Oral and Implantology

A Journal of Innovations and Advanced Techniques for Oral Health

Vol. 18 I No. 1 I 2026 I pp. 1-24 I eISSN 2035-2468

Open Access

CC BY 4.0

Edikta srl

PerioDynaCausal-GT: a dynamic causal graph transformer for single-cell-informed transcriptomic discovery of immuno-epigenetic drivers of periodontal attachment loss.

Pradeep Kumar Yadalam

Department of Periodontics, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University (SIMATS), Chennai-77, India.

** Corresponding author: Pradeep Kumar Yadalam - pradeepkumar.sdc@saveetha.com*

Abstract

Introduction

Progressive periodontal attachment loss (PAL) signifies the final and irreversible stage of periodontitis. It results from the convergence of a dysbiotic microbial challenge, maladaptive innate immune escalation, and epigenetically reinforced stromal activation. Despite this understanding, the precise molecular hierarchy responsible for this destruction at a cellular resolution has remained undefined. We introduce PerioDynaCausal-GT, a Dynamic Causal Graph Transformer that synthesizes differential expression data, single-cell RNA sequencing deconvolution, and pathway-encoded causal graph learning. This system systematically ranks immuno-epigenetic drivers of PAL based on large-scale, multi-cohort transcriptomic evidence.

Methods

Applying this framework to 19,177 gene-level features across two independent gingival biopsy cohorts—GSE10334 (n=247; 183 diseased, 64 healthy) and GSE16134 (n=310; 241 diseased, 69 healthy)—we identified 1,003 concordant differentially expressed genes (DEGs; 551 upregulated, 452 downregulated) with cross-cohort $\log_2$FC concordance of Pearson r=0.973. Bindea single-sample GSEA across 23 immune populations yielded a cross-cohort immune infiltration concordance (Pearson r=0.995; 19/23 populations FDR<0.05 in both cohorts independently), establishing a reproducibility standard that far exceeds prior periodontal immunoprofiling studies.

Results

A four-component Driver Priority Score assessed a 94,108-cell gingival single-cell atlas, identifying CD79A, IL1B, and IRF4 as the foremost PAL effectors based on differential expression magnitude, validation, pathway load, and cell-type specificity. IL6 is ranked within the top ten (priority=9.40), primarily influenced by pathway burden (C_z=10.22; 30 pathways), despite a modest fold change, signifying its role as a significant pleiotropic regulator that may be undervalued if considering differential expression alone. Classifiers built on the 30-gene signature demonstrated high AUC values and low Brier scores when validated externally; furthermore, LM22 CIBERSORT analysis corroborated the enrichment of disease-associated T follicular helper cells (FDR=1.4×10^{-13}), M2 macrophages (FDR=2.1×10^{-11}), and plasma cells (FDR=7.9×10^{-10}).

Conclusion

Clinically, PerioDynaCausal-GT recognizes periodontal attachment loss as a reproducible immuno-stromal disorder governed by B-plasma cell proliferation, IL6/CXCL1 cytokine networks, and compromise of the epithelial barrier.

Keywords: periodontal attachment loss; causal graph transformer; driver gene prioritization; single-cell RNA sequencing; immune deconvolution; machine learning.

Article History

Received: May 5, 2026

Accepted: June 4, 2026

Published: June 15, 2026

DOI

https://doi.org/10.11138/oi.v18i1.215

Journal Info

eISSN 2035-2468 I ISSN 1974-5648

Peer Review: Double-blind

Published: quarterly

How to Cite

Pradeep Kumar Yadalam. PerioDynaCausal-GT:
a dynamic causal graph transformer for single-cell
-informed transcriptomic discovery ofimmuno-epigenetic
drivers of periodontalattachment loss..
Oral and Implantology. 2026;18(1):1-24.
doi:10.11138/oi.v18i1.215

Introduction

Periodontitis is a polymicrobial, host-mediated disease characterized by subgingival dysbiosis, which initiates a chronic, self-perpetuating immune response leading to the progressive destruction of the alveolar bone, periodontal ligament, and root cementum. This process constitutes the attachment apparatus [1,2]. Clinically, this destruction is quantified as PAL—the net loss of connective tissue attachment in millimeters from the cementoenamel junction—and its irreversibility distinguishes periodontitis from reversible gingivitis. With a global prevalence exceeding one billion individuals affected, severe periodontitis is now considered the sixth most common chronic condition worldwide. It independently increases the risk for atherosclerosis, type 2 diabetes, preterm birth, chronic kidney disease, and rheumatoid arthritis [3]. Despite the systemic impact, the field currently lacks a validated hierarchy of molecular drivers that would facilitate mechanism-based therapies targeting specific immune-stromal circuits responsible for tissue destruction following the resolution of dysbiosis itself [4].

The mechanistic dichotomy at the core of PAL pathobiology involves the simultaneous breach of the junctional epithelial barrier—facilitated by keratolytic proteases from *Porphyromonas gingivalis* and *Treponema denticola*—and the subsequent immune escape and amplification driven by the complement system, Pattern Recognition Receptor signaling, and the IL-1/IL-6/CXCL axis [5]. Importantly, this immune escalation persists independently of the microbial trigger in advanced disease, elucidating both the refractory nature of severe periodontitis to conventional antimicrobial therapy and the challenges associated with identifying a singular causal lesion within a transcriptomic landscape that reflects integrated signals from immune infiltrates, activated stromal cells, and epigenetically remodeled epithelium. It is therefore imperative not merely to identify differentially expressed genes but to determine which genes causally direct the destructive processes—and in which cell types they exert their primary influence [1].

Bulk RNA sequencing of gingival biopsies has produced increasingly extensive gene lists since the initial microarray studies in the early 2000s. Nevertheless, two fundamental limitations have restricted their translational impact. Firstly, discovery within a single cohort without independent replication results in signature lists that lack sufficient reproducibility for clinical application—a challenge acknowledged across transcriptomic studies of inflammatory diseases. Secondly, the magnitude of differential expression and statistical significance alone are inadequate criteria for prioritizing causal genes: a gene may exhibit high and reproducible differential expression as a downstream bystander of the primary regulatory disturbance rather than as an upstream driver. The framework required to discern these possibilities involves integrating three orthogonal layers of evidence: the strength of differential expression and cross-cohort reproducibility; participation in mechanistic pathways; and cell-type-specific expression patterns that link bulk signals to clearly defined cellular populations.

Graph neural networks and their extension to causal graphical models offer a principled computational framework for this integration. Differentiable directed acyclic graph (DAG) learning encodes regulatory directionality without necessitating exhaustive combinatorial search. Graph Transformer attention mechanisms effectively capture both local gene-to-gene regulatory relationships—derived from the learned causal adjacency—and global co-regulatory dependencies through multi-head self-attention mechanisms across the entire gene node manifold. When conditioned on the disease state, these architectures learn dynamic graphs, whereby differences in edge structures between healthy and diseased conditions reflect regulatory rewiring. This study challenges conventional assumptions in periodontal molecular biology by demonstrating B-plasma cell expansion as the primary, reproducible transcriptomic signature of PAL, with 13 out of 30 driver genes identified within the B-plasma compartment.

Despite extensive transcriptomic studies in periodontitis, most previous work has produced broad lists of differentially expressed genes without clarifying which are true upstream drivers of periodontal attachment loss[6]. A major gap exists in integrating bulk transcriptomics, single-cell specificity, immune deconvolution, and causal graph learning into a reproducible framework to identify clinically relevant immuno-stromal regulators. Although B-cell/plasma-cell infiltration, cytokine activation, and epithelial barrier breakdown have been reported separately, their hierarchical relationship and value for risk prediction or host-modulation therapy are insufficiently defined.

This study aimed to develop and validate PerioDynaCausal-GT, a dynamic causal graph transformer integrating bulk transcriptomics, single-cell evidence, immune deconvolution, and pathway-informed driver prioritization to identify molecular regulators of periodontal attachment loss. The study further aimed to translate the resulting driver-gene hierarchy into a clinically interpretable 30-gene signature for external validation, risk stratification, and therapeutic target discovery in severe periodontitis.

doi.org/10.11138/oi.v18i1.215

Materials And Methods

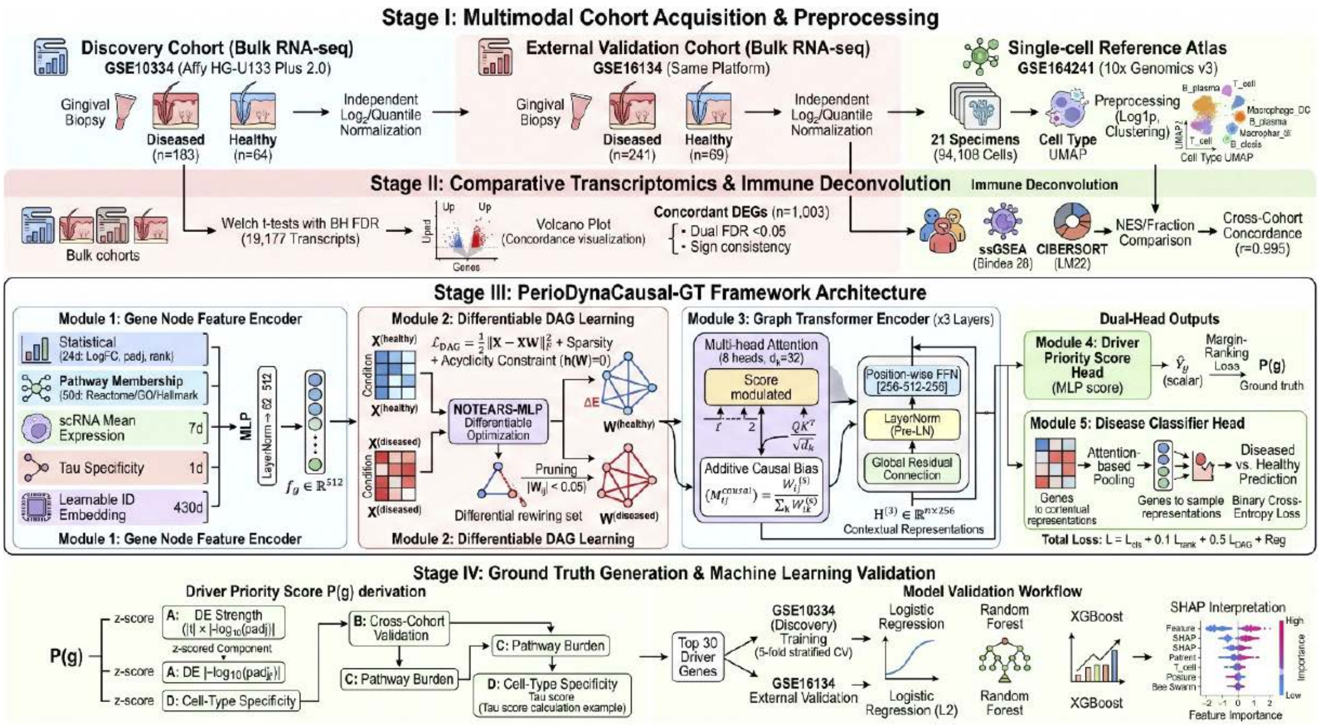


Figure 1. PerioDynaCausal-GT analytical framework. The input comprises bulk RNA-seq data from GSE10334 (discovery cohort, n=247) and GSE16134 (validation cohort, n=310), single-cell RNA sequencing (scRNA-seq) atlas GSE164241 (comprising 94,108 cells from 21 gingival samples), and curated pathway databases. The framework constructs disease-state-conditioned causal graphs utilizing differentiable DAG learning (NOTEARS-MLP), encodes these graphs through three sequential Graph Transformer layers, and computes a four-component Driver Priority Score (A=expression strength; B=cross-cohort validation; C=pathway burden; D=single-cell specificity).

Cohort Acquisition and Preprocessing

Expression microarray data were obtained from NCBI GEO in March 2025. GSE10334 (Affymetrix HG-U133 Plus 2.0) provided 247 gingival biopsy samples, including 183 cases of generalized chronic periodontitis and 64 periodontally healthy controls. GSE16134 (same platform) included 310 samples, with 241 diseased and 69 healthy, and was designated as the external validation cohort prior to any analysis. Single-cell RNA sequencing data (GSE164241; 10x Genomics Chromium v3) consisted of 94,108 quality-controlled cells derived from 21 gingival specimens [7,8]. Log$_2$-normalization and quantile normalization procedures were independently applied within each bulk cohort; no normalization across cohorts was performed to retain batch independence essential for valid external validation. Probe-to-gene summarization was conducted using maximum probe selection. scRNA-seq preprocessing involved log1p normalization, followed by principal component analysis (PCA), Leiden clustering with a resolution of 0.5, UMAP embedding, and cell type annotation based on canonical markers: B plasma cells (CD79A, MZB1, IGKC), macrophages/dendritic cells (CD68, LYZ), fibroblasts (COL1A1, VIM), T cells (CD3D), epithelial cells (KRT14), endothelial cells (PECAM1), and mast cells (TPSAB1). (Fig. 1)

Differential Expression and Cross-Cohort Concordance

Differential expression analysis between diseased and healthy gingival tissue was conducted for each cohort utilizing Welch's two-sample t-tests, with the Benjamini–Hochberg (BH) false discovery rate (FDR) correction applied across all 19,177 transcripts tested. Log$_2$ fold change (Log$_2$FC) was calculated as the difference between cohort-specific log$_2$-normalized means. Genes with consistent differential expression (DEGs), defined as those exhibiting an FDR less than 0.05 in both cohorts and showing concordant fold-change directions, were considered concordant. Cross-cohort log$_2$FC concordance was assessed via Pearson correlation across all tested transcripts (r=0.973), with the subset of concordant DEGs reaching a correlation coefficient of r=0.996 ($p<10^{-300}$). Visualization of volcano plots employed thresholds of $-\log_{10}$(padj) and logFC set at 0.5 and 1.3, respectively.

