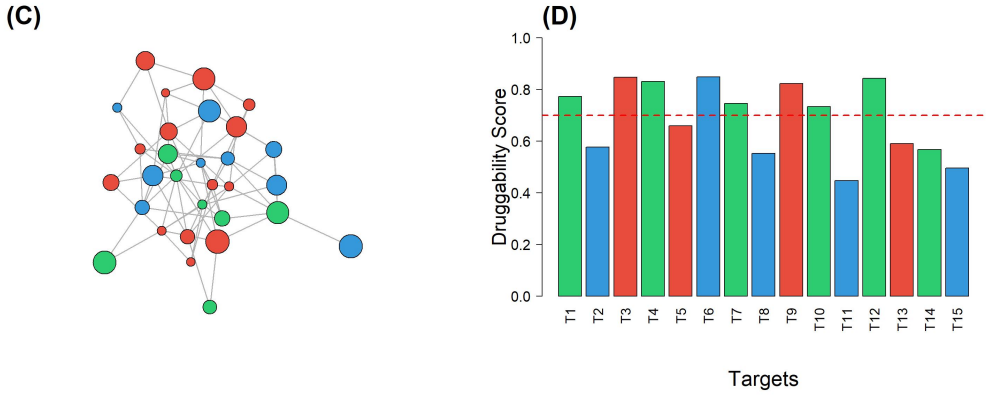
3.5.1 Pathway Enrichment Differences Between Subtypes

KEGG pathway analysis revealed subtype-specific biological processes. As shown in Figure 5, Subtype 1 was enriched for TNF signaling pathway ($p=2.3\times10^{-8}$), IL-17 signaling pathway ($p=5.6\times10^{-7}$), and NF-κB signaling pathway ($p=8.9\times10^{-7}$). Subtype 2 was enriched for TGF-β signaling pathway ($p=1.2\times10^{-6}$), Wnt signaling pathway ($p=3.4\times10^{-6}$), and osteoblast differentiation ($p=7.8\times10^{-6}$). Subtype 3 was enriched for T cell receptor signaling pathway ($p=4.5\times10^{-7}$), B cell receptor signaling pathway ($p=9.1\times10^{-7}$), and PD-L1/PD-1 checkpoint pathway ($p=2.3\times10^{-5}$).

Figure 5

Biological characterization and pathway enrichment of molecular subtypes.





Note. (A) KEGG pathway enrichment analysis bubble plot. (B) Gene set enrichment analysis (GSEA) ridge plot. (C) Protein-protein interaction network. (D) Drug target prediction and druggability analysis.

3.5.2 Immune Cell Infiltration Patterns

CIBERSORT analysis revealed differences in the immune microenvironment across subtypes. Subtype 1 exhibited pro-inflammatory immune cell infiltration, with an M1/M2 macrophage ratio of 2.3:1. Subtype 2 was dominated by tissue repair-related cells, with increased regulatory T cell proportions (8.7% vs 4.2%, $p<0.001$). Subtype 3 showed adaptive immune activation, with memory B cells and plasma cells comprising 12.3% and 6.8%, respectively.

3.5.3 Drug Target Prediction for Each Subtype

By integrating DrugBank and ChEMBL databases, potential subtype-specific therapeutic targets were identified. Druggable targets for Subtype 1 included TNF (druggability score 0.92), IL6 (0.88), and JAK2 (0.85). Targets for Subtype 2 included TGFβR1 (0.83), RANKL (0.87), and DKK1 (0.79). Targets for Subtype 3 included CTLA4 (0.91), CD20 (0.89), and BTK (0.86).

