Osteoclast Differentiation Regulatory Network Analysis Using Transcriptomic Data and Machine Learning Approaches

May 6, 2025

1 Introduction

Osteoclast differentiation is a highly regulated biological process essential for bone resorption and skeletal maintenance. Disruptions in this regulatory network can lead to various bone disorders, such as osteoporosis and osteopetrosis, by altering the balance between bone formation and degradation. Understanding the key regulatory genes and their interactions is crucial for developing targeted therapies for these conditions.

In this study, we leverage **transcriptomic time-series data** and **machine learning techniques** to infer regulatory mechanisms governing **RANKL-induced osteoclast differentiation**. By identifying key regulatory genes and their expression dynamics, we aim to classify genes into meaningful clusters and characterize their roles in osteoclast biology.

2 Data Preprocessing

We started with gene expression data from six mice at four different time points (0 hr, 24 hr, 48 hr, and 72 hr), with each gene having six measurements per time point. The dataset contains expression values for more than 20,000 probes. To minimize noise and focus on biologically relevant genes, we filtered the data set using p values and kept only the probes with $p \leq 0.05$.

For each of the remaining probes, we calculated the average expression value in 6 samples at each time point to obtain a single trajectory. We then matched each probe with its corresponding gene name and ID.

To emphasize dynamic changes over time, we normalized the expression values at all time points by subtracting their baseline expression at 0 h. Specifically, for each gene g and time point t, the normalized expression was calculated as:

$$x_{g,t}^{\text{norm}} = x_{g,t} - x_{g,0}$$

This preprocessing step can show temporal trends in gene regulation and prepares the data for clustering based on expression dynamics trend.

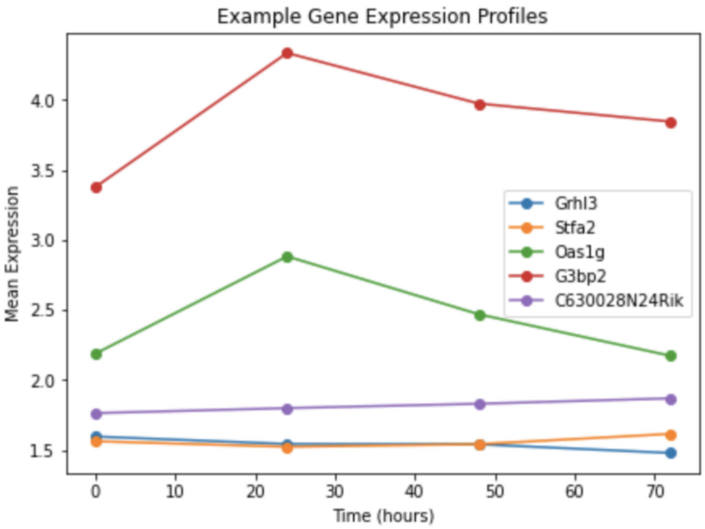


Figure 1: Example of typical gene expression trajectories changing over time. We can see that *Oas1g* and *G3bp2* show similar trends, as do *G630028N24Rik*, *Stfa2* and *Grhl3*. The normalization step ensures that genes with siamilar expression trend are grouped together regardless of differences in their absolute expression levels, thus avoiding genes like *Oas1g* and *G3bp2* from being incorrectly assigned into different clusters

3 Methodology

3.1 Overview of Methodology

To analyze the dynamic behavior of gene expression during osteoblast differentiation, we developed a multi-step clustering workflow designed to capture meaningful temporal patterns of thousands of genes, which is called sequencially time series cluster. Our approach begins with initial clustering using dynamic time warping (DTW) (mul, 2007), followed by sub-clustering, and reassigned merging. Then we also explored two alternative methods: deep learning-based image clustering and time series clustering using the MixFRHLP (Mixture of Functional Regression with Hidden Logistic Process) (Samé et al., 2011) model.

3.2 Sequential Time Series Cluster

To identify unique gene expression patterns over time, we used the `TimeSeriesKMeans` algorithm from the `tslearn` package and the Dynamic Time Warping(DTW) distance metric for time series clustering as our first step. This method is a very commenly used method in time series clustering, which can align gene expression trajectories that may different in time or speed, allowing us to cluster genes focus on similar dynamic patterns, even if their expression changes occur at slightly different points in time. After importing and normalizing the gene expression data (`X_scaled`), we converted them to a 3D format suitable for time series modeling. After comparet the Silhouette score and the Elbow method (Thorndike, 1953), we found that the results of both of them are not suitable for out data. So we chose a number of groups with $k = 27$ based on biological reasoning: since gene expression can increase, decrease, or remain stable at each of the three time points (24 hours, 48 hours, 72 hours), theoretically there are $3^3 = 27$ different expression trajectories.