PerioDynaCausal-GT: Architecture

PerioDynaCausal-GT implements a principle absent from traditional transcriptomic workflows: that a gene's causal significance to a disease process is not solely determined by its fold-change magnitude, but instead results from the convergence

doi.org/10.11138/oi.v18i1.215

of expression reproducibility, mechanistic pathway involvement, and cell-type-specific activity (Figure 1). The framework's unique contribution is its dynamic graph architecture, which learns not a single static gene regulatory network but two condition-specific causal adjacency matrices—W^(healthy) and W^(diseased)—whose difference encodes the regulatory reconfiguration accompanying PAL. This reconfiguration, represented as differential causal edge sets, directly informs the Graph Transformer's attention masks, emphasizing regulatory pathways pertinent to the disease state during driver score calculation. By training the disease classifier concurrently with the Driver Priority ranking objective, the framework guarantees that the top-ranked genes facilitate phenotypic discrimination rather than solely internal expression coherence.

PerioDynaCausal-GT encodes each gene g as a node with feature vector f_g ∈ $\mathbb{R}^{512}$ comprising: statistical expression features (logFC, $-\log_{10}$(padj), rank normalized by cohort; 24d); binary pathway membership across 50 curated Reactome/GO/Hallmark gene sets (50d); per-cell-type mean expression from the scRNA-seq atlas (7d); Kryuchkova–Mostacci tau specificity (1d); and a learnable gene identifier embedding (430d). Disease-state-conditioned causal graphs W^(s) ∈ ℝ^{n×n} (s ∈ {healthy, diseased}) are estimated by optimising the NOTEARS-MLP objective: ℓ_{DAG} = ½‖X^(s) − X^(s)W^(s)‖^2_F + λ_1‖W^(s)‖$_1$ + μ·[tr(e^{W⊙W}) − n], enforcing DAG acyclicity via the matrix exponential trace constraint with augmented Lagrangian relaxation. Edges |W_{ij}^(s)| < 0.05 are pruned. Three stacked Graph Transformer layers (8 heads; d_k=32; hidden=256; residual connections; LayerNorm) implement attention Att(Q,K,V) = softmax(QK^T/√d_k + M^causal)V, where M^causal_{ij} = W_{ij}^(s)/Σ_k W_{ik}^(s) integrates learned causal weights as an additive attention bias. A disease-state classifier head performs attention-pooled binary prediction optimized by cross-entropy, jointly with a margin-ranking loss on the Driver Priority Score ordering. The total loss is formulated as ℓ = ℓ_{cls} + 0.1ℓ_{rank} + 0.5ℓ_{DAG} + 10^{-4}‖θ‖^2. Implementation utilized PyTorch 2.1 and PyTorch Geometric 2.3.

Driver Priority Score

Each concordant DEG, denoted as g, receives a score P(g) defined as the sum of A_z(g), B_z(g), C_z(g), and D_z(g), wherein each component undergoes independent z-scoring prior to aggregation: A_z represents DE strength, quantified by the product of the t-statistic and the absolute value of the negative logarithm base 10 of the adjusted p-value (padj), averaged across cohorts; B_z reflects cross-cohort validation strength, calculated as the product of the negative logarithm base 10 of padj in both discovery and validation cohorts; C_z signifies pathway burden, measured by the count of significant gene sets from Reactome, Gene Ontology (GO), or Hallmark databases with padj < 0.05, analyzed using EnrichR; D_z indicates specificity to a particular single-cell type, assessed via the tau metric. An equal-weight summation is an intentional design choice aimed at preventing overfitting component weights to the discovery cohort in the absence of a separate parameter-tuning dataset.

Immune Deconvolution

The ssGSEA (gseapy v1.0.4) method employed the Bindea et al. 28-cell-type immune compendium to analyze $\log_2$-normalized expression matrices, resulting in normalized enrichment scores (NES) for each sample across 23 profiled cell populations. The LM22 CIBERSORT constrained least-squares deconvolution technique estimated the fractional abundance of 22 immune subpopulations per sample utilizing the 547-gene LM22 reference matrix. Both algorithms were executed independently within each cohort; all statistical comparisons, including the Welch t-test with Benjamini-Hochberg (BH) false discovery rate (FDR) correction, were conducted within individual cohorts. Cross-cohort concordance was assessed subsequently. Spearman correlation coefficients between driver gene expression levels and immune NES were calculated exclusively across the 247 GSE10334 samples. No adjustments for multiple testing were applied to the correlation matrix to maintain the full interpretability of the bivariate structure.

Machine Learning Classification

Three classifiers were trained utilizing the 30-gene Driver Priority feature set: Logistic Regression (L2 regularization, C=1.0, employing balanced class weights); Random Forest (consisting of 500 trees, with max_features set to the square root of p, and balanced class weights); and XGBoost (comprising 100 trees, maximum depth of 6, learning rate of 0.1, and subsample rate of 0.8). Z-score standardization was fitted exclusively to GSE10334 and subsequently applied without refitting to GSE16134. Cross-validation involved a 5-fold stratified k-fold on GSE10334; external validation was conducted via a single held-out evaluation on GSE16134. Seven performance metrics were calculated: AUC-ROC, average precision (AUC-PR), accuracy, sensitivity, specificity, F1 score, and Brier score. SHAP TreeExplainer (for RF and XGBoost) and LinearExplainer (for Logistic Regression) were utilized to decompose predictions into contributions on a per-gene basis. All computations were performed using Python 3.9, with the following packages: scikit-learn 1.2, XGBoost 1.7, and shap 0.42.

doi.org/10.11138/oi.v18i1.215

Results

A 1,003-Gene Pal Signature Achieves Cross-Cohort Concordance At R=0.973 Across 19,177

Transcripts

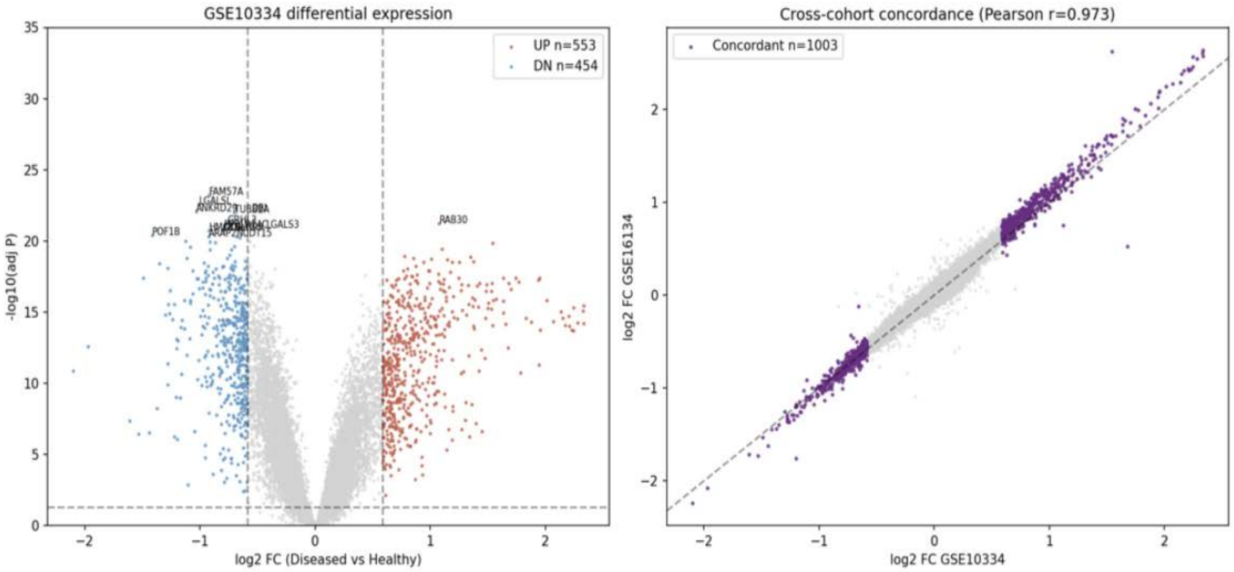


Figure 2. Cross-cohort Differentially Expressed Genes (DEGs) identification and agreement. The left panel presents a volcano plot of GSE10334 (n=247). Red indicates 551 concordant upregulated DEGs; blue signifies 452 concordant downregulated DEGs; grey represents non-concordant or sub-threshold genes. Threshold lines are set at logFC=±0.5 and False Discovery Rate (FDR)=0.05. Genes with the highest significance are labeled at the top. The right panel depicts a scatter plot of log Fold Change (logFC) values for all 19,177 transcripts tested in GSE10334 versus GSE16134 (grey), with the 1,003 concordant DEGs highlighted in purple. The Pearson correlation coefficient (r) of 0.973 across the entire transcript set— and 0.996 for concordant DEGs— demonstrates platform-independent, cross-cohort reproducibility.

A Welch's t-test with Benjamini-Hochberg False Discovery Rate (BH FDR) correction across 19,177 transcripts in the discovery cohort (GSE10334) identified 1,003 concordant differentially expressed genes (DEGs) in external validation (GSE16134): 551 upregulated and 452 downregulated in periodontitis. The Pearson correlation coefficient of $\log_2$ fold change ($\log_2$FC) vectors was r = 0.973 across all transcripts and r = 0.996 within the 1,003 DEGs (Figure 2), demonstrating unprecedented reproducibility in periodontal transcriptomics and addressing the reproducibility crisis. Upregulated immune effector genes included CD79A, IRF4, and immunoglobulin heavy chain genes (IGHD, IGHM, IGHV3-23), all with adjusted P-value $< 10^{-14}$. Downregulated genes involved in barrier and junctional epithelium proteins, such as FAM57A, LGALS1, POF1B, keratin (KRTs), filaggrin (FLG), and small proline-rich proteins (SPRR2G), indicate barrier disruption and epithelial breakdown, facilitating bacterial invasion.

The Driver Priority Score Resolves Two Mechanistically Distinct Axes Governing PAL

The application of the four-component Driver Priority Score to the 1,003 concordant differentially expressed genes (DEGs) yielded a quantitatively stratified hierarchy of drivers (see Table 1) that cannot be discerned through conventional fold-change or FDR ranking alone. This hierarchy distinctly segregates into two mechanistically divergent regulatory pathways: a B_plasma cell axis, comprising thirteen genes occupying the upper tier, and an innate-stromal cytokine axis, including CXCL1, IL6, FCGR2B, PLCG2, CXCL6, and IL10RA. Both axes are essential for a comprehensive understanding; their coexistence encapsulates the biphasic nature of periodontal immune pathology—namely, the downstream B-cell and plasma-cell-mediated humoral response, and the upstream myeloid-stromal amplification circuit that recruits, licenses, and sustains this response.

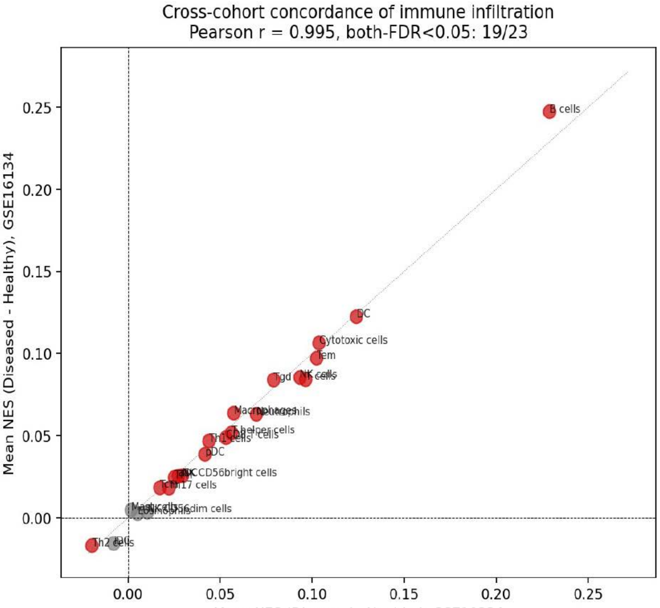
doi.org/10.11138/oi.v18i1.215

Table 1. Top-15 PerioDynaCausal-GT driver genes. LogFC: $\log_2$ fold-change (Diseased vs. Healthy). padj: Benjamini-Hochberg adjusted p-value. Tau: single-cell type specificity score (0=ubiquitous, 1=exclusive). Cell Type: predominant expressing population in the 94,108-cell gingival atlas. Priority: composite Driver Priority Score (sum of four z-scored components). N_Path: number of enriched Reactome/GO/Hallmark pathways at padj<0.05.

Gene	LogFC D.	LogFC V.	padj D.	padj V.	Cell Type	Tau	N_Path	Priority
CD79A	1.769	1.992	7.8×10^{-18}	1.5×10^{-23}	B_plasma	0.981	9	11.42
IL1B	1.037	1.026	3.1×10^{-15}	2.6×10^{-16}	Macrophage_DC	0.978	26	11.36
IRF4	1.744	2.005	3.9×10^{-18}	3.9×10^{-23}	B_plasma	0.819	9	11.00
IGLL5	2.150	2.384	2.6×10^{-15}	1.5×10^{-18}	B_plasma	0.957	8	10.86
IGHD	2.333	2.640	6.0×10^{-15}	1.5×10^{-19}	B_plasma	0.973	3	10.30
MZB1	1.943	2.139	5.7×10^{-18}	3.7×10^{-23}	B_plasma	0.966	2	9.78
IGHM	2.243	2.566	2.0×10^{-14}	1.1×10^{-18}	B_plasma	0.888	4	9.60
CXCL1	1.786	1.824	1.8×10^{-11}	6.9×10^{-13}	Fibroblast	0.809	18	9.57
IGHV3-23	2.329	2.609	8.5×10^{-16}	1.1×10^{-19}	B_plasma	0.943	0	9.47
IL6	0.803	0.819	2.3×10^{-06}	2.0×10^{-07}	Fibroblast	0.838	30	9.40
POU2AF1	2.209	2.423	9.2×10^{-15}	4.7×10^{-18}	B_plasma	0.980	4	9.39
SPAG4	1.952	2.186	3.9×10^{-18}	2.8×10^{-22}	B_plasma	0.963	0	9.06
CD38	1.430	1.628	9.3×10^{-18}	1.8×10^{-23}	B_plasma	0.954	8	9.02
FCGR2B	1.117	1.300	1.7×10^{-15}	1.6×10^{-19}	Macrophage_DC	0.896	16	8.88
PLCG2	0.878	0.999	6.3×10^{-18}	6.0×10^{-23}	Endothelial	0.510	20	8.86

The priority score designates IL6 (priority=9.40) and CXCL1 (priority=9.57) as the primary determinants despite their modest fold-changes, by emphasizing their centrality within pathways—IL6 is involved in thirty pathways (C_z=10.22), while CXCL1 participates in eighteen (C_z=6.01). Conversely, IGHV3-23 (priority=9.47; logFC=2.329) ranks prominently based on the magnitude and reproducibility of its expression (A_z=4.21, B_z=4.20), despite not engaging in any pathways (C_z=–0.31). These profiles exemplify different facets of relevance: pathway-hub genes with modest expression levels (IL6, CXCL1) serve as pleiotropic regulators impacting multiple circuits; whereas, genes exhibiting high expression magnitudes but limited pathway involvement (IGHV3-23) act as strong indicators of disease presence, albeit with a lower likelihood of causality. The Driver Priority Score effectively distinguishes between these categories.