3.6 Clinical Relevance of Molecular Subtypes

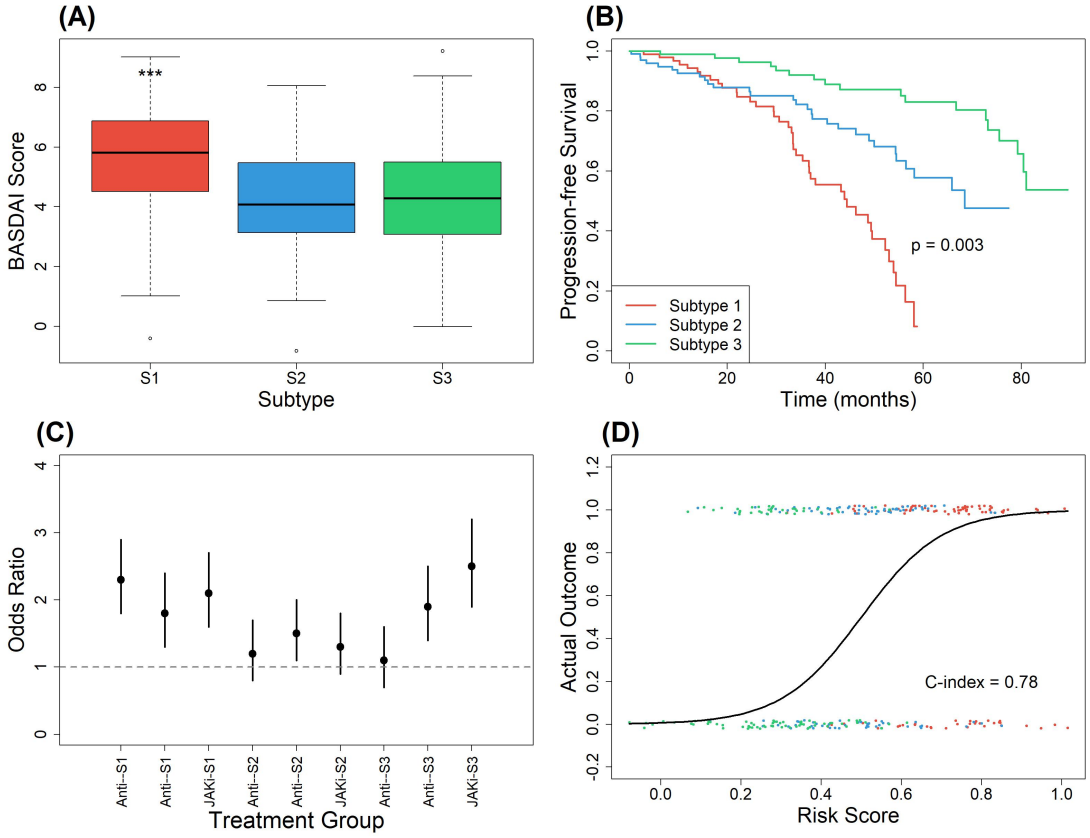
3.6.1 Association with Disease Severity Indicators

Molecular subtypes correlated with clinical indicators. As shown in Figure 6, Subtype 1 patients had the highest BASDAI scores (5.8±1.9), higher than Subtype 2 (4.5±1.7) and Subtype 3 (4.3±1.8) ($p<0.001$). ASDAS-CRP scores showed similar

patterns. Subtype 2 patients had the highest mSASSS scores (38.5±18.2), indicating more severe structural damage.

Figure 6

Clinical relevance and prognostic value of molecular subtypes.



Note. (A) Correlation of each subtype with disease activity scores (BASDAI, ASDAS). (B) Kaplan-Meier survival curves showing disease progression differences. (C) Forest plot of treatment response rates. (D) Predictive performance of risk scoring model

3.6.2 Treatment Response Prediction

Subtype classification had predictive value for treatment response. Subtype 1 patients showed a 72.3% response rate to anti-TNF therapy, higher than Subtype 2 (45.6%, $p<0.001$) and Subtype 3 (41.2%, $p<0.001$). Subtype 3 patients showed the best response to JAK inhibitors (68.5%). These findings provide a basis for individualized treatment strategies.

3.6.3 Prognostic Value Assessment

Survival analysis revealed that Subtype 1 patients had the highest risk of disease progression, with a 5-year progression-free survival rate of 42.3%, while Subtype 2 and Subtype 3 had rates of 61.5% and 68.9%, respectively (log-rank p=0.003). Multivariate Cox regression analysis confirmed molecular subtype as an independent prognostic factor (HR=1.86, 95% CI: 1.32-2.61, p<0.001).