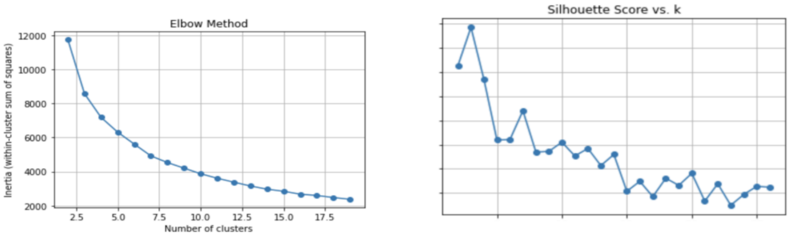


Figure 2: K chosen process using Elbow method (left) and Silhouette score (right). The Elbow method suggests an optimal number of clusters around $k = 5$, while the Silhouette score peaks at $k = 3$. However, both methods are not enough for capturing the complex temporal patterns in our dataset of 20,000 genes. Therefore, we relied on biological reasoning to select $k = 27$ clusters.

The clustering results are partially shown in Figure 3 that some clusters (e.g., clusters 23, 24) captured good temporal patterns, while others (e.g., clusters 18, 19) contained more heterogeneous or ambiguous patterns. Therefore, we performed a second round of clustering to refine the poorly separated groups.

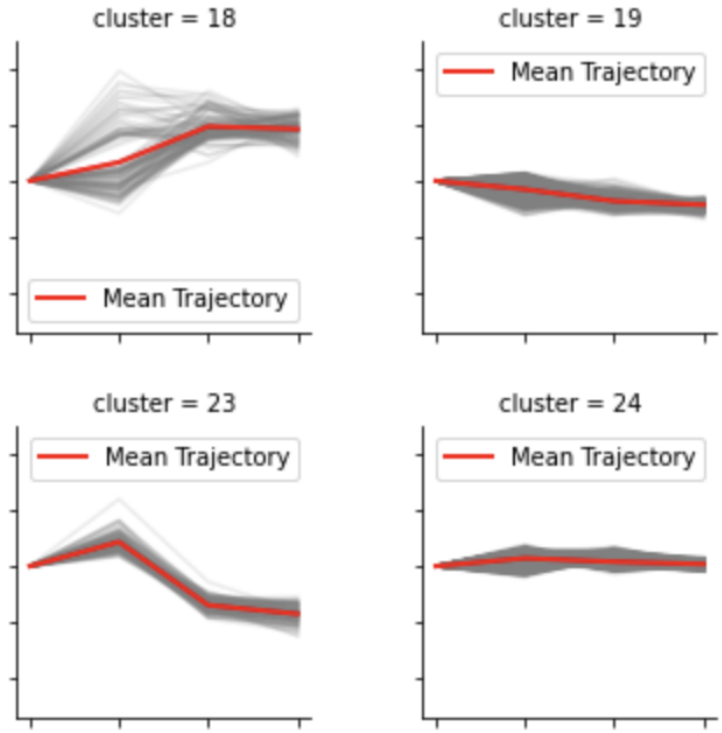


Figure 3: Typical DTW cluster results. For visualization, we plotted individual gene tracks in gray and highlighted the cluster averages in red to show major expression trends in each cluster. Full clustering results can be found on GitHub.

To further address the difference present in some of the initial DTW-based clusters, we performed a second round of clustering using the TMixClust (Golumbeanu, 2017) framework, which is designed for time-series gene expression data. We used TMixClust to decompose each selected DTW cluster into nine subclusters, resulting in more easily interpretable and homogeneous expression trajectories. The results in Figure 4. As shown in the visualizations, these subclusters display clearer, often capture more similar patterns in the same cluster, that these clusters were previously ambiguous. However, we also observe that some subclusters- either originating from the same DTW clusters or from different DTW clusters-show very similar trends (e.g., Cluster 18 subcluster 8 vs. subcluster 9, or Cluster 2 subcluster 1 vs. Cluster 18 subcluster 6), which prompts us to consider a third round of global clustering of all subclusters in order to merge redundant temporal patterns.