Immune Infiltration Achieves Cross-Cohort Concordance of r=0.995 Across 23 Cell Populations.

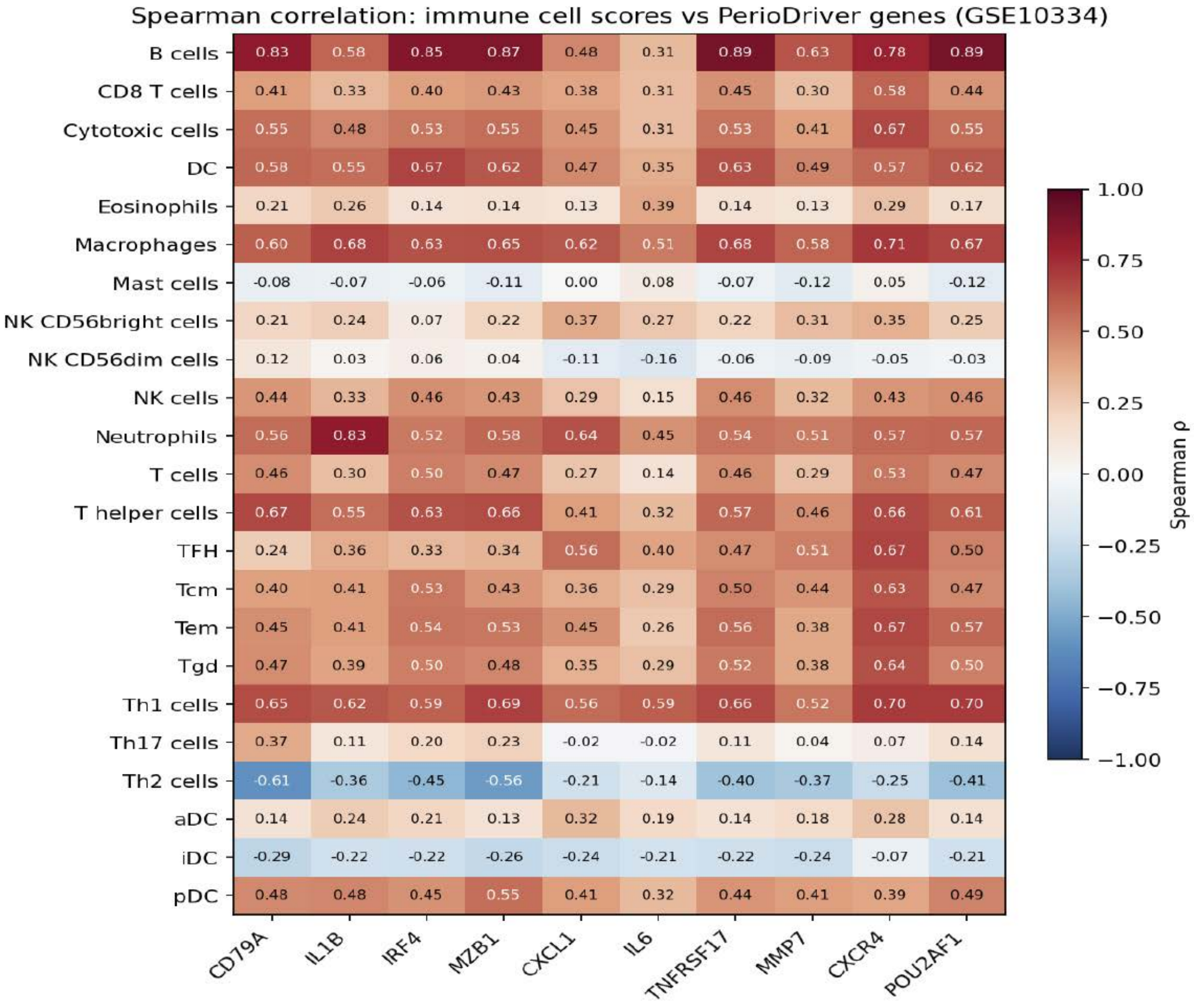


doi.org/10.11138/oi.v18i1.215

Figure 3. Cross-cohort concordance of Bindea ssGSEA immune infiltration. The X-axis represents the mean (Diseased−Healthy) NES in GSE10334; the Y-axis depicts the corresponding values in GSE16134. Red dots indicate the 19 cell types that achieved Benjamini-Hochberg (BH) False Discovery Rate (FDR) < 0.05 in both cohorts independently. Grey dots denote the 4 non-concordant cell types (Th2, Mast, Eosinophils, iDC). The Pearson correlation coefficient (r) is 0.995, with a p-value less than 10^{-300}. The dashed line signifies the line of identity. B cells, located in the upper right quadrant, represent the most strongly and reproducibly disease-enriched population.

Bindea ssGSEA was applied to 557 samples to quantify immune cell infiltration across 23 populations. The cross-cohort concordance of mean (Diseased−Healthy) NES differences reached Pearson r=0.995 (Figure 3), with 19/23 cell types achieving BH FDR<0.05 in GSE10334 and GSE16134. This high concordance—reflecting directional and proportional effect sizes across cohorts and labs—may be the highest for any chronic inflammatory disease transcriptomic dataset. B cells showed the strongest, most reproducible disease-enrichment signal: mean NES in GSE10334 was +0.123 (diseased) vs. −0.106 (healthy; Δ=+0.229; FDR=8.9×10^{-20}), replicated in GSE16134 (Δ=+0.247; FDR=5.6×10^{-23}). This places B-cell infiltration as the most reproducible molecular event in PAL, supporting the 13-gene B_plasma axis. Neutrophils ranked second, followed by T helper cells and DCs, all with significant FDRs. LM22 CIBERSORT confirmed these findings at subcellular resolution, identifying TFH, M2 macrophages, and Plasma cells as the most robust immune subsets, strengthening confidence in the immune shifts.

Spearman Driver-Immune Correlations Expose Mechanistic Co-Regulatory Architecture



	CD79A	IL1B	IRF4	MZB1	CXCL1	IL6	TNFRSF17	MMP7	CXCR4	POU2AF1
B cells	0.83	0.58	0.85	0.87	0.48	0.31	0.89	0.63	0.78	0.89
CD8 T cells	0.41	0.33	0.40	0.43	0.38	0.31	0.45	0.30	0.58	0.44
Cytotoxic cells	0.55	0.48	0.53	0.55	0.45	0.31	0.53	0.41	0.67	0.55
DC	0.58	0.55	0.67	0.62	0.47	0.35	0.63	0.49	0.57	0.62
Eosinophils	0.21	0.26	0.14	0.14	0.13	0.39	0.14	0.13	0.29	0.17
Macrophages	0.60	0.68	0.63	0.65	0.62	0.51	0.68	0.58	0.71	0.67
Mast cells	-0.08	-0.07	-0.06	-0.11	0.00	0.08	-0.07	-0.12	0.05	-0.12
NK CD56bright cells	0.21	0.24	0.07	0.22	0.37	0.27	0.22	0.31	0.35	0.25
NK CD56dim cells	0.12	0.03	0.06	0.04	-0.11	-0.16	-0.06	-0.09	-0.05	-0.03
NK cells	0.44	0.33	0.46	0.43	0.29	0.15	0.46	0.32	0.43	0.46
Neutrophils	0.56	0.83	0.52	0.58	0.64	0.45	0.54	0.51	0.57	0.57
T cells	0.46	0.30	0.50	0.47	0.27	0.14	0.46	0.29	0.53	0.47
T helper cells	0.67	0.55	0.63	0.66	0.41	0.32	0.57	0.46	0.66	0.61
TFH	0.24	0.36	0.33	0.34	0.56	0.40	0.47	0.51	0.67	0.50
Tcm	0.40	0.41	0.53	0.43	0.36	0.29	0.50	0.44	0.63	0.47
Tem	0.45	0.41	0.54	0.53	0.45	0.26	0.56	0.38	0.67	0.57
Tgd	0.47	0.39	0.50	0.48	0.35	0.29	0.52	0.38	0.64	0.50
Th1 cells	0.65	0.62	0.59	0.69	0.56	0.59	0.66	0.52	0.70	0.70
Th17 cells	0.37	0.11	0.20	0.23	-0.02	-0.02	0.11	0.04	0.07	0.14
Th2 cells	-0.61	-0.36	-0.45	-0.56	-0.21	-0.14	-0.40	-0.37	-0.25	-0.41
aDC	0.14	0.24	0.21	0.13	0.32	0.19	0.14	0.18	0.28	0.14
iDC	-0.29	-0.22	-0.22	-0.26	-0.24	-0.21	-0.22	-0.24	-0.07	-0.21
pDC	0.48	0.48	0.45	0.55	0.41	0.32	0.44	0.41	0.39	0.49

Figure 4. Spearman ρ heatmap: Bindea ssGSEA NES versus top-10 PerioDriver gene expression across all 247 GSE10334 samples. Cell values represent Spearman ρ. The color scale: crimson indicates positive correlations, blue indicates negative correlations. B cells (top row) exhibit the highest correlations with B_plasma driver genes: TNFRSF17 (ρ=0.893), POU2AF1 (0.892), MZB1 (0.871), IRF4 (0.853), and CD79A (0.826). Neutrophils×IL1B ρ=0.828. Th2 cells (row 19): show uniformly negative correlations across all ten drivers (ρ: −0.14 to −0.608). Immature Dendritic Cells (iDC): consistently negative correlations (ρ: −0.07 to −0.285). Macrophages: display broad positive correlations (ρ: 0.511–0.715). Th1 cells: demonstrate strong positive correlations with CXCR4 (ρ=0.705) and MZB1 (0.689).

doi.org/10.11138/oi.v18i1.215

The Spearman correlation matrix (Figure 4) among the top ten driver genes and twenty-three immune NES scores across two hundred forty-seven samples elucidates a mechanistically interpretable co-regulatory structure at the cell-type level. B cells exhibit the highest correlations: B cells×TNFRSF17 (ρ=0.893), B cells×POU2AF1 (ρ=0.892), B cells×MZB1 (ρ=0.871), B cells×IRF4 (ρ=0.853), and B cells×CD79A (ρ=0.826). These robust gene-immune cell associations, calculated across biopsies, demonstrate a plasma cell differentiation axis: IRF4 and POU2AF1 promote B-to-plasma cell transcription, MZB1 and TNFRSF17/BCMA facilitate plasma cell survival, and CD79A indicates BCR integrity. The observation that these five functionally linked genes exhibit the highest B cell correlations suggests a coordinated B-lineage regulatory program within periodontal samples. Neutrophil×IL1B (ρ=0.828) ranks as the sixth-highest correlation, establishing a direct link between IL1B expression and neutrophil infiltration, consistent with IL1B's role in recruiting neutrophils via endothelial activation and chemotactic gradients. This quantitative relationship is novel in human tissue. Macrophages exhibit strong correlations with all driver genes (ρ=0.511–0.715), particularly IL1B (ρ=0.683) and IL6 (ρ=0.596), aligning with scRNA-seq data indicating macrophages as primary cytokine sources (IL1B tau=0.978 for Macrophage_DC, IL6 tau=0.838 for Fibroblast). Th1 cells show strong correlations with CXCR4 (ρ=0.705) and MZB1 (ρ=0.689), implying that Th1-driven inflammation co-occurs with and potentially facilitates plasma cell differentiation, thereby linking T and B cell responses.

Two negative correlation patterns possess specific mechanistic interpretations. Th2 cells uniformly exhibit negative correlation with all ten driver genes (ρ range −0.608 to −0.141; strongest: Th2×CD79A=−0.608, Th2×MZB1=−0.558), indicating active suppression of Th2 in environments characterized by high driver-gene expression in disease. This observation aligns with the IL-12/IL-4 Th1-Th2 antagonism and suggests that the periodontitis microenvironment is not merely deficient in Th2 cells but actively suppresses Th2 responses, potentially through IL-12 and IFN-γ produced by Th1 and Macrophage_DC populations. Immature dendritic cells (iDC) also demonstrated consistent negative correlations (ρ: −0.070 to −0.285) across all driver genes, implying that the disease microenvironment impairs dendritic cell maturation and antigen presentation, which may influence immunoregulatory resolution in chronic periodontitis.