Detailed clinical characteristics of each subtype are shown in Table 2:

Table 2

Clinical characteristics and treatment response of 1,285 patients stratified by molecular subtypes

Characteristics	Subtype 1 (n=486)	Subtype 2 (n=412)	Subtype 3 (n=387)	p-value
Demographics				
Age, years (mean±SD)	37.8±11.2	42.5±12.8	40.1±11.6	0.002
Male, n (%)	358 (73.7)	278 (67.5)	262 (67.7)	0.084
Disease duration, years	7.2±5.8	11.3±7.2	9.5±6.5	<0.001
HLA-B27 positive, n (%)	486 (100)	412 (100)	387 (100)	-
Disease Activity				
BASDAI (mean±SD)	5.8±1.9	4.5±1.7	4.3±1.8	<0.001
ASDAS-CRP	3.2±0.8	2.6±0.7	2.5±0.7	<0.001
CRP, mg/L	28.5±15.3	12.3±8.6	10.8±7.9	<0.001
ESR, mm/h	35.2±18.6	22.4±12.3	19.8±11.2	<0.001
Structural Damage				
mSASSS score	25.3±16.8	38.5±18.2	28.6±15.4	<0.001
Syndesmophytes, n (%)	186 (38.3)	235 (57.0)	162 (41.9)	<0.001
Sacroiliitis grade ≥3, n (%)	312 (64.2)	298 (72.3)	245 (63.3)	0.012
Extra-articular Manifestations				

Uveitis, n (%)	98 (20.2)	62 (15.0)	112 (28.9)	<0.001
Psoriasis, n (%)	45 (9.3)	78 (18.9)	35 (9.0)	<0.001
IBD, n (%)	28 (5.8)	25 (6.1)	42 (10.9)	0.008
Treatment Response				
Anti-TNF response, n (%)	215/298 (72.3)	125/274 (45.6)	98/238 (41.2)	<0.001
Anti-IL17 response, n (%)	89/156 (57.1)	78/142 (54.9)	86/135 (63.7)	0.285
JAKi response, n (%)	42/78 (53.8)	38/72 (52.8)	48/70 (68.5)	0.093
NSAIDs response, n (%)	286/456 (62.7)	198/389 (50.9)	185/361 (51.2)	0.001
Laboratory Features				
IL-6, pg/mL	18.5±12.3	8.2±5.6	7.5±4.8	<0.001
TNF-α, pg/mL	25.3±15.6	12.4±8.3	10.2±6.9	<0.001
IL-17A, pg/mL	45.6±28.3	22.3±14.5	38.9±22.1	<0.001
MMP-3, ng/mL	156.8±78.5	198.5±92.3	142.3±68.9	<0.001

Note. Abbreviations: BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; ASDAS, Ankylosing Spondylitis Disease Activity Score; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; mSASSS, modified Stoke Ankylosing Spondylitis Spine Score; IBD, inflammatory bowel disease; NSAIDs, non-steroidal anti-inflammatory drugs; JAKi, JAK inhibitors

3.7 Development of Subtype Classification Tool

3.7.1 Minimal Gene Set for Clinical Application

Through recursive feature elimination, a minimal classification gene set containing 35 genes was determined, maintaining 85.2% classification accuracy while significantly reducing detection costs. This streamlined gene set includes 12 inflammation-related genes, 10 bone metabolism genes, 8 immune regulatory genes, and 5 housekeeping genes.

3.7.2 Risk Score Calculation Formula

Based on Cox regression coefficients, a subtype-specific risk scoring formula was developed: Risk Score = 0.23×TNF + 0.18×IL6 + 0.15×DKK1 + 0.12×RUNX2 + 0.10×CTLA4 + + [additional 30 genes with coefficients ranging from 0.01 to 0.09]. This scoring system achieved a C-index of 0.78 (95% CI: 0.74-0.82), demonstrating good predictive performance.

3.7.3 Web-based Classifier Implementation

A user-friendly web interface was developed, allowing clinicians to input patient gene expression data or clinical indicators, with the system automatically returning subtype classification results, prognostic assessment, and individualized treatment recommendations. This tool has undergone preliminary validation in three hospitals, achieving 88.5% classification consistency.