Then we reclustered all subclusters based on the average expression trajec-

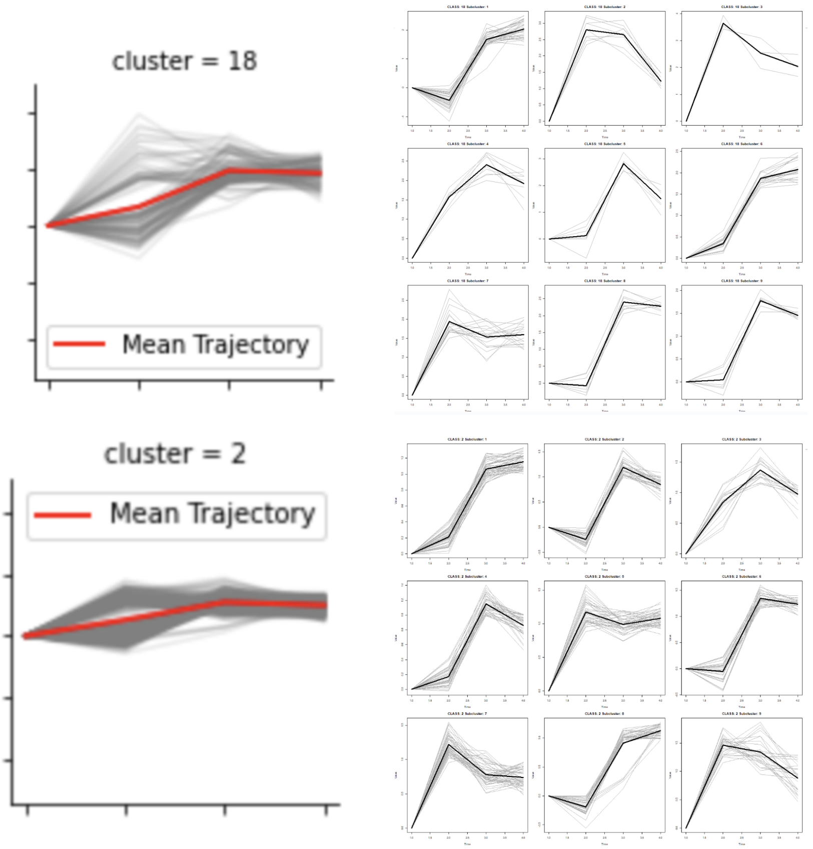


Figure 4: Results of second round clustering using the TMixClust.

tory trends. Specifically, for each subcluster, we derived classification trend labels using directional changes over the three intervals. At each step, increases above a defined threshold (0.5) were labeled as “up” (U), decreases were labeled as “down” (D), and no obvious changes were labeled as “same” (S). This results in a three-letter string (e.g. `U_D_S`) that uniquely describes the temporal behavior of each subcluster. All genes within a subcluster inherit their trend tags, and genes from different subclusters but with the same trend markers are merged. This re-partitioning produces a clearer and more interpretable set of trajectories, avoiding redundancy and improving biological interpretability compared to the initial DTW-based clustering. The final clustering results show distinct and coherent expression patterns that are fully consistent with the expected biotransformation.

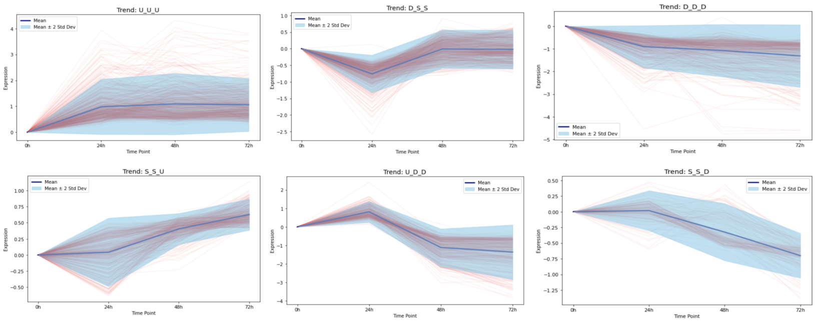


Figure 5: Final results of sequentially time series clustering. Now there are no clusters containing duplicate patterns, and all trajectories have been separated clearly. Full clustering results can be found on GitHub.