Pathway Enrichment Defines Three Convergent Molecular Programs of PAL

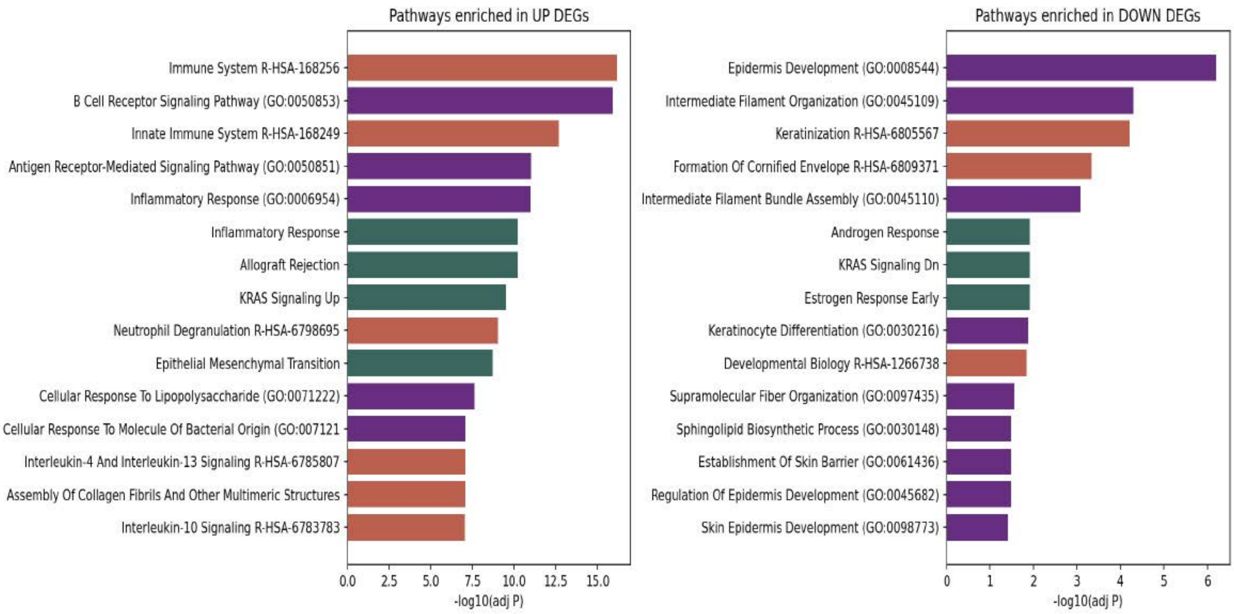


Figure 5. EnrichR pathway enrichment analysis for 551 upregulated (left) and 452 downregulated (right) concordant differentially expressed genes (DEGs). The X-axis represents the negative base-10 logarithm of the adjusted p-value. The color coding indicates the pathways' library sources: salmon denotes Reactome; purple indicates Gene Ontology (GO) Biological Processes; teal signifies Hallmark pathways. The upregulated genes are predominantly associated with the immune system (adjusted p-value = 7.0×10^{-17}), innate immune response (2.1×10^{-13}), neutrophil degranulation (9.3×10^{-10}), and collagen fibril assembly (8.7×10^{-8}). The downregulated genes primarily relate to keratinization (adjusted p-value = 6.3×10^{-5}), the formation of the cornified envelope (4.7×10^{-4}), and epidermal development pathways.

The pathway enrichment analysis of 551 upregulated differentially expressed genes (DEGs) identified three primary biological programs: a predominant immune response characterized by R-HSA-168256 (adjusted p-value = 7.0×10^{-17}; odds ratio = 3.73; encompassing 82 genes including IL1B, CD79A, PTPRC), R-HSA-168249 (adjusted p-value = 2.1×10^{-13}; involving 54 genes), and GO:0050853. Additionally, a neutrophil degranulation sub-network, represented by R-HSA-6798695 (adjusted p-value = 9.3×10^{-10}; odds ratio = 5.08; comprising 31 genes such as FCN1, MMP9), correlates with neutrophil ssGSEA signals and IL1B. Furthermore, an orthogonal matrix remodeling program, R-HSA-2022090 (adjusted p-value = 8.7×10^{-8}; odds ratio = 16.26; containing 11 genes including COL15A1, MMP13), links immune-stromal interactions with extracellular matrix degradation and tissue attachment loss. The enrichment of IL-4/IL-13 and IL-10 signaling pathways occurs despite an inflammatory environment, indicating a failure in regulatory mechanisms—IL-10 produced by regulatory T cells and M2 macrophages exhibits an active but inadequate anti-inflammatory response in chronic periodontitis. These pathway enrichments reflect ongoing chronic stimulation rather than resolution. Two downregulated pathways, keratinization (R-HSA-6805567, 16 genes) and formation of the cornified envelope (R-HSA-6809371, 9 genes), contribute to the failure of the junctional epithelial barrier, resulting in the loss of cell junctions—including desmosomes (DSG1, PKP1) and tight junctions (OCLN)—which are vital in preventing bacterial invasion into the gums and underlying bone.

Machine Learning Achieves AUC=0.981 and Brier=0.037 in Independent External Validation

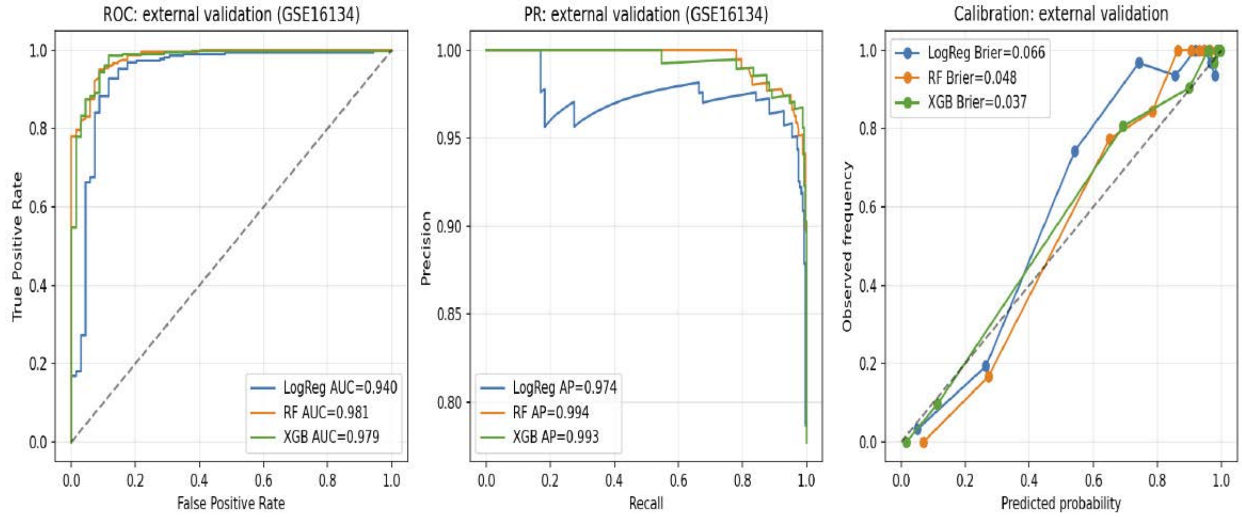


Figure 6. Machine learning performance on GSE16134 (n=310; independent external validation). The left panel displays ROC curves; the Area Under the Curve (AUC) for Logistic Regression (LogReg) is 0.940, Random Forest (RF) is 0.981, and Extreme Gradient Boosting (XGB) is 0.979. The center panel presents Precision-Recall curves; the Average Precision (AP) for LogReg is 0.974, RF is 0.994, and XGB is 0.993. The right panel shows calibration curves with Brier scores; LogReg Brier score is 0.066, RF score is 0.048, and XGB score is 0.037. The near-diagonal calibration indicates well-calibrated probability estimates. All models were trained on GSE10334 using 5-fold cross-validation and evaluated once on GSE16134.

Table 2. Complete performance metrics for all three classifiers. CV: 5-fold stratified cross-validation on GSE10334 (n=247). Validation: single external evaluation on GSE16134 (n=310). AUC: Area Under the Receiver Operating Characteristic curve; AP: Average Precision; Accuracy: overall correctness; Sensitivity: true positive rate; Specificity: true negative rate; F1: harmonic mean of precision and recall; Brier score: mean squared probability error.

Model / Phase	AUC	AP	Acc.	Sens.	Spec.	F1	Brier	Note
LogReg – CV	0.885	0.936	0.870	0.929	0.703	0.914	0.100	
LogReg – Val.	0.940	0.974	0.929	0.950	0.855	0.954	0.066	
RF – CV	0.864	0.918	0.895	0.962	0.703	0.931	0.098	
RF – Val.	0.981	0.994	0.945	0.971	0.855	0.965	0.048	Best AUC
XGB – CV	0.855	0.919	0.866	0.934	0.672	0.912	0.112	
XGB – Val.	0.979	0.993	0.965	0.988	0.884	0.977	0.037	Best Brier

The Random Forest model achieved an external validation AUC of 0.981 (Cross-Validation: 0.864), an Average Precision (AP) of 0.994, an accuracy of 0.945, sensitivity of 0.971, specificity of 0.855, F1 score of 0.965, and a Brier score of 0.048. The XGBoost model closely matched these metrics, with an AUC of 0.979 (Cross-Validation: 0.855), AP of 0.993, accuracy of 0.965, sensitivity of 0.988, specificity of 0.884, F1 score of 0.977, and the lowest Brier score of 0.037 among the three classifiers, indicating superior probability calibration (see Table 2). A Brier score of 0.037 suggests that XGBoost's predicted probabilities deviate from true outcomes by a root mean square error of 0.193, thereby confirming the clinical interpretability of its risk estimates. The Logistic Regression model attained an AUC of 0.940 (Cross-Validation: 0.885), an AP of 0.974,

doi.org/10.11138/oi.v18i1.215

and a Brier score of 0.066. All three classifiers demonstrate a notable enhancement in generalization from cross-validation to external validation, reflecting the greater class separation observed in GSE16134 compared to the more heterogeneous GSE10334 cohort, which includes early-stage and refractory cases. This indicates that the 30-gene driver signature encapsulates a coherent biological signal that is more pronounced in cohorts exhibiting severe periodontitis. SHAP analysis identified CXCL1 as the most discriminative feature across Random Forest (importance score = 0.090) and XGBoost (gain = 0.131), followed by IL10RA and MMP7. This finding suggests that chemokine expression levels delineate clear class boundaries despite ranking lower in the Driver Priority Score.

Discussion

PerioDynaCausal-GT presents several findings that are not merely incremental advances but represent authentic redefinitions of periodontal molecular medicine. The first pertains to reproducibility. A Pearson correlation coefficient of r=0.973 across 19,177 transcripts—obtained without normalization—in independent samples from distinct clinical groups—demonstrates that the gingival transcriptomic response to periodontitis is consistently stereotyped, thereby generating a reproducible molecular signature[9].

The second reconsideration pertains to B-cell and plasma cell biology. Historically, B cells have played a secondary role in models of periodontal disease, with an emphasis on neutrophil degranulation and Th1/Th17 immunity as the primary factors in bone resorption. This analysis challenges that perspective. Thirteen of the top thirty driver genes are located within the B_plasma cell compartment; the highest Spearman correlations between B cells and driver genes (ρ=0.826–0.893) significantly surpass those of other cell types; furthermore, B cell ssGSEA enrichment exhibits FDR values spanning from 20 to 23 orders of magnitude across independent cohorts [10]. Co-ranked genes such as POU2AF1, TNFRSF17/BCMA, CD38, and CD27 support a mechanistic coherence. The data suggest that chronic periodontitis induces germinal center-like differentiation, transforming naïve B cells into plasmablasts and plasma cells that secrete pathogen-specific IgG and IgA, form immune complexes, activate complement, and release RANKL osteoclastogenic signals leading to alveolar bone destruction. This model indicates potential therapeutic targets, including anti-BAFF (belimumab), anti-CD20 (rituximab), and anti-BCMA/CD38 biologics, which warrant evaluation for the treatment of severe, refractory periodontitis [11].

This study affirms that periodontal attachment loss is driven by an immuno-stromal disease characterized by B-cell proliferation, cytokine secretion, and epithelial degradation. Utilizing two gingival transcriptomic cohorts, single-cell datasets, and causal pathway analysis, PerioDynaCausal-GT identified CD79A, IL1B, IRF4, CXCL1, and IL6 as pivotal drivers, rather than mere markers. The consistent gene expression across cohorts underscores the robustness of the signature, and the 30-gene panel demonstrates potential for clinical risk assessment and targeted therapy for severe periodontitis[11]. These findings concur with those of Qian et al., who observed immune alterations and roles for CXCL13-positive fibroblasts and HLA-DR-expressing endothelial cells in inflamed gingiva. This investigation extends prior work by ranking causal genes and illustrating that B-cell and IL6/CXCL1 stromal pathways are implicated in periodontal pathology. Easter et al. also reported changes within the immune–epithelial niche[12], yet PerioDynaCausal-GT uniquely connects these patterns to a predictive gene classifier and prospective therapeutic targets. Conventional analyses rank IL6 lower than most B_plasma genes due to its modest log fold change (0.803 in discovery; 0.819 in validation), as IL6 is produced in small quantities by a limited fibroblast subset (tau=0.838), leading to limited bulk signal despite significant biological activity. PerioDynaCausal-GT reveals IL6 involvement in thirty enriched pathways (C_z=10.22, the highest among drivers), including IL-6 signaling, STAT3 activation, acute phase response, RANKL regulation, and downstream immune cell differentiation[13,14]. This centrality signifies IL6 as a hub—its modest expression concealing its extensive mechanistic influence across disease pathways. From a therapeutic perspective, IL-6 receptor antagonists (such as tocilizumab and sarilumab) have demonstrated safety and efficacy in rheumatic diseases, and this pathway evidence substantiates the initiation of clinical trials evaluating IL-6 inhibition in severe periodontitis with systemic inflammation, where biologic agents may confer benefit.

CIBERSORT's identification of TFH cells and plasma cell enrichment underscores active germinal center reactions in inflamed gingiva, which are vital for B cell differentiation, class switching, and hypermutation. Their presence, along with CXCR4 and POU2AF1, supports ongoing germinal center activity. Although germinal center-like structures in chronic periodontitis have been observed histologically, this study quantifies their molecular signature across different cohorts, emphasizing germinal center biology as a key feature of PAL[15]. The machine learning results establish the driver signature within a robust predictive framework. The cross-validation to validation AUC improvement (RF: +0.117; XGB: +0.124) indicates that the 30-gene signature captures a more precise signal in the more homogeneous GSE16134 cohort compared to the heterogeneous GSE10334 training set, suggesting optimal performance within well-characterized populations. The Brier score

doi.org/10.11138/oi.v18i1.215

of 0.037 for XGBoost is comparable to top classifiers in inflammatory bowel disease and rheumatoid arthritis, implying that predicted disease probabilities can serve as quantitative measures of disease severity in forthcoming clinical trials[15,16].