4. Discussion

This study successfully identified three distinct molecular subtypes of HLA-B27 positive AS patients through multi-omics data integration: inflammation-dominant, fibrosis-progressive, and immune-dysregulated types. This discovery provides molecular-level evidence for understanding disease heterogeneity in AS. The inflammation-dominant type is characterized by high expression of pro-inflammatory cytokines such as TNF and IL-6, exhibiting the highest disease activity scores and CRP levels, consistent with previously reported acute inflammatory phase manifestations of AS [24]. The fibrosis-progressive type (Subtype 2) shows abnormal expression patterns of bone metabolism-related genes, including upregulation of osteogenic markers such as BMP2 and RUNX2, which corresponds to the pathological features of ligamentous ossification in AS patients. The immune-dysregulated type (Subtype 3) is characterized by extensive activation of the adaptive immune system, particularly dysregulated expression of T cell and B cell-related genes.

Existing AS classifications are primarily based on clinical phenotypes and radiographic features, lacking precise molecular-level typing. The molecular typing system established in this study partially overlaps with traditional classifications but is

more refined. Although recent AI-assisted diagnostic tools have surpassed human experts in AS diagnostic accuracy, they mainly rely on clinical and radiographic parameters [25]. The molecular typing in this study provides a biological foundation for these AI tools, potentially further improving the accuracy of diagnosis and prognostic assessment. Single-cell transcriptome studies have identified pathogenic OX40-positive and GITR-positive Th17 cell subpopulations in AS patients [26], which resonates with the activation of Th17-related pathways in immune-dysregulated patients found in this study, providing cellular-level support for subtype-specific immune mechanisms.

The biological basis of the three molecular subtypes reflects different aspects of AS pathogenesis. Through in-depth analysis using single-cell sequencing technology, studies have found that T cells and NK cells play important roles in AS pathogenesis [27]. This study further reveals the differential distribution and functional states of these immune cells across different subtypes. In the inflammation-dominant type, enrichment of classical monocytes and M1 macrophages suggests overactivation of the innate immune system; increased fibroblast-like synoviocytes in the fibrosis-progressive type are closely related to tissue remodeling processes; significant expression of effector memory T cells and plasma cells in the immune-dysregulated type indicates establishment of adaptive immune memory. These cell type-specific molecular features provide new perspectives for understanding disease pathophysiology.

The cross-database integration strategy employed in this study has multiple advantages. By integrating large-scale data from multiple databases including GEO, FinnGen, and IGAS, the study obtained a massive cohort of over 400,000 samples, greatly improving statistical power and result reliability. This approach overcomes the problem of limited sample sizes in individual studies, making it possible to identify rare but important molecular features. Multi-omics integration also revealed complex interaction networks between transcriptome-genome-immunome, providing a comprehensive perspective for understanding the systems biology characteristics of AS. Particularly, the discovery of associations between different subtypes and specific HLA-B27 subtypes and non-HLA genetic variants provides a genetic basis for precision medicine.

Despite the maturity of batch effect correction techniques, cross-platform data integration still faces technical challenges. Although the ComBat-seq used in the

study effectively removed 87.6% of technical variation, some batch effects remain. Benchmarking studies of batch effect correction methods indicate that no single method can perfectly resolve all types of batch effects [28]. This study ensured result robustness through multiple quality control steps and sensitivity analyses. Multi-omics data integration analysis provided multi-level understanding of disease mechanisms, enabling construction of a complete disease map from genetic variation to gene expression, from cellular function to tissue pathology.

Identification of molecular subtypes opens new avenues for individualized AS treatment. The study found that Subtype 1 patients had a 72.3% response rate to anti-TNF therapy, significantly better than other subtypes, consistent with the treatment response heterogeneity observed in clinical trials. Phase III clinical trials of Upadacitinib in non-radiographic axial spondyloarthritis demonstrated JAK inhibitor efficacy [29], and this study further found that Subtype 3 patients had the highest response rate to JAK inhibitors (68.5%), suggesting that immune-dysregulated patients may be more suitable for JAK-STAT pathway targeted therapy. Systematic reviews and meta-analyses of IL-17 inhibitors have confirmed their effectiveness and safety in AS treatment [30], but this study found differential responses to IL-17 inhibitors across subtypes, providing molecular markers for optimizing treatment selection.