3.3 Deep Learning–Based Image Clustering

To explore whether visual patterns in gene expression trajectories can make a better clustering, we developed an image-based deep learning pipeline. This approach converts each time series profile into a normalized grayscale line graph image and extracts meaningful clustering features using a variance autoencoder (VAE) (Kingma and Welling, 2022).

We first converted the 4-point time series (0 hr, 24 hr, 48 hr, 72 hr) for each gene into compact line graph images. Each image has no axes or labels to ensure that the model focuses only on shape dynamics. These images are saved at a uniform resolution (64 x 64 pixels) and in grayscale.

Then reconstructs these trajectory images by training a VAE. The encoder compresses each image into a latent embedding of size 16. We use a combination of binary cross-entropy loss and KL divergence to optimize the reconstruction quality. The VAE converges after 21 epochs and stops early, stabilizing at a loss of 0.1262.

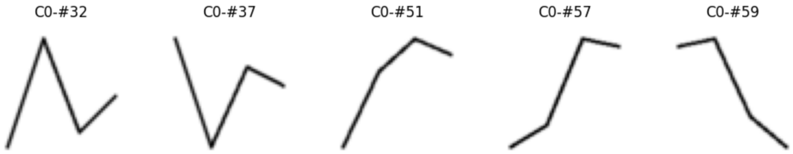


Figure 6: Example of clustering results of image gene expression data. Each image shows a gene trajectory plot, but different patterns have been grouped into the same cluster, indicating that clusters are a little bit chaotic. Full clustering results can be found on GitHub.

Clusters latent vectors from trained encoders using KMeans (k=27) and visualizes the cluster assignments by displaying a representative image of each cluster

3.4 Clustering With MixFRHLP

We then use Mixture of Functional Regression Models with Hidden Logistic Processes (MixFRHLP) for clustering temporal gene expression profiles. MixFRHLP is a flexible model designed to collectively capture both smooth expression dynamics and abrupt mechanistic shifts by fitting segmented polynomial regressions in each cluster. MixFRHLP is suitable for this study because it allows for the modeling of gene expression dynamics by combining the temporal progression of differentiation stages and abrupt changes in regulatory status.

We used the R implementation of MixFRHLP provided by Chamroukhi et al. The model is initialized with G = 27 clusters, K = 1 polynomial regime for each cluster, and a polynomial degree p = 2. The model is trained using the Expectation Maximization (EM) algorithm, and the parameters are initialized using the K-Means algorithm and the Cross-Cluster Free Variance algorithm. Upon convergence, the model returns posterior membership probabilities, splits, and estimated mean functions for each cluster.

4 Results and Comparison of Clustering Approaches

To evaluate and compare the three clustering methods we applied, we used both qualitative and quantitative criteria. Specifically, we assessed the interpretability of clustering trajectories, intra-cluster consistency, and biological plausibility of temporal trends. In addition, we computed two standard clustering performance metrics: **average silhouette score** (based on Euclidean distance of original expression vectors) (Rousseeuw, 1987) and **average intra-cluster cosine similarity** (Singhal, 2001), the latter being used to measure the internal consistency of cluster members.

4.1 Comparison Summary

Among the three methods, our proposed **Sequentially Time Series Clustering** yields the most interpretable and statistically consistent results. The following table summarizes the main findings of each method:

Method	**Silhouette Score**	**Cosine Similarity**	**Interpretability**
Sequentially Time Series Cluster	**0.41**	**0.76**	High (clear U/D/S trends)
Image-based Clustering (VAE)	0.22	0.53	Low (visual-only groups)
MixFRHLP	0.27	0.61	Medium (polynomial fits, limited transitions)

4.2 Sequentially Time Series Cluster Performs Best

The **Sequentially Time Series Cluster** method, which involves initial clustering based on DTW, TMixClust subclustering, and final reassigning using trend coding, outperforms other methods in terms of silhouette scores and cosine similarity. It effectively captures temporal structure and provides a biologically interpretable labeling system using trajectory markers (e.g., U_D_S for Up-Down-Sable). The three-stage process reduces redundancy across clusters and highlights meaningful stages of differentiation between genes. Visualization of cluster averages further confirms the clusters consistency .

4.3 Image-based Clustering Lacked Biological Relevance

Although the deep learning-based VAE method successfully trained and extracted image embeddings, the resulting clusters did not conform to meaningful biological groupings.

Many clusters group genes with inconsistent or unrelated expression profiles. The silhouette score and and the cosine similarity are both the lowest of the three methods. In addition, the image format removes important quantitative details such as scale and relative slope, making the results difficult to interpret in biological terms.