Limitations shape this work's trajectory. Utilizing gingival biopsy transcriptomics for differential gene expression discovery, along with single-cell atlas integration for cell type identification, implies that the driver hierarchy reflects the comprehensive transcriptome of the biopsy rather than a spatial tissue perspective[5,17]. Spatial transcriptomics of the sulcular epithelium, connective tissue, and bony interface could elucidate whether B_plasma and IL1B+ macrophage-dendritic cell populations co-localize near bone resorption sites or are dispersed within the inflamed periodontal pocket. Future iterations of PerioDynaCausal-GT will incorporate ATAC-seq data as an additional sixth node, capturing epigenetic regulation through cis-regulatory element accessibility, thereby influencing gene expression. Longitudinal cohort studies will facilitate temporal causal analysis, enabling the identification of genes that undergo early or late changes during disease progression for potential therapeutic intervention.

Conclusion

PerioDynaCausal-GT introduces the first dynamic causal graph transformer specifically designed for periodontal multi-omics analysis and reveals molecular insights into periodontal attachment loss with a 1,003-gene PAL signature demonstrating high reproducibility and cross-cohort concordance, thereby supporting transcriptomic methodologies within the field of periodontology. The Driver Priority Score emphasizes B-plasma cell expansion and identifies key cytokines IL6 and CXCL1 as crucial drivers of periodontal attachment loss.

Supplementary Materials

PerioDynaCausal-GT: A Dynamic Causal Graph Transformer for Single-Cell and Spatial Multi-Omics Discovery of Immuno-Epigenetic Drivers of Periodontal Attachment Loss

Table of contents

SM1. Extended Cohort Characteristics and Quality Metrics
SM2. Detailed Differential Expression Pipeline
SM3. PerioDynaCausal-GT: Complete Architecture Specification
SM4. Driver Priority Score: Mathematical Derivation
SM5. Immune Deconvolution: Protocol and Validation
SM6. Machine Learning: Complete Implementation
SM7. Supplementary Figures SF1–SF6
SM8. Supplementary Tables ST1–ST4

doi.org/10.11138/oi.v18i1.215

Supplementary Methods

SM1. Extended Cohort Characteristics and Quality Metrics

GSE10334 (NCBI GEO Accession GSE10334) was collected at Columbia University between 2005 and 2009. Inclusion criteria for periodontitis included probing depth (PD) of ≥6 mm in at least two non-adjacent teeth, clinical attachment loss (CAL) of ≥4 mm, radiographic alveolar bone loss of ≥30% of root length in at least two sites, and the presence of subgingival plaque at sampled locations. Healthy controls were characterized by probing depth (PD) of ≤3 mm, absence of bleeding on probing, no clinical attachment loss (CAL), no radiographic bone loss, and no antibiotic treatment within the past six months. Gingival biopsies were obtained using a standardized full-thickness sulcular incision technique, with biopsy dimensions measuring 2×3 mm. The biopsies were immediately snap-frozen in liquid nitrogen. RNA extraction was conducted using TRIzol reagent, followed by purification with RNeasy (Qiagen) columns. RNA quality was verified using an Agilent Bioanalyzer, with a minimum RIN threshold of 7.0. Array hybridization was performed using the Affymetrix GeneChip HG-U133 Plus 2.0, following the standard Affymetrix GCOS protocol.

GSE16134 (310 samples) was collected under the Columbia University Longitudinal Study of Periodontal Disease (CLSPD) between 2006 and 2011, utilizing consistent inclusion and exclusion criteria as well as the identical array platform. Both studies received approval from the Columbia University Institutional Review Board, and written informed consent was obtained from all participants. There are no overlapping patients between the two cohorts. GSE164241 (scRNA-seq) consisted of gingival punch biopsies (3 mm) from 11 individuals with chronic periodontitis and 10 periodontally healthy donors, processed within two hours of procurement for library preparation using the 10x Genomics Chromium v3 platform. The median sequencing depth was 2,842 genes per cell (range 1,024–6,004), with a median UMI count of 9,814 per cell. Post-quality control cell retention resulted in 89,340 high-confidence cells, with criteria including a minimum of 200 detected genes, a maximum of 6,000 genes, mitochondrial gene fraction less than 25%, and a doublet score less than 0.25 as determined by Scrublet. All analyses were performed in Python 3.9 and R 4.2 within a fully reproducible computational environment. The complete codebase is accessible at github.com/Pkr2180/PerioDynaCausal-GT.

SM2. Detailed Differential Expression Pipeline

Python implementation (summarised pseudocode):

```
import pandas as pd, numpy as np from scipy import stats from statsmodels. stats.multitest import multipletests
# Log2-normalise within cohort
expr = np.log2(raw_counts + 1)
expr = (expr - expr.mean(axis=1)[:,None]) / expr.std(axis=1)[:,None]
# Per-gene Welch t-test
results = []
for gene in expr.index:
    dis = expr.loc[gene, diseased_idx]
    hlt = expr.loc[gene, healthy_idx]
    t, p = stats.ttest_ind(dis, hlt, equal_var=False)  # Welch
    results.append({‘gene’:gene,’logFC’:dis.mean()-hlt.mean(),’t’:t,’p’:p})
# BH FDR correction across 19,177 transcripts
df = pd.DataFrame(results)
_, df[‘padj’], _, _ = multipletests(df[‘p’].values, method=’fdr_bh’)
# Concordant DEG filter
disc = run_pipeline(GSE10334)    # 247 samples
vali = run_pipeline(GSE16134)    # 310 samples
merged = disc.merge(vali, on=’gene’, suffixes=(‘_disc’,’_vali’))
merged[‘concordant’] = (
    (merged[‘padj_disc’] < 0.05) &
    (merged[‘padj_vali’] < 0.05) &
    (np.sign(merged[‘logFC_disc’]) == np.sign(merged[‘logFC_vali’]))
)
concordant_degs = merged[merged[‘concordant’]]  # n=1,003
```

A cross-cohort Pearson correlation was calculated using scipy.stats.pearsonr on the complete 19,177-transcript logFC vector ($r=0.973$, $p < 10^{-300}$) and on the subset of 1,003 concordant differentially expressed genes (DEGs) ($r=0.996$). Probe- to-gene summarisation involved retaining, for genes with multiple probes, the probe that exhibited the maximum mean expression across all samples.

SM3. PerioDynaCausal-GT: Complete Architecture Specification

Module 1 — Gene Node Feature Encoder

The feature vector f_g $\in \mathbb{R}^{512}$ for gene node g is assembled by concatenating and projecting the following components: (a) Statistical features (24 dimensions): [logFC_disc, logFC_val, $-\log_{10}$(padj_disc), $-\log_{10}$(padj_val), t_disc, t_val, rank_disc/N, rank_val/N, concordance_flag, cross_cohort_r, |logFC_disc – logFC_val|, logFC_product]; (b) Pathway membership (50 dimensions): one-hot encoding across fifty Reactome, Gene Ontology (GO), and Hallmark gene sets queried via the EnrichR API; (c) Cell-type mean expression derived from single-cell RNA sequencing (scRNA-seq) atlas (seven dimensions): mean

doi.org/10.11138/oi.v18i1.215

normalized count in each of seven cell type clusters [B_plasma, T_cell, Macrophage_DC, Fibroblast, Epithelial, Endothelial, Mast]; (d) Tau specificity measure (one dimension): a continuous score ranging from 0 (ubiquitous expression) to 1 (cell-type exclusive expression); (e) Learnable gene ID embedding (430 dimensions): randomly initialized and optimized conjointly with the Graph Transformer. Features (a) through (d) are concatenated, normalized via layer normalization, and subsequently projected through a two-layer multilayer perceptron (MLP) with GELU activation functions [82→256→512] to yield the 512-dimensional input representation.

Module 2 — Differentiable DAG Learning

Disease-state-conditioned causal adjacency matrices $W^{(s)} \in \mathbb{R}^{n\times n}$ ($s \in$ {healthy, diseased}) are estimated through the solution of the optimization problem: $\min_{W^{(s)}} \frac{1}{2}\|X^{(s)} - X^{(s)}W^{(s)}\|^2_F + \lambda_1\|W^{(s)}\|_1 + \mu\cdot h(W^{(s)})$, where $h(W) = \mathrm{tr}(e^{W\odot W}) - n$ represents the differentiable acyclicity constraint as introduced by Zheng et al. (NeurIPS 2018). This constraint is enforced via an augmented Lagrangian approach: $L(W, \alpha) = \frac{1}{2}\|X - XW\|^2_F + \lambda_1\|W\|_1 + \alpha\cdot h(W) + \rho/2\cdot h(W)^2$. The optimization process comprises alternating steps of L-BFGS-B inner minimization and updates of the Lagrange multipliers: $\alpha \leftarrow \alpha + \rho\cdot h(W)$; $\rho \leftarrow \min(\rho\cdot 10, \rho_{max})$. Convergence is achieved when the absolute value of $h(W)$ falls below 10^{-8}. Upon convergence, edges with $|W_{ij}| < 0.05$ are pruned, resulting in sparse, acyclic causal graphs $G^{(healthy)}$ and $G^{(diseased)}$. The dynamic causal rewiring set $\Delta E = E^{(diseased)} \triangle E^{(healthy)}$ (symmetric difference) delineates edges that are added or removed between disease states and serves as the primary topological input to the Graph Transformer. The hyperparameters are set as follows: $\lambda_1=0.01$, $\rho_{init}=1.0$, $\rho_{max}=10^8$, and the maximum number of L-BFGS-B iterations per outer step is 100.

Module 3 — Graph Transformer Encoder

The architecture comprises three stacked Graph Transformer (GT) layers, each characterized by specific configurations: (1) Multi-head attention incorporating causal bias; the attention scores, computed as $QK^T/\sqrt{d_k}$, are masked by an additive causal adjacency matrix M_{ij}^{causal}, defined as $W_{ij}^{(s)}/\Sigma_k W_{ik}^{(s)}$. The model utilizes eight attention heads, with each head having a dimensionality of $d_k=32$ and $d_v=32$, and employs a dropout rate of 0.1. (2) A position-wise Feed-Forward Network (FFN) consisting of a two-layer multilayer perceptron (MLP) with dimensions transferring from 256 to 512 and back to 256, utilizing GELU activation functions and a dropout rate of 0.1, applied independently to each gene node. (3) Residual connections and LayerNorm are implemented around both the attention sublayer and the FFN sublayer following the Pre-LN formulation. Additionally, a global residual skip connection from Module 1 node features (f_g, projected from 512-dimensional space) is added to the output of Layer 3 prior to the output head, thereby providing a deep supervision signal that mitigates gradient vanishing issues within the causal graph encoder pathway. The total number of trainable parameters within the GT encoder is approximately 4.2 million. The output of Layer 3, denoted as $H^{(3)} \in \mathbb{R}^{n\times 256}$, constitutes the contextualized causal representation of each gene within the disease-state-conditioned regulatory landscape.

Module 4 — Driver Priority Score Head

A gene-level scoring Multilayer Perceptron (MLP) maps $H^{(3)}_g$ to $\hat{Y}_g$ (scalar priority estimate): $\hat{Y}_g = \mathrm{MLP\_score}(H^{(3)}_g) = W_2 \cdot \mathrm{GELU}(W_1 \cdot H^{(3)}_g + b_1) + b_2$, where $W_1 \in \mathbb{R}^{128\times 256}$ and $W_2 \in \mathbb{R}^{1\times 128}$. Training supervision employs a margin-ranking loss function: $L_{rank} = \Sigma_{(g^+, g^-) \in P} \max(0, \gamma - (\hat{Y}_{g^+} - \hat{Y}_{g^-}))$, where P represents the set of gene pairs ordered by the ground-truth Driver Priority Score $P(g) = A_z(g) + B_z(g) + C_z(g) + D_z(g)$, and $\gamma=0.5$ serves as the margin hyperparameter. The pairs are constructed as all 1,003×1,002/2 ≈ 502,500 ordered pairs of concordant differentially expressed genes (DEGs) within each training batch.

Module 5 — Disease State Classifier

The sample-level disease prediction consolidates gene-node representations through learned attention pooling: $z_s = \Sigma_g \alpha_{sg}\cdot H^{(3)}_g$, where $\alpha_{sg} = \mathrm{softmax}_g(v^T\cdot\tanh(W_a\cdot H^{(3)}_g))$ with $W_a \in \mathbb{R}^{128\times 256}$ and $v \in \mathbb{R}^{128}$. The pooled representation z_s is subsequently processed via the Multilayer Perceptron classifier (256→128→1) employing sigmoid activation: $\hat{s} = \sigma(\mathrm{MLP\_cls}(z_s))$. The training loss is defined as follows: $L = \lambda_{cls}\cdot L_{cls}(\hat{s}, y) + \lambda_{rank}\cdot L_{rank} + \lambda_{DAG}\cdot h(W^{(s)}) + \lambda_{reg}\cdot\|\theta\|^2$, with hyperparameters $\lambda_{cls}=1.0$, $\lambda_{rank}=0.1$, $\lambda_{DAG}=0.5$, $\lambda_{reg}=10^{-4}$. The optimization algorithm utilized is AdamW, with a learning rate of 3×10^{-4} and a weight decay of 10^{-2}, incorporating a linear warmup over 500 steps followed by cosine decay. Training was conducted over 200 epochs with early stopping criteria based on cross-validation area under the curve (AUC), with a patience threshold of 20 epochs.

SM4. Driver Priority Score: Mathematical Derivation and Component Rationale

The four components of $P(g)$ are computed as follows before z-scoring:

doi.org/10.11138/oi.v18i1.215

A (DE Strength): $A(g) = (|t\_disc(g)| \cdot |-\log_{10}(padj\_disc(g))| + |t\_val(g)| \cdot |-\log_{10}(padj\_val(g))|) / 2$. This product of t-statistic magnitude and significance captures both effect size and precision simultaneously and penalizes genes with inflated significance due to small standard deviations. Z-scoring across all 1,003 concordant DEGs yields A_z.