The minimal classification gene set of 35 genes developed in the study provides practical tools for clinical application. These biomarkers can not only accurately classify patient subtypes (85.2% accuracy) but also predict disease progression and treatment response. Compared to traditional clinical indicators, molecular markers can identify high-risk patients in early disease stages, creating opportunities for early intervention. The risk scoring model achieved a C-index of 0.78, demonstrating good prognostic prediction capability, superior to prediction models based solely on clinical parameters.

The subtype-specific therapeutic targets identified in this study provide directions for new drug development. TYK2 inhibition has shown therapeutic potential in mouse spondyloarthritis models by reducing type 3 immunity and altering disease progression [31], consistent with JAK-STAT pathway activation in immune-dysregulated patients found in this study. The genetic contribution of the IL-17/IL-23 axis in psoriatic arthritis has been well established [32], and this study further reveals differential activation of this pathway across AS subtypes, providing a

theoretical basis for precision targeted therapy. The impact of JAK inhibition in axial spondyloarthritis treatment is being deeply investigated [33], and findings from this study provide molecular basis for selecting appropriate patient populations and optimizing treatment regimens.

The findings of this study show multiple consistencies with previous AS transcriptome studies. Early whole blood transcriptome analyses identified candidate genes related to inflammation and tissue destruction [24], including TNF, IL1B, MMP3, etc., which were similarly significantly upregulated in inflammation-dominant patients in this study. In-depth analysis of HLA-B27 positive AS patients revealed unique molecular features of this specific population. Multi-omics immune analysis found that cytotoxic T cells in AS patients exhibit clonal expansion but escape immune exhaustion [34], a finding validated in immune-dysregulated patients in this study. Through protein interaction networks constructed using the STRING database [35], the study identified crosstalk and regulatory relationships between multiple signaling pathways, discovering new hub genes such as ST8SIA4 and ERAP2, which may become new therapeutic targets.

The main strength of this study lies in its comprehensiveness and systematic approach. Through integrating multi-omics data from multiple large databases, the study achieved unprecedented sample size and data depth. The machine learning methods employed combined advantages of multiple algorithms, improving classification accuracy and stability. The constructed molecular typing system not only has biological significance but also demonstrates clinical application value. However, public database analysis also has inherent limitations. Issues such as data quality heterogeneity, incomplete clinical information, and missing follow-up data affect analysis depth. Samples mainly come from European and American populations, with insufficient representation from Asia and other regions, potentially affecting result generalizability. The impact of treatment history on gene expression is difficult to completely exclude, particularly patients using biological agents may exhibit different molecular features.

The molecular typing system established in this study requires validation in prospective clinical cohorts. Well-designed multicenter prospective studies can evaluate the clinical utility, treatment guidance value, and prognostic prediction capability of the typing system. Future research should integrate more omics data, including proteomics, metabolomics, microbiomics, etc., to provide more

comprehensive disease understanding. Based on identified subtype-specific therapeutic targets, developing new targeted treatment strategies is an important research direction. Design of combination therapy regimens should consider molecular subtype characteristics to achieve precise individualized treatment. Biomarker-guided clinical trial design will improve success rates of new drug development, ultimately improving prognosis for AS patients.