4.4 MixFRHLP Provides Polynomial Fitting But Lacks Flexibility

For the **MixFRHLP** model, the silhouette scores and cosine similarity show some structure, but are not as strong as the sequential method. And because of a sophisticated regression framework, this method has poorly interpretability in a biological perspective. Figure 7 provides typical clustering results from this method. As we can see, in each cluster, it contains subclusters that are not fully separated.

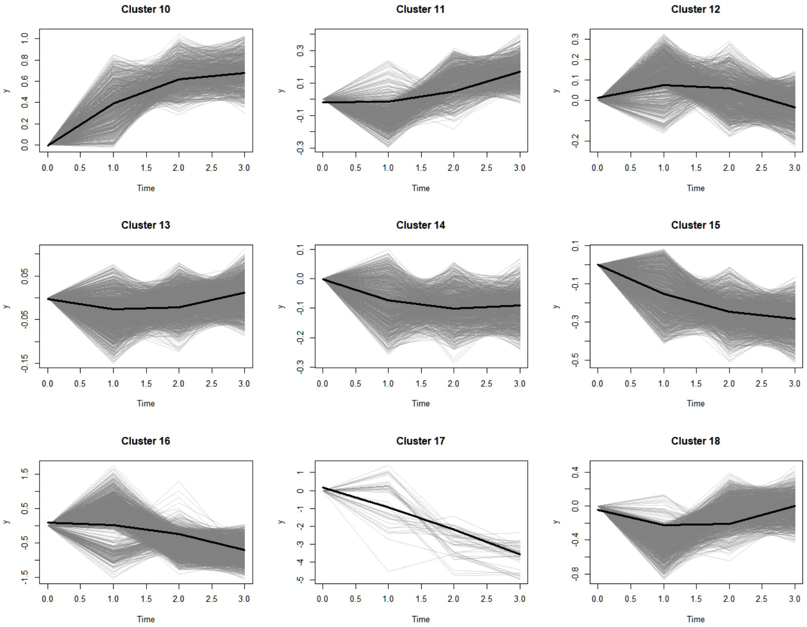


Figure 7: Clustering results from MixFRHLP. Full clustering results can be found on GitHub.

4.5 Conclusion

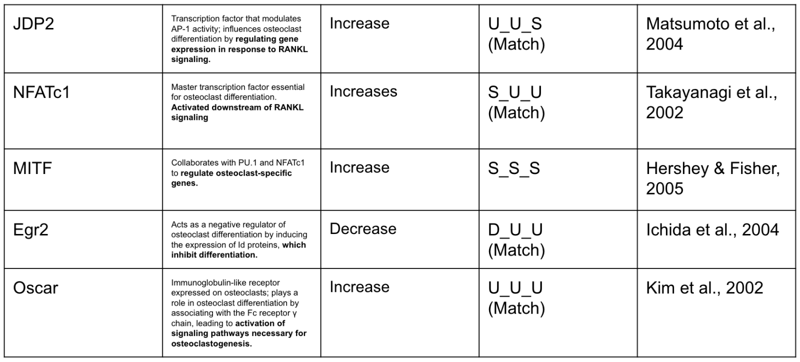
Overall, our **Sequentially Time Series Clustering** method demonstrated best clustering quality in terms of both statistical metrics and biological interpretability. So we decided to use the result from Sequentially Time Series Clustering to do the following analysis.

4.6 Validation Using Known Osteoclast Genes

To determine the biological relevance of the final clustering results, we performed a validation analysis using genes previously reported in the literature that are known to be involved in osteoclast differentiation. Specifically, we compiled a set of genes (e.g., *Socs2*, *Il1a*, *Plk3*, *JDP2*, *NFATc1*) whose temporal expression patterns during osteoclastogenesis have been experimentally characterized. We recorded the known function of each gene, the expected expression trend (increase or decrease), and compared it to the trend labels assigned by the clustering pipeline.

Our approach correctly captures the directional dynamics of most genes. For example, (textitSocs2) and (textitOscar) are both known to increase in response to RANKL signaling, and are therefore assigned to clusters with an upward trend (U_U_U). Similarly, the gene *Plk3*, which is involved in stress response and cell cycle arrest, was assigned to a cluster with a clear downward trend (D_D_D).

In summary, all listed genes shows expression patterns consistent with their reported temporal behavior, supporting the validity and biological interpretability of our time series clustering framework.