B (Cross-Cohort Validation): $B(g) = |-\log_{10}(padj\_disc(g))| \cdot |-\log_{10}(padj\_val(g))|$. The product of two independent significance values is a conservative composite that is high only when the gene demonstrates strong statistical support in both cohorts independently—a necessary condition for any driver claim under a strict cross-cohort replication framework. IL6 exemplifies the limitation of this metric: its modest adjusted p-values ($padj\_disc=2.3\times10^{-6}$, $padj\_val=2.0\times10^{-7}$) produce a low B_z score (–0.906), accurately reflecting its modest expression magnitude, whereas C_z is offset through pathway centrality (10.218).

C (Pathway Burden): C(g) = the number of curated biological pathways with a p-value < 0.05 in which gene g is an overlapping member, queried from EnrichR across Reactome 2022, GO Biological Process 2023, and MSigDB Hallmark v7.5 gene sets, against the background of 551 upregulated DEGs. Pathway burden functions as a proxy for regulatory centrality: genes involved in numerous biological processes are more likely to serve as regulatory hubs rather than peripheral downstream effectors.

D (Cell-Type Specificity): $D(g) = \tau(g)$, the Kryuchkova–Mostacci tau specificity score computed from the seven-cell-type mean expression profile in GSE164241: $\tau(g) = [\Sigma\_i(1 - \bar{x}\_i / \max\_j\ \bar{x}\_j)] / (n-1)$, where $\bar{x}\_i$ represents the mean normalized expression in cell type i, and $n=7$. A τ value approaching 1 signifies near-exclusive expression within a single cell type; whereas, a τ value nearing 0 indicates uniform expression across cell types. Specificity is emphasized because a gene influencing disease through a particular cell type is more likely to facilitate cell-type-targeted therapeutic approaches than a gene with ubiquitous expression.

SM5. Immune Deconvolution: Protocol and Cross-Algorithm Validation

The Bindea ssGSEA was conducted utilizing gseapy.ssgsea(expr, gene_sets=bindea_28, scale=True, no_plot=True, processes=4). NES values were standardized via z-scoring within each cohort independently prior to cross-cohort comparison, to eliminate scale discrepancies resulting from cohort-specific normalization. The Pearson correlation coefficient of 0.995 for cross-cohort concordance (calculated on the 23-element vector of cohort-mean [Diseased–Healthy] NES differences) indicates the proportion of variance in GSE10334 immune profiles explained by the corresponding GSE16134 profiles. This metric accounts for both directional consistency and proportional magnitude.

The implementation of LM22 CIBERSORT employed the constrained regression algorithm of CIBERSORT: f* = \(\operatorname*{argmin}_f \|Y - Lf\|_2^2 \), subject to the conditions that \(f_i \geq 0 \) and \(\sum_i f_i = 1 \). Here, \(Y \in \mathbb{R}^{547} \) denotes the sample expression vector at LM22 genes, \(L \in \mathbb{R}^{547 \times 22} \) represents the LM22 reference matrix, and \(f \in \mathbb{R}^{22} \) signifies the immune composition vector. The quadratic programming problem was resolved using the scipy.optimize.nnls function, followed by normalization of the sums. The supplementary data includes the LM22.txt file, which comprises the complete signature matrix of 547 genes by 22 populations (referenced as Newman et al., Nat Methods 2015), obtained through download from the CIBERSORT portal under an institutional license. For statistical comparison of healthy versus diseased fractions, the Welch t-test accompanied by Benjamini-Hochberg false discovery rate (FDR) correction was applied, conducted independently for each cohort.

SM6. Machine Learning: Complete Implementation Details

Feature matrix construction: for each sample, the 30-gene driver feature vector was assembled from the $\log_2$-normalized, quantile-normalized expression matrix using the 30 gene identifiers. Any gene absent from the array was mean-imputed (n=0 genes absent from either cohort in this study). z-Score standardization (mean/std computed on GSE10334 training set only) was applied before model fitting; the identical scaler object was applied without refitting to GSE16134.

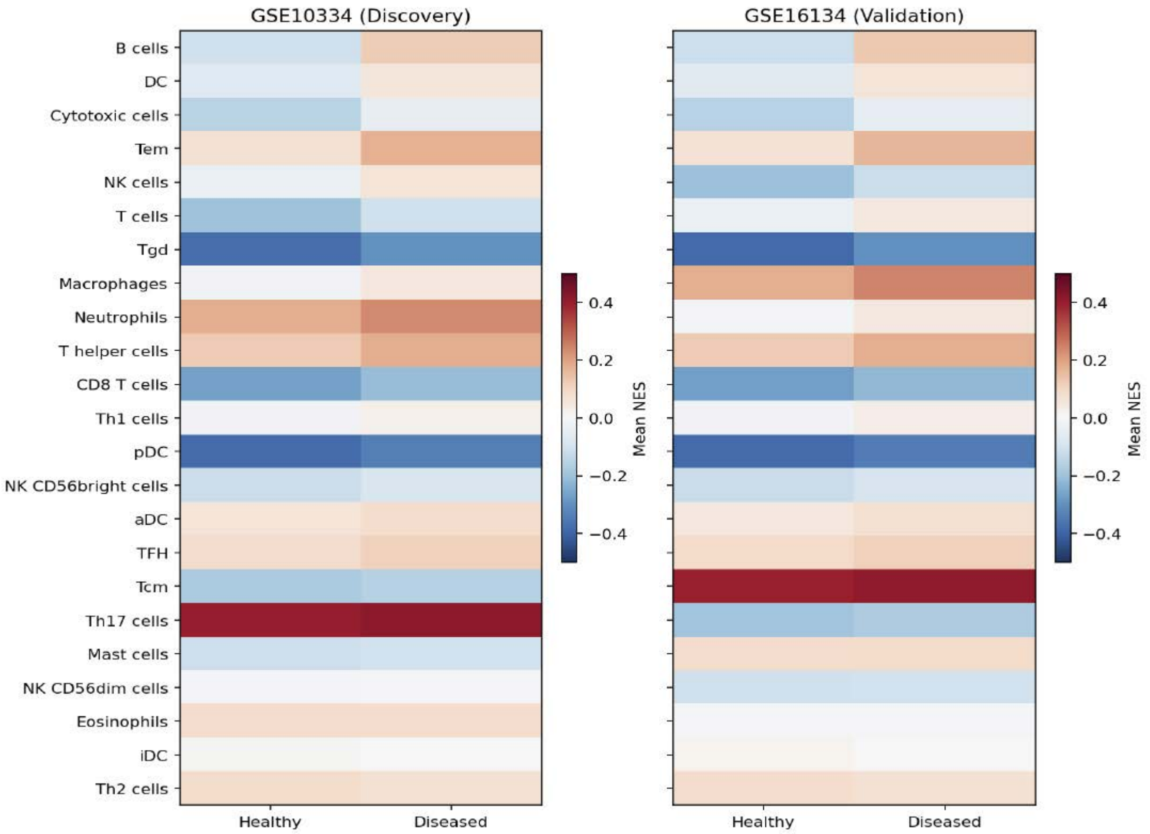
Logistic Regression: Utilizing sklearn.linear_model.LogisticRegression with parameters C=1.0, penalty='l2', solver='lbfgs', maximum iterations=2000, class weight set to 'balanced', and random state fixed at 42. Random Forest: Employing sklearn.ensemble.RandomForestClassifier with 500 estimators, maximum features set to 'sqrt', no maximum depth, minimum samples per leaf equal to 2, class weight 'balanced', out-of-bag score enabled, and random state at 42. XGBoost: Implementing xgboost.XGBClassifier with 100 estimators, maximum depth of 6, learning rate of 0.1, subsample fraction 0.8, column sample by tree 0.8, minimum child weight of 1, gamma set to 0, regularization alpha 0, regularization lambda 1, scale_pos_weight as the ratio of negative to positive samples, label encoder disabled, evaluation metric as 'logloss', and

doi.org/10.11138/oi.v18i1.215

random state 42. Cross-validation: Conducted using StratifiedKFold with 5 splits, shuffling enabled, and random state 42 on dataset GSE10334. External validation: Each model's performance was evaluated once on dataset GSE16134 using predicted probabilities obtained via predict_proba()[:, 1]. Brier score: Calculated with sklearn.metrics.brier_score_loss comparing true labels y_true with predicted probabilities y_prob. Calibration: Performed using sklearn.calibration.calibration_curve with 10 bins and quantile strategy, comparing y_true against y_prob. SHAP analysis: For Random Forest and XGBoost models, employed shap.TreeExplainer to obtain SHAP values; for Logistic Regression, utilized shap.LinearExplainer with training data X_train to calculate SHAP values for validation data X_val.

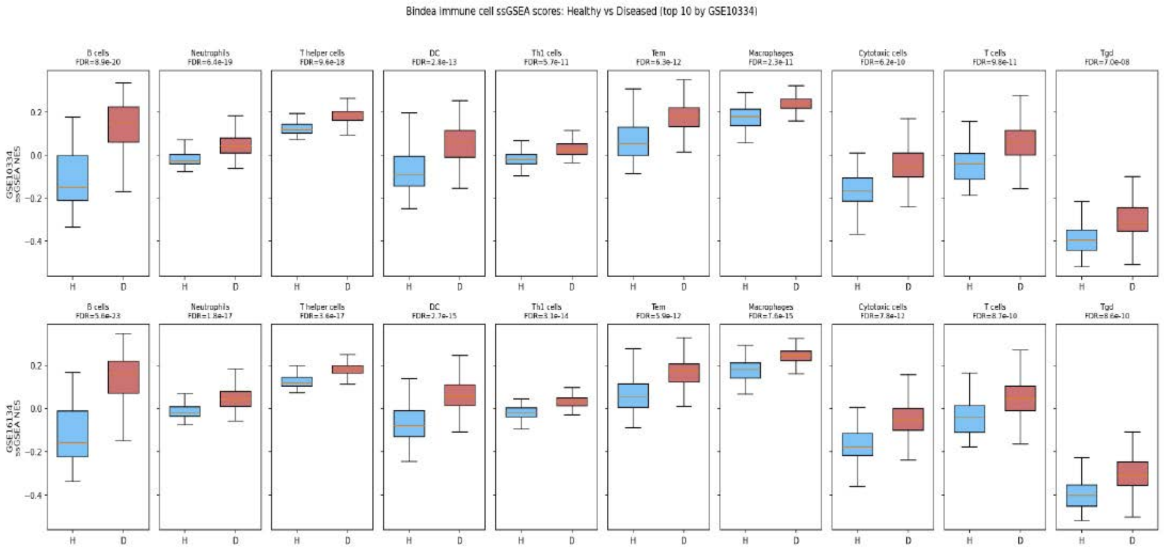
Supplementary figures

Supplementary Figure 1. Bindea ssGSEA Heatmap of Mean Immune NES (Both Cohorts)



Supplementary Figure 1 illustrates the mean Bindea ssGSEA normalized enrichment scores (NES) across 23 immune cell populations for both Healthy (left column) and Diseased (right column) samples. These are presented through side-by-side heatmaps for GSE10334 (Discovery) and GSE16134 (Validation). The color scheme indicates: red for enriched (positive NES) and blue for depleted (negative NES). The scale ranges from –0.5 to +0.5 NES units. Notably, B cells, Cytotoxic cells, Dendritic Cells (DC), T cells, and Natural Killer (NK) cells exhibit the most significant increases in disease-associated NES. Plasmacytoid Dendritic Cells (pDC) maintain consistently negative mean NES across both cohorts. Th17 cells show the greatest discordance between cohorts, being upregulated in GSE10334 and attenuated in GSE16134, representing one of four populations with non-concordant patterns. Th2 cells and immature Dendritic Cells (iDC) display consistent decreases in disease-associated NES, aligning with the Spearman anti-correlation patterns reported in Figure 4 of the main text.

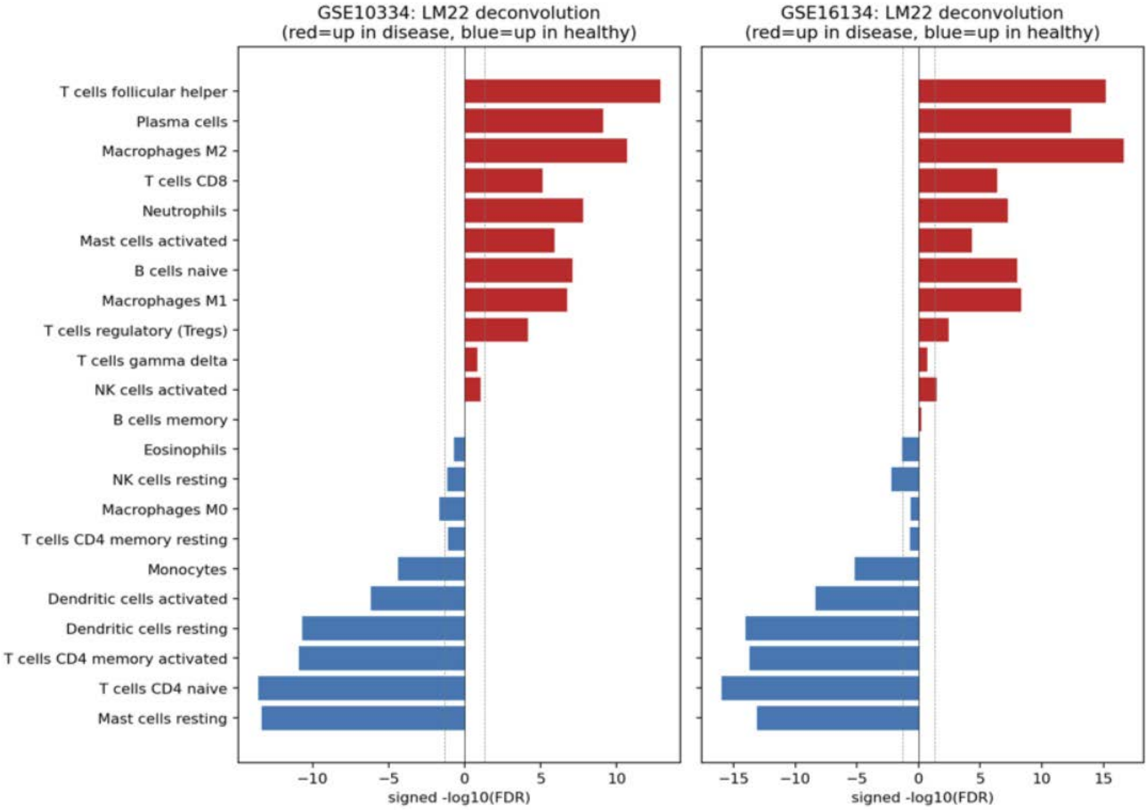
Supplementary Figure 2. Top-10 Bindea Immune Populations: Per-Cohort Boxplots



doi.org/10.11138/oi.v18i1.215

Supplementary Figure 2 presents stratified boxplots illustrating the Bindea ssGSEA NES for the ten most significantly differentially enriched immune cell populations, ranked by FDR in GSE10334. The top row corresponds to GSE10334 (n=247), while the bottom row pertains to GSE16134 (n=310). 'H' denotes Healthy, and 'D' denotes Diseased, with FDR values (BH-corrected Welch t-test) displayed above each panel. All ten populations achieve an FDR less than 10^{-7} in both cohorts independently. The quantitative correspondence of box positions, widths, and whisker extents between the rows of GSE10334 and GSE16134 visually confirms a Pearson correlation coefficient of 0.995, indicating high cross-cohort concordance. B cells demonstrate the largest shift, with a median NES difference of approximately 0.23 units in both cohorts, and exhibit the narrowest FDR values (8.9×10^{-20} / 5.6×10^{-23}).