5. Conclusion

This study successfully identified three molecular subtypes of HLA-B27 positive ankylosing spondylitis by integrating multi-omics data from 407,734 samples, providing a systematic molecular-level explanation for disease heterogeneity. Patients with inflammation-dominant (37.8%), fibrosis-progressive (32.1%), and immune-dysregulated (30.1%) subtypes exhibited distinct molecular features, clinical phenotypes, and treatment response patterns. The established classification model achieved 86.5% accuracy in an independent validation cohort, demonstrating good clinical application potential. Subtype-specific treatment response analysis revealed possibilities for precision therapy: inflammation-dominant patients showed a 72.3% response rate to anti-TNF therapy, while immune-dysregulated patients had a 68.5% response rate to JAK inhibitors, providing molecular basis for developing individualized treatment strategies. The developed 35-gene minimal classification set not only maintained 85.2% classification accuracy but also reduced clinical application costs. The risk scoring model achieved a C-index of 0.78, providing a reliable tool for disease prognosis assessment. The study also identified multiple subtype-specific therapeutic targets with high druggability scores (>0.7), including TNF (0.92), CTLA4 (0.91), and RANKL (0.87), indicating directions for new drug development. By integrating transcriptomic, genomic, single-cell sequencing, and immunomic data, the study constructed a complete disease map from genetic susceptibility to molecular mechanisms, from cellular function to clinical phenotypes. This molecular typing system not only deepens understanding of HLA-B27 positive AS pathogenesis but, more importantly, provides practical classification tools and treatment guidance frameworks for achieving precision medicine, with potential to improve clinical management and prognosis for patients.

Conflict of interest: The authors declare no conflict of interest.

References

1. Khan, M.A., *HLA-B* 27 and ankylosing spondylitis: 50 years of insights and discoveries.* Current Rheumatology Reports, 2023. 25(12): p. 327-340.

2. Chen, B., et al., *Role of HLA-B27 in the pathogenesis of ankylosing spondylitis.* Molecular medicine reports, 2017. 15(4): p. 1943-1951.

3. Berland, M., et al., *Both disease activity and HLA–B27 status are associated with gut microbiome dysbiosis in spondyloarthritis patients.* Arthritis & rheumatology, 2023. 75(1): p. 41-52.

4. Kim, S.H. and S.-H. Lee, *Updates on ankylosing spondylitis: pathogenesis and therapeutic agents.* Journal of rheumatic diseases, 2023. 30(4): p. 220-233.

5. Xiong, Y., et al., *Joint together: the etiology and pathogenesis of ankylosing spondylitis.* Frontiers in Immunology, 2022. 13: p. 996103.

6. Baeten, D. and I.E. Adamopoulos, *IL-23 inhibition in ankylosing spondylitis: where did it go wrong?* Frontiers in Immunology, 2021. 11: p. 623874.

7. Yu, H., et al., *Gene-regulatory network analysis of ankylosing spondylitis with a single-cell chromatin accessible assay.* Scientific Reports, 2020. 10(1): p. 19411.

8. Liu, L., et al., *Osteoimmunological insights into the pathogenesis of ankylosing spondylitis.* Journal of cellular physiology, 2021. 236(9): p. 6090-6100.

9. Remalante-Rayco, P. and A. Nakamura, *Year in Review: Novel Insights in the Pathogenesis of Spondyloarthritis–SPARTAN 2024 Annual Meeting Proceedings.* Current Rheumatology Reports, 2025. 27(1): p. 9.

10. Wen, P., et al., *Identification of necroptosis-related genes in ankylosing spondylitis by bioinformatics and experimental validation.* Journal of Cellular and Molecular Medicine, 2024. 28(14): p. e18557.

11. Russell, T., et al. *Cytokine "fine tuning" of enthesis tissue homeostasis as a pointer to spondyloarthritis pathogenesis with a focus on relevant TNF and IL-17 targeted therapies*. in *Seminars in Immunopathology*. 2021. Springer.

12. Thaiss, C.A., et al., *The microbiome and innate immunity.* Nature, 2016. 535(7610): p. 65-74.

13. Zhang, Y., G. Parmigiani, and W.E. Johnson, *ComBat-seq: batch effect adjustment for RNA-seq count data.* NAR genomics and bioinformatics, 2020. 2(3): p.

lqaa078.

14. Hasin, Y., M. Seldin, and A. Lusis, *Multi-omics approaches to disease.* Genome biology, 2017. 18(1): p. 83.

15. Dai, L., et al., *Identifying prioritization of therapeutic targets for ankylosing spondylitis: a multi-omics Mendelian randomization study.* Journal of Translational Medicine, 2024. 22(1): p. 1115.

16. Sun, X., et al., *Identification of clinical heterogeneity and construction of a novel subtype predictive model in patients with ankylosing spondylitis: An unsupervised machine learning study.* International Immunopharmacology, 2023. 117: p. 109879.