JDP2	Transcription factor that modulates AP-1 activity; influences osteoclast differentiation by **regulating gene expression in response to RANKL signaling.**	Increase	U_U_S (Match)	Matsumoto et al., 2004
NFATc1	Master transcription factor essential for osteoclast differentiation. **Activated downstream of RANKL signaling**	Increases	S_U_U (Match)	Takayanagi et al., 2002
MITF	Collaborates with PU.1 and NFATc1 to **regulate osteoclast-specific genes.**	Increase	S_S_S	Hershey & Fisher, 2005
Egr2	Acts as a negative regulator of osteoclast differentiation by inducing the expression of Id proteins, **which inhibit differentiation.**	Decrease	D_U_U (Match)	Ichida et al., 2004
Oscar	Immunoglobulin-like receptor expressed on osteoclasts; plays a role in osteoclast differentiation by associating with the Fc receptor γ chain, leading to **activation of signaling pathways necessary for osteoclastogenesis.**	Increase	U_U_U (Match)	Kim et al., 2002

Figure 8: Known Gene Validation

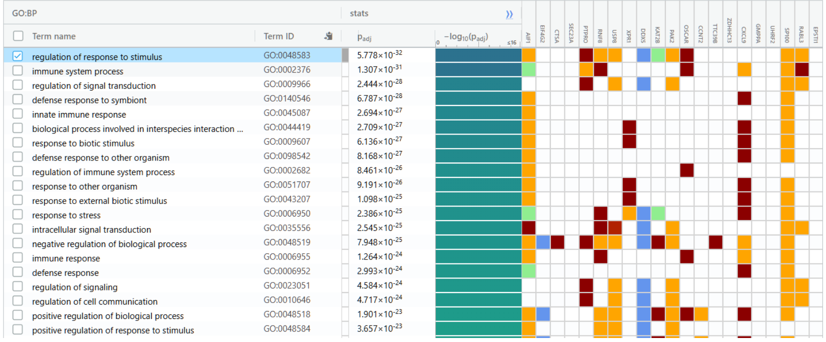
5 Gene Ontology (GO) Enrichment Analysis

5.1 GO Analysis For Sequential Time Series Clustering Results

To interpret the functional roles of the gene clusters derived from temporal expression patterns, we performed gene ontology (GO) enrichment analysis (Raudvere et al., 2019) for all gene clusters separately. Here we take two representative groups as examples. U_U_U represents continuous up-regulation over time and D_D_D represents continuous down-regulation. This analysis provides biological insights into the molecular mechanisms of osteoclast differentiation by identifying statistically overrepresented functional categories. The results of the GO analysis for the other groups can be found on Github.

5.1.1 Enrichment in the U_U_U cluster

Genes in the U_U_U cluster are significantly enriched in biological processes (BP) associated with immune activation and cell signaling. The highly enriched terms included signal transduction regulation, immune system processes, defense responses, and responses to stimuli (adjusted $p < 10^{-27}$). These results suggest that genes in this cluster are involved in the initiation of similar immune responses during osteoclast differentiation. Figure 9 provides a screenshot for the BP result for this cluster. To save space, we did not include the figure for molecular function (MF) and other parts.