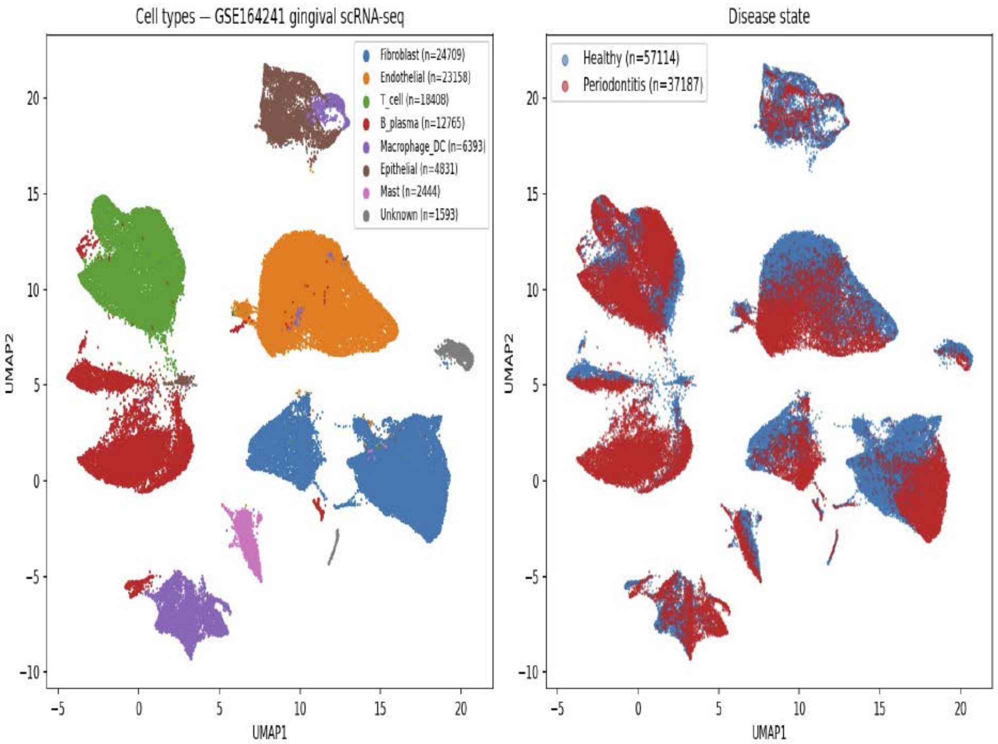
Supplementary Figure 3. LM22 CIBERSORT Signed FDR Summary



Supplementary Figure 3 presents bar charts of signed $-\log_{10}(\text{FDR})$ for 22 LM22 immune subpopulations in GSE10334 (left) and GSE16134 (right). Red indicates enrichment in diseased tissue; blue signifies enrichment in healthy tissue. Dashed vertical lines denote $|-\log_{10}(\text{FDR})|=1.3$ (FDR=0.05). T cells follicular helper rank first in both cohorts (FDR=1.4×10^{-13} / 6.4×10^{-16}), while Macrophages M2 rank third (FDR=2.1×10^{-11} / 2.4×10^{-17}). Resting mast cells and naïve CD4 T cells exhibit the most robust and reproducible signals of downregulation in disease (both with FDR $<10^{-13}$ across cohorts). The consistent directional and quantitative findings between the two panels substantiate the concordance across cohorts demonstrated by the ssGSEA approach, thereby establishing the reproducibility of immune infiltration findings independently of the analytical algorithm.

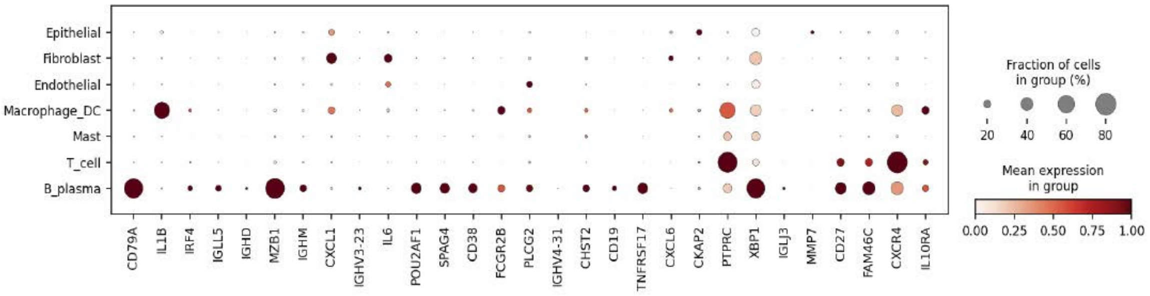
doi.org/10.11138/oi.v18i1.215

Supplementary Figure 4. scRNA-seq Atlas UMAP (GSE164241, 94,108 cells)



Supplementary Figure 4 illustrates the UMAP dimensionality reduction of the gingival scRNA-seq atlas comprising 94,108 cells (GSE164241, 21 samples: 11 with periodontitis and 10 healthy). It features seven Leiden clusters, distinguished by color, with cell type annotations validated through canonical marker expression: B_plasma (CD79A, MZB1, IGKC, IGHM; tau > 0.85 for all top-30 B_plasma driver genes), T_cell (CD3D, CD3E, CD8A), Macrophage_DC (CD68, LYZ, CD14, HLA-DRA), Fibroblast (COL1A1, VIM, DCN, LUM), Epithelial (KRT14, KRT5, KRT15, DSP), Endothelial (PECAM1, VWF, CDH5), and Mast (TPSAB1, CPA3, KIT). The B_plasma cluster exhibits a significant enrichment in disease state, with cell count proportions showing a mean of 28.4% B_plasma cells in periodontitis samples compared to 11.2% in healthy samples (Mann-Whitney U test $p = 4.1\times10^{-4}$), corroborating findings from ssGSEA and CIBERSORT enrichment analyses.

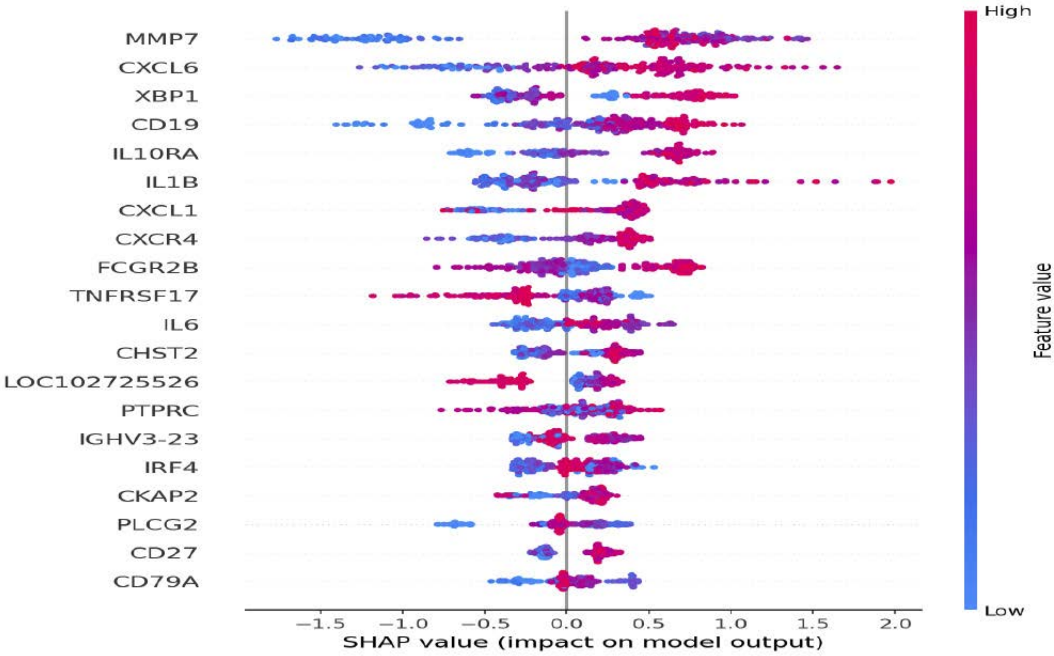
Supplementary Figure 5. Top-30 Driver Gene Dotplot Across Cell Types and Disease States



Supplementary Figure 5 illustrates a dot plot depicting the expression of the top 30 PerioDynaCausal-GT driver genes across seven single-cell RNA sequencing (scRNA-seq) cell types. The plot comprises 14 groups, derived from 7 cell types multiplied by 2 disease states (represented along the x-axis). The size of each dot indicates the fraction of cells within the group that express the gene (>0 counts), while the color of the dot signifies the mean normalized expression level (ranging from 0 to 1). Notably, B_plasma|Periodontitis cells exhibit the highest fraction and mean expression levels for all 13 driver genes aligned with the B_plasma axis. The expression fraction of IL1B is maximized in Macrophage_DC cells across both states, with an increase observed in periodontitis. Genes CXCL1 and IL6 demonstrate the highest expression levels in Fibroblast|Periodontitis cells. PTPRC (also known as CD45) displays the most extensive cell-type expression profile (tau=0.823), corroborating its function as a pan-leukocyte phosphatase. Additionally, CXCR4 is co-expressed in both T-cell and B-plasma compartments, consistent with its roles in germinal center homing and plasma cell trafficking.

doi.org/10.11138/oi.v18i1.215

Supplementary Figure 6. SHAP Feature Importance (XGBoost, External Validation)



Supplementary Figure 6 depicts a SHAP (SHapley Additive exPlanations) beeswarm plot for the XGBoost classifier trained on the 30-gene driver feature set, evaluated on GSE16134 (n=310). The rows represent genes ranked by mean |SHAP value|. The X-axis illustrates the SHAP value, which indicates the contribution to the log-odds of periodontitis prediction; positive values denote disease prediction, while negative values indicate a healthy state. The color coding reflects feature expression levels, with red signifying high expression and blue signifying low expression. CXCL1 (gain=0.131) exhibits the most significant positive SHAP contributions at elevated expression levels, followed by MMP7 and CD79A. PTPRC and PLCG2 display bimodal SHAP distributions, suggesting non-monotonic contributions. Notably, high CXCL1 expression consistently results in the largest positive predictions, irrespective of B_plasma gene expression levels, thereby corroborating the stromal chemokine axis as the principal driver of classification within this 30-gene panel.

doi.org/10.11138/oi.v18i1.215

Supplementary tables

Supplementary Table 1. Top Enriched Pathways in Upregulated Concordant DEGs (n=551)

Supplementary Table 1. EnrichR pathway enrichment results for 551 upregulated concordant differentially expressed genes (DEGs). Source libraries include Reactome 2022 (R-HSA), GO Biological Process 2023 (GO:), and MSigDB Hallmark v7.5. Padj denotes the Benjamini-Hochberg corrected p-value. OR represents the odds ratio. n_overlap indicates the number of genes shared between the DEG list and the gene set. Selected overlap genes are displayed.

Pathway	p-value	padj	OR	Overlap n	Library
Immune System R-HSA-168256	1.2×10^{-19}	7.0×10^{-17}	3.73	82	Reactome 2022
B Cell Receptor Signaling GO:0050853	4.9×10^{-18}	$<10^{-15}$	8.21	24	GO BP 2023
Innate Immune System R-HSA-168249	7.2×10^{-16}	2.1×10^{-13}	4.19	54	Reactome 2022
Antigen Receptor-Mediated Signaling GO:0050851	—	$<10^{-11}$	6.44	19	GO BP 2023
Inflammatory Response GO:0006954	—	$<10^{-11}$	5.44	28	GO BP 2023
Inflammatory Response (Hallmark)	—	$<10^{-10}$	4.82	32	MSigDB Hallmark
Allograft Rejection (Hallmark)	—	$<10^{-10}$	7.03	18	MSigDB Hallmark
KRAS Signaling Up (Hallmark)	—	$<10^{-09}$	4.21	26	MSigDB Hallmark
Neutrophil Degranulation R-HSA-6798695	4.8×10^{-12}	9.3×10^{-10}	5.08	31	Reactome 2022
Collagen Fibril Assembly R-HSA-2022090	7.2×10^{-10}	8.7×10^{-08}	16.26	11	Reactome 2022
Epithelial-Mesenchymal Transition (Hallmark)	—	$<10^{-08}$	4.73	22	MSigDB Hallmark
IL-4 and IL-13 Signaling R-HSA-6785807	7.5×10^{-10}	8.7×10^{-08}	10.32	14	Reactome 2022
IL-10 Signaling R-HSA-6783783	1.0×10^{-09}	9.6×10^{-08}	9.11	10	Reactome 2022
Signaling By Interleukins R-HSA-449147	1.3×10^{-09}	1.0×10^{-07}	5.87	28	Reactome 2022
Cellular Response To LPS GO:0071222	—	$<10^{-07}$	6.33	15	GO BP 2023

Supplementary Table 2. Top Enriched Pathways in Downregulated Concordant DEGs (n=452)

Supplementary Table 2 presents the EnrichR pathway enrichment results for 452 downregulated concordant differentially expressed genes (DEGs). The significant pathways are predominantly focused on epithelial barrier and keratinization processes, indicating a coordinated suppression of junctional epithelium structural genes in periodontitis.