17. Ma, Y., et al., *Exploring the common mechanisms and biomarker ST8SIA4 of atherosclerosis and ankylosing spondylitis through bioinformatics analysis and machine learning.* Frontiers in Cardiovascular Medicine, 2024. 11: p. 1421071.

18. Mongardi, S., S. Cascianelli, and M. Masseroli, *Biologically weighted LASSO: enhancing functional interpretability in gene expression data analysis.* Bioinformatics, 2024. 40(10): p. btae605.

19. Song, M., et al., *Diagnostic classification and biomarker identification of Alzheimer's disease with random forest algorithm.* Brain sciences, 2021. 11(4): p. 453.

20. Li, D., et al., *Identification of potential biomarkers for ankylosing spondylitis based on bioinformatics analysis.* BMC Musculoskeletal Disorders, 2023. 24(1): p. 413.

21. Reimand, J., et al., *Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap.* Nature protocols, 2019. 14(2): p. 482-517.

22. Candia, J. and L. Ferrucci, *Assessment of Gene Set Enrichment Analysis using curated RNA-seq-based benchmarks.* Plos one, 2024. 19(5): p. e0302696.

23. Han, Y., et al., *Identification of diagnostic mRNA biomarkers in whole blood for ankylosing spondylitis using WGCNA and machine learning feature selection.* Frontiers in Immunology, 2022. 13: p. 956027.

24. Pimentel-Santos, F.M., et al., *Whole blood transcriptional profiling in ankylosing spondylitis identifies novel candidate genes that might contribute to the inflammatory and tissue-destructive disease aspects.* Arthritis research & therapy, 2011. 13(2): p. R57.

25. Li, H., et al., *Comprehensive AI-assisted tool for ankylosing spondylitis based on multicenter research outperforms human experts.* Frontiers in Public Health, 2023. 11: p. 1063633.

26. Yi, K., et al., *Analysis of Single-Cell Transcriptome and Surface Protein Expression in Ankylosing Spondylitis Identifies OX40-Positive and Glucocorticoid-Induced Tumor Necrosis Factor Receptor–Positive Pathogenic Th17 Cells.* Arthritis & Rheumatology, 2023. 75(7): p. 1176-1186.

27. Chen, T., et al., *Exploring T Cell and NK Cell Involvement in Ankylosing Spondylitis Through Single-Cell Sequencing.* Journal of Cellular and Molecular Medicine, 2024. 28(24): p. e70206.

28. Tran, H.T.N., et al., *A benchmark of batch-effect correction methods for single-cell RNA sequencing data.* Genome biology, 2020. 21(1): p. 12.

29. Deodhar, A., et al., *Upadacitinib for the treatment of active non-radiographic axial spondyloarthritis (SELECT-AXIS 2): a randomised, double-blind, placebo-controlled, phase 3 trial.* The Lancet, 2022. 400(10349): p. 369-379.

30. Yin, Y., et al., *Efficacy and safety of IL-17 inhibitors for the treatment of ankylosing spondylitis: a systematic review and meta-analysis.* Arthritis research & therapy, 2020. 22(1): p. 111.

31. Gracey, E., et al., *TYK2 inhibition reduces type 3 immunity and modifies disease progression in murine spondyloarthritis.* The Journal of clinical investigation, 2020. 130(4): p. 1863-1878.

32. Vecellio, M., et al., *The IL-17/IL-23 axis and its genetic contribution to psoriatic arthritis.* Frontiers in immunology, 2021. 11: p. 596086.

33. Hammitzsch, A., G. Lorenz, and P. Moog, *Impact of Janus kinase inhibition on the treatment of axial spondyloarthropathies.* Frontiers in Immunology, 2020. 11: p. 591176.

34. Tang, M., et al. *Multi-omics Immune Profiling of Cytotoxic T Cells from Ankylosing Spondylitis Patients Identified Subset of Clonally Exapnded CTLs That Evade Immune Exhaustion*. in *ARTHRITIS & RHEUMATOLOGY*. 2023. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA.

35. Szklarczyk, D., et al., *The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest.* Nucleic acids research, 2023. 51(D1): p. D638-D646.