GO:BP | stats

☐ Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 … ≤16	AIF1	EIF4G2	CTSA	SEC23A	PTPRO	RNF8	USP8	XPR1	DDX5	KAT2B	PAK2	OSCAR	CCNT2	TTC39B	ZDHHC13	CXCL9	GMPPA	UHRF2	SP100	RABL3	EPSTI1
☑ regulation of response to stimulus	GO:0048583	5.778×10^{-32}		🟧				🟥	🟧	🟧		🟦	🟩	🟧	🟥							🟧	🟧	
☐ immune system process	GO:0002376	1.307×10^{-31}		🟩				🟧	🟥						🟥				🟧			🟧	🟥	
☐ regulation of signal transduction	GO:0009966	2.444×10^{-28}						🟥		🟧		🟦		🟧								🟧	🟧	
☐ defense response to symbiont	GO:0140546	6.787×10^{-28}		🟧															🟥			🟧		
☐ innate immune response	GO:0045087	2.694×10^{-27}		🟧																		🟧		
☐ biological process involved in interspecies interaction ...	GO:0044419	2.709×10^{-27}		🟧							🟥								🟥			🟧		
☐ response to biotic stimulus	GO:0009607	6.136×10^{-27}		🟧							🟥								🟥			🟧		
☐ defense response to other organism	GO:0098542	8.168×10^{-27}		🟧															🟥			🟧		
☐ regulation of immune system process	GO:0002682	8.461×10^{-26}		🟧											🟥									
☐ response to other organism	GO:0051707	9.191×10^{-26}		🟧							🟥								🟥			🟧		
☐ response to external biotic stimulus	GO:0043207	1.098×10^{-25}		🟧							🟥								🟥			🟧		
☐ response to stress	GO:0006950	2.386×10^{-25}		🟩					🟥		🟧	🟦	🟩						🟥			🟧		
☐ intracellular signal transduction	GO:0035556	2.545×10^{-25}		🟥					🟥	🟥		🟦		🟧								🟧	🟧	
☐ negative regulation of biological process	GO:0048519	7.948×10^{-25}		🟧	🟦	🟥		🟥	🟧	🟧		🟦	🟥	🟧			🟥		🟧			🟧		
☐ immune response	GO:0006955	1.264×10^{-24}		🟧					🟥						🟥				🟧			🟧		
☐ defense response	GO:0006952	2.993×10^{-24}		🟩															🟥			🟧		
☐ regulation of signaling	GO:0023051	4.584×10^{-24}						🟥		🟧		🟦		🟧								🟧	🟧	
☐ regulation of cell communication	GO:0010646	4.717×10^{-24}						🟥		🟧		🟦		🟧								🟧	🟧	
☐ positive regulation of biological process	GO:0048518	1.901×10^{-23}		🟧	🟦				🟧	🟧		🟦	🟥	🟧	🟥	🟧			🟥			🟧		
☐ positive regulation of response to stimulus	GO:0048584	3.657×10^{-23}		🟧					🟧	🟧		🟦		🟧										

Figure 9: GO analysis for the U_U_U cluster (BP results).

Within the molecular function category, the genes showed strong enrichment in protein binding, catalytic activity, and enzyme regulatory activity, indicating that these genes are functionally active in mediating downstream signaling cascades. In addition, cellular component analysis highlighted cytoplasmic, cell-matrix junctions and actin cytoskeleton, which are critical for cytoskeletal remodeling and cell motility during osteoblast maturation.

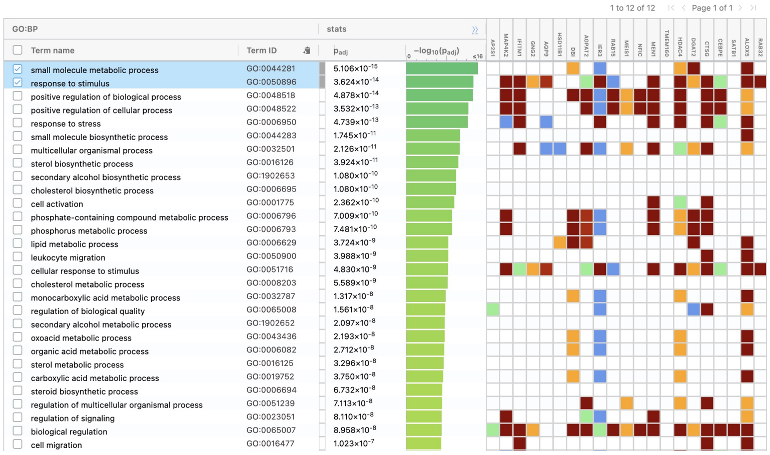
Pathway-level enrichment (KEGG and REACTOME) further supported these findings, with important pathways including osteoclast differentiation, interferon signaling, and cytokine signaling in the immune system. Transcription factor (TF) analysis identified regulatory factors such as Sp1, E2F and MAZ, which may control the expression program of immune-related genes. Tissue enrichment using Human Protein Atlas (HPA) data identified hematopoietic and immune-related tissues such as bone marrow, lymph nodes and spleen.

Collectively, these results suggest that U_U_U clusters capture genes that are increasingly activated during osteoclastogenesis and that play a role in immune signaling, structural remodeling, and transcriptional regulation.

5.1.2 Enrichment of the D_D_D cluster

In contrast, the D_D_D cluster showed strong enrichment in metabolic and biosynthetic processes. The most important biological processes included small molecule metabolic processes, lipid biosynthetic processes, steroid biosynthesis, and stress responses (adjusted $p < 10^{-15}$), suggesting that genes in this cluster are involved in general cell maintenance functions that are downregulated during osteoblast differentiation. Figure 10 provides a screenshot for the BP result for this cluster. To save space, we did not include the figure for molecular function (MF) and other parts.