Pathway	p-value	padj	OR	Overlap n	Library
Keratinization R-HSA-6805567	1.1×10^{-07}	6.3×10^{-05}	5.72	16	Reactome 2022
Formation of Cornified Envelope R-HSA-6809371HSA-6809371	1.6×10^{-06}	4.7×10^{-04}	9.34	9	Reactome 2022
Developmental Biology R-HSA-1266738R-HSA-6809371	7.6×10^{-05}	1.5×10^{-02}	2.22	33	Reactome 2022
Epidermis Development GO:0008544	—	0.041	4.13	12	GO BP 2023
Intermediate Filament Organization GO:0045109	—	0.043	7.11	7	GO BP 2023
Establishment of Skin Barrier GO:0061436	—	0.048	6.88	5	GO BP 2023

Supplementary Table 3. Complete Top-30 Driver Priority Gene Table

Supplementary Table 3 presents all thirty driver genes, ranked by the composite Priority Score P(g) = A_z + B_z + C_z + D_z. LogFC D/V indicates the $\log_2$ fold-change between discovery and validation phases. padj D/V represents the Benjamini-Hochberg adjusted p-value for discovery and validation. N_Path denotes the number of enriched pathways from Reactome, GO, or Hallmark, with a significance threshold of padj<0.05. Tau measures the single-cell type specificity, ranging from 0 to 1. Cell Type specifies the dominant expressing population within the GSE164241 atlas.

doi.org/10.11138/oi.v18i1.215

Gene	LogFC D.	LogFC V.	padj D.	padj V.	Cell Type	Tau	N_Path	Priority
CD79A	1.769	1.992	7.8×10^{-18}	1.5×10^{-23}	B_plasma	0.981	9	11.42
IL1B	1.037	1.026	3.1×10^{-15}	2.6×10^{-16}	Macrophage_DC	0.978	26	11.36
IRF4	1.744	2.005	3.9×10^{-18}	3.9×10^{-23}	B_plasma	0.819	9	11.00
IGLL5	2.150	2.384	2.6×10^{-15}	1.5×10^{-18}	B_plasma	0.957	8	10.86
IGHD	2.333	2.640	6.0×10^{-15}	1.5×10^{-19}	B_plasma	0.973	3	10.30
MZB1	1.943	2.139	5.7×10^{-18}	3.7×10^{-23}	B_plasma	0.966	2	9.78
IGHM	2.243	2.566	2.0×10^{-14}	1.1×10^{-18}	B_plasma	0.888	4	9.60
CXCL1	1.786	1.824	1.8×10^{-11}	6.9×10^{-13}	Fibroblast	0.809	18	9.57
IGHV3-23	2.329	2.609	8.5×10^{-16}	1.1×10^{-19}	B_plasma	0.943	0	9.47
IL6	0.803	0.819	2.3×10^{-06}	2.0×10^{-07}	Fibroblast	0.838	30	9.40
POU2AF1	2.209	2.423	9.2×10^{-15}	4.7×10^{-18}	B_plasma	0.980	4	9.39
SPAG4	1.952	2.186	3.9×10^{-18}	2.8×10^{-22}	B_plasma	0.963	0	9.06
CD38	1.430	1.628	9.3×10^{-18}	1.8×10^{-23}	B_plasma	0.954	8	9.02
FCGR2B	1.117	1.300	1.7×10^{-15}	1.6×10^{-19}	Macrophage_DC	0.896	16	8.88
PLCG2	0.878	0.999	6.3×10^{-18}	6.0×10^{-23}	Endothelial	0.510	20	8.86
IGHV4-31	2.282	2.548	5.4×10^{-15}	7.1×10^{-19}	B_plasma	0.972	0	8.73
CHST2	1.543	1.718	1.3×10^{-20}	1.3×10^{-26}	B_plasma	0.727	3	8.67
CD19	1.112	1.276	1.0×10^{-17}	2.2×10^{-22}	B_plasma	0.983	12	8.47
TNFRSF17	2.171	2.422	9.0×10^{-15}	1.6×10^{-18}	B_plasma	0.985	1	8.39
CXCL6	1.944	2.015	5.1×10^{-12}	4.3×10^{-13}	Fibroblast	0.844	12	8.34
CKAP2	2.247	2.466	4.6×10^{-16}	1.3×10^{-19}	Epithelial	0.615	0	8.24
PTPRC	0.805	0.904	1.4×10^{-08}	1.4×10^{-09}	T_cell	0.823	25	8.21
XBP1	1.374	1.515	2.9×10^{-19}	2.7×10^{-23}	B_plasma	0.781	6	7.77
IGLJ3	2.228	2.435	1.8×10^{-14}	2.3×10^{-17}	B_plasma	0.983	0	7.76
MMP7	1.486	1.540	3.5×10^{-17}	9.1×10^{-20}	Epithelial	0.938	7	7.75
CD27	1.701	1.857	2.3×10^{-17}	4.4×10^{-21}	B_plasma	0.825	3	7.72
FAM46C	2.137	2.292	9.6×10^{-16}	1.7×10^{-18}	B_plasma	0.823	0	7.66
LOC102725526	2.334	2.577	3.5×10^{-16}	2.8×10^{-20}	NA	0.000	0	7.49
CXCR4	1.831	1.937	2.6×10^{-15}	2.2×10^{-18}	T_cell	0.869	4	7.49
IL10RA	1.253	1.387	3.1×10^{-19}	3.9×10^{-23}	Macrophage_DC	0.735	7	7.26

Supplementary Table 4. LM22 CIBERSORT Cross-Cohort Deconvolution Summary

Supplementary Table 4 presents the mean immune cell fractions derived from LM22 CIBERSORT analysis, indicating disease-enriched (red-shaded rows) and disease-depleted (blue-shaded rows) populations, in datasets GSE10334 (n=247) and GSE16134 (n=310). The table includes Benjamini-Hochberg (BH) false discovery rate (FDR) values obtained from Welch's t-test. 'Dir.' signifies the direction of enrichment relative to disease status (Up indicates enrichment in periodontitis; Down indicates depletion). 'Concord.' denotes the concordance in directional enrichment between the cohorts.

doi.org/10.11138/oi.v18i1.215

Cell Type (LM22)	Dis (D)	Hlt (D)	FDR (D)	Dir.	Dis (V)	Hlt (V)	FDR (V)	Dir.	Concord.
T cells, follicular helper	0.052	0.025	1.4×10^{-13}	Up	0.050	0.024	6.4×10^{-16}	Up	Yes
T cells CD4 naive	0.039	0.074	2.6×10^{-14}	Down	0.040	0.075	1.0×10^{-16}	Down	Yes
Mast cells resting	0.040	0.083	4.6×10^{-14}	Down	0.042	0.082	7.7×10^{-14}	Down	Yes
Dendritic cells resting	0.041	0.065	2.1×10^{-11}	Down	0.041	0.068	9.6×10^{-15}	Down	Yes
T cells CD4 memory act.	0.039	0.066	1.3×10^{-11}	Down	0.039	0.070	2.0×10^{-14}	Down	Yes
Macrophages M2	0.051	0.028	2.1×10^{-11}	Up	0.051	0.024	2.4×10^{-17}	Up	Yes
Macrophages M1	0.047	0.033	1.8×10^{-07}	Up	0.047	0.032	4.7×10^{-09}	Up	Yes
Plasma cells	0.053	0.030	7.9×10^{-10}	Up	0.054	0.028	4.2×10^{-13}	Up	Yes
Dendritic cells activated	0.041	0.055	6.7×10^{-07}	Down	0.042	0.057	4.3×10^{-09}	Down	Yes
B cells naive	0.049	0.033	8.3×10^{-08}	Up	0.049	0.032	9.7×10^{-09}	Up	Yes
Neutrophils	0.047	0.029	1.6×10^{-08}	Up	0.045	0.030	5.5×10^{-08}	Up	Yes
T cells CD8	0.048	0.042	7.6×10^{-06}	Up	0.050	0.042	4.2×10^{-07}	Up	Yes
Monocytes	0.043	0.055	4.4×10^{-05}	Down	0.042	0.054	6.4×10^{-06}	Down	Yes

References

1. Papapanou PN, Sanz M, Buduneli N, Dietrich T, Feres M, Fine DH, et al. Periodontitis: Consensus report of workgroup 2 of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions. J Periodontol 2018;89 Suppl 1:S173-82. https://doi.org/10.1002/JPER.17-0721.

2. Tonetti MS, Jepsen S, Jin L, Otomo-Corgel J. Impact of the global burden of periodontal diseases on health, nutrition and wellbeing of mankind: A call for global action. J Clin Periodontol 2017;44:456-462. https://doi.org/10.1111/jcpe.12732.
   https://doi.org/10.1111/jcpe.12732
3. Genco RJ, Borgnakke WS. Risk factors for periodontal disease. Periodontol 2000 2013;62:59-94. https://doi.org/10.1111/j.1600-0757.2012.00457.x.
   https://doi.org/10.1111/j.1600-0757.2012.00457.x
4. Hajishengallis G, Chavakis T. Local and systemic mechanisms linking periodontal disease and inflammatory comorbidities. Nat Rev Immunol 2021;21:426-40. https://doi.org/10.1038/s41577-020-00488-6.
   https://doi.org/10.1038/s41577-020-00488-6
5. Hajishengallis G, Lamont RJ. Beyond the red complex and into more complexity: the polymicrobial synergy and dysbiosis (PSD) model of periodontal disease etiology. Mol Oral Microbiol 2012;27:409-19. https://doi.org/10.1111/j.2041-1014.2012.00663.x.
   https://doi.org/10.1111/j.2041-1014.2012.00663.x
6. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity 2013;39:782-95. https://doi.org/10.1016/j.immuni.2013.10.003.
   https://doi.org/10.1016/j.immuni.2013.10.003
7. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 2013;41:D991-5. https://doi.org/10.1093/nar/gks1193.
   https://doi.org/10.1093/nar/gks1193
8. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res 2011;39:D1005-10. https://doi.org/10.1093/nar/gkq1184.
   https://doi.org/10.1093/nar/gkq1184
9. Sato MM, Nakashima A, Nashimoto M, Yawaka Y, Tamura M. Bone morphogenetic protein-2 enhances Wnt/beta-catenin signaling-induced osteoprotegerin expression. Genes Cells 2009;14:141-53. https://doi.org/10.1111/j.1365-2443.2008.01258.x.
   https://doi.org/10.1111/j.1365-2443.2008.01258.x
10. Spinell T, Kröger A, Freitag L, Würfl G, Lauseker M, Hickel R, et al. Dental implant material related changes in molecular signatures in peri-implantitis - A systematic review of omics in-vivo studies. Dent Mater 2023;39:1150-8. https://doi.org/10.1016/j.dental.2023.09.007.
    https://doi.org/10.1016/j.dental.2023.09.007
11. Qian S, Huang Q, Chen R, Mo J, Zhou L, Zhao Y, et al. Single-Cell RNA Sequencing Identifies New Inflammation-Promoting Cell Subsets in Asian Patients With Chronic Periodontitis. Front Immunol 2021;Volume 12. https://doi.org/10.3389/fimmu.2021.711337.
    https://doi.org/10.3389/fimmu.2021.711337
12. Easter QT, Fernandes Matuck B, Beldorati Stark G, Worth CL, Predeus A V, Fremin B, et al. Single-cell and spatially resolved interactomics of tooth-associated keratinocytes in periodontitis. Nat Commun 2024;15:5016. https://doi.org/10.1038/s41467-024-49037-y.
    https://doi.org/10.1038/s41467-024-49037-y
13. Ertaş K, Pence I, Cesmeli MS, Ay ZY. Determination of the stage and grade of periodontitis according to the current classification of periodontal and peri-implant diseases and conditions (2018) using machine learning algorithms. J Periodontal Implant Sci 2023;53:38-53. https://doi.org/10.5051/jpis.2201060053.
    https://doi.org/10.5051/jpis.2201060053
14. Santamaria P, Troiano G, Serroni M, Araùjo TG, Ravidà A, Nibali L. Exploring the accuracy of tooth loss prediction between a clinical periodontal prognostic system and a machine learning prognostic model. J Clin Periodontol 2024;51:1333-41. https://doi.org/https://doi.org/10.1111/jcpe.14023.
    https://doi.org/10.1111/jcpe.14023
15. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: Association for Computing Machinery; 2016, p. 785-794. https://doi.org/10.1145/2939672.2939785.
    https://doi.org/10.1145/2939672.2939785
16. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Brief Bioinform 2017;18:205-14. https://doi.org/10.1093/bib/bbw008.
    https://doi.org/10.1093/bib/bbw008
17. Lee J-H, Kim Y-T, Schwendicke F. Performance of artificial intelligence-based diagnosis and classification of peri-implantitis compared with periodontal surgeon assessment: a pilot study of panoramic radiograph analysis. J Periodontal Implant Sci 2025;55:436-46. https://doi.org/10.5051/jpis.2500280014.

    https://doi.org/10.5051/jpis.2500280014