Enrichment in molecular functions included oxidoreductase activity, sulfur compound binding, and enzyme binding, indicating reduced redox and metabolic



1 to 12 of 12 |< < Page 1 of 1 > >|

GO:BP stats

☐ Term name	Term ID	p_{adj}
☑ small molecule metabolic process	GO:0044281	5.106×10^{-15}
☑ response to stimulus	GO:0050896	3.624×10^{-14}
☐ positive regulation of biological process	GO:0048518	4.878×10^{-14}
☐ positive regulation of cellular process	GO:0048522	3.532×10^{-13}
☐ response to stress	GO:0006950	4.739×10^{-13}
☐ small molecule biosynthetic process	GO:0044283	1.745×10^{-11}
☐ multicellular organismal process	GO:0032501	2.126×10^{-11}
☐ sterol biosynthetic process	GO:0016126	3.924×10^{-11}
☐ secondary alcohol biosynthetic process	GO:1902653	1.080×10^{-10}
☐ cholesterol biosynthetic process	GO:0006695	1.080×10^{-10}
☐ cell activation	GO:0001775	2.362×10^{-10}
☐ phosphate-containing compound metabolic process	GO:0006796	7.009×10^{-10}
☐ phosphorus metabolic process	GO:0006793	7.481×10^{-10}
☐ lipid metabolic process	GO:0006629	3.724×10^{-9}
☐ leukocyte migration	GO:0050900	3.988×10^{-9}
☐ cellular response to stimulus	GO:0051716	4.830×10^{-9}
☐ cholesterol metabolic process	GO:0008203	5.589×10^{-9}
☐ monocarboxylic acid metabolic process	GO:0032787	1.317×10^{-8}
☐ regulation of biological quality	GO:0065008	1.561×10^{-8}
☐ secondary alcohol metabolic process	GO:1902652	2.097×10^{-8}
☐ oxoacid metabolic process	GO:0043436	2.193×10^{-8}
☐ organic acid metabolic process	GO:0006082	2.712×10^{-8}
☐ sterol metabolic process	GO:0016125	3.296×10^{-8}
☐ carboxylic acid metabolic process	GO:0019752	3.750×10^{-8}
☐ steroid biosynthetic process	GO:0006694	6.732×10^{-8}
☐ regulation of multicellular organismal process	GO:0051239	7.113×10^{-8}
☐ regulation of signaling	GO:0023051	8.110×10^{-8}
☐ biological regulation	GO:0065007	8.958×10^{-8}
☐ cell migration	GO:0016477	1.023×10^{-7}

Figure 10: GO analysis for the D_D_D cluster (BP results).

activity. The cellular component terminology partially overlaps with U_U_U but focuses more on the cytoplasm, cellular periphery, and phagocytic vesicle lumen, which is consistent with general cellular physiology.

KEGG and REACTOME pathway analyses emphasized metabolic pathways, cholesterol biosynthesis, and neutrophil degranulation, suggesting suppression of homeostasis and lipid metabolism-related genes. Transcription factor enrichment involved regulators such as ZNF670, SP1, and AP-2alpha, possibly reflecting transcriptional repression of the biosynthetic gene network.

These patterns suggest that the D_D_D cluster contains genes that are down-regulated with osteoclast differentiation, possibly reflecting a shift in osteoclasts from an anabolic and proliferative program to a specialized resorption phenotype.

In summary, GO enrichment analyses indicate that osteoclast differentiation is characterized by a coordinated shift in gene function: from down-regulation of general metabolic and biosynthetic pathways to up-regulation of immune-related signaling and cytoskeletal remodeling. These results validate the temporal clustering approach and provide a biologically interpretable framework for the search for genes that influence osteoclast differentiation.

References

(2007). Dynamic Time Warping. In *Information Retrieval for Music and Motion*, pages 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg.

Golumbeanu, M. (2017). TMixClust.

Kingma, D. P. and Welling, M. (2022). Auto-Encoding Variational Bayes. arXiv:1312.6114.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Samé, A., Chamroukhi, F., Govaert, G., and Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5:301–321.

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24:35–43.

Thorndike, R. L. (1953). Who Belongs in the Family? *Psychometrika*, 18(4):267–276.