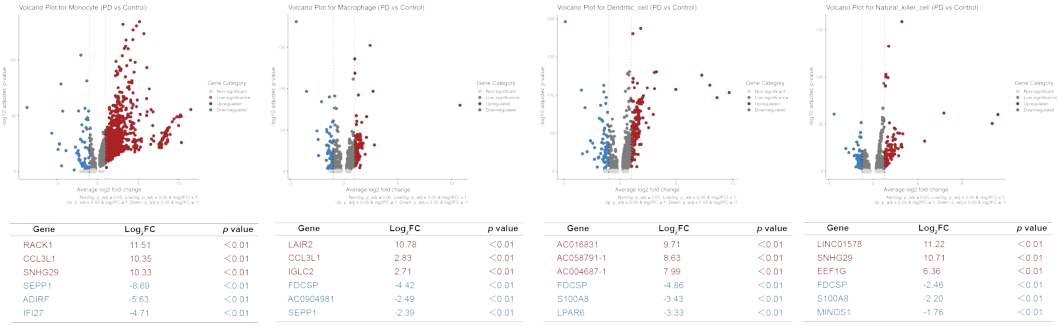
SUPPLEMENTARY DOCUMENTS

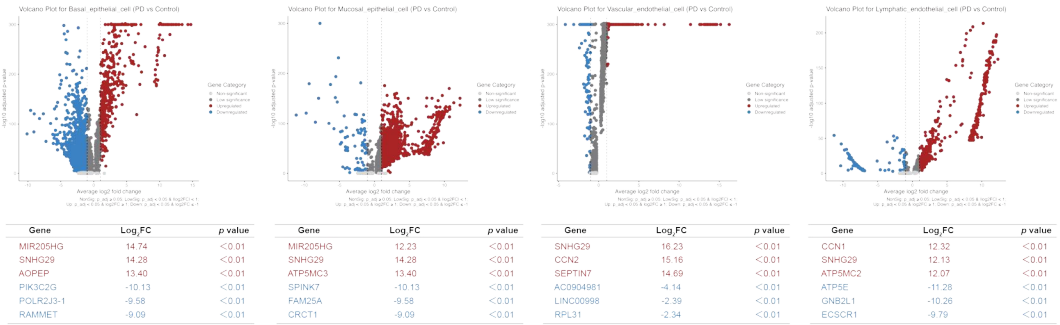


Gene	Log_2FC	*p* value
RACK1	11.51	<0.01
CCL3L1	10.35	<0.01
SNHG29	10.33	<0.01
SEPP1	-8.69	<0.01
ADIRF	-5.63	<0.01
IFI27	-4.71	<0.01

Gene	Log_2FC	*p* value
LAIR2	10.78	<0.01
CCL3L1	2.83	<0.01
IGLC2	2.71	<0.01
FDCSP	-4.42	<0.01
AC090498.1	-2.49	<0.01
SEPP1	-2.39	<0.01

Gene	Log_2FC	*p* value
AC016831	9.71	<0.01
AC058791-1	8.63	<0.01
AC004687-1	7.99	<0.01
FDCSP	-4.86	<0.01
S100A8	-3.43	<0.01
LPAR6	-3.33	<0.01

Gene	Log_2FC	*p* value
LINC01578	11.22	<0.01
SNHG29	10.71	<0.01
EEF1G	6.36	<0.01
FDCSP	-2.46	<0.01
S100A8	-2.20	<0.01
MINOS1	-1.76	<0.01

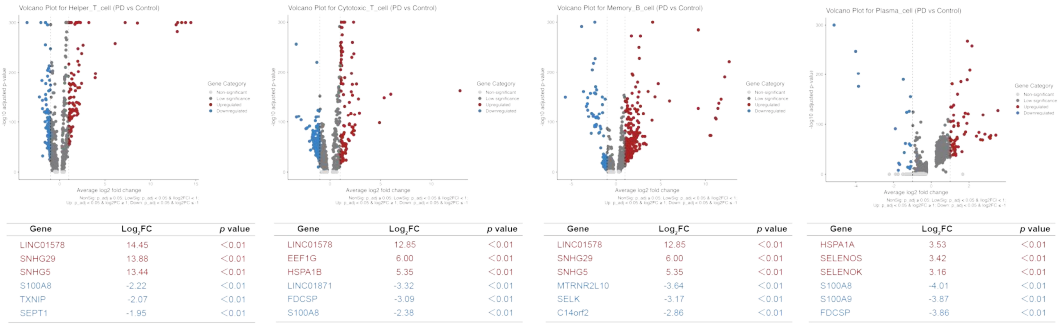


Gene	Log_2FC	*p* value
MIR205HG	14.74	<0.01
SNHG29	14.28	<0.01
AOPEP	13.40	<0.01
PIK3C2G	-10.13	<0.01
POLR2J3-1	-9.58	<0.01
RAMMET	-9.09	<0.01

Gene	Log_2FC	*p* value
MIR205HG	12.23	<0.01
SNHG29	14.28	<0.01
ATP5MC3	13.40	<0.01
SPINK7	-10.13	<0.01
FAM25A	-9.58	<0.01
CRCT1	-9.09	<0.01

Gene	Log_2FC	*p* value
SNHG29	16.23	<0.01
CCN2	15.16	<0.01
SEPTIN7	14.69	<0.01
AC090498.1	-4.14	<0.01
LINC00998	-2.39	<0.01
RPL31	-2.34	<0.01

Gene	Log_2FC	*p* value
CCN1	12.32	<0.01
SNHG29	12.13	<0.01
ATP5MC2	12.07	<0.01
ATP5E	-11.28	<0.01
GNB2L1	-10.26	<0.01
ECSCR1	-9.79	<0.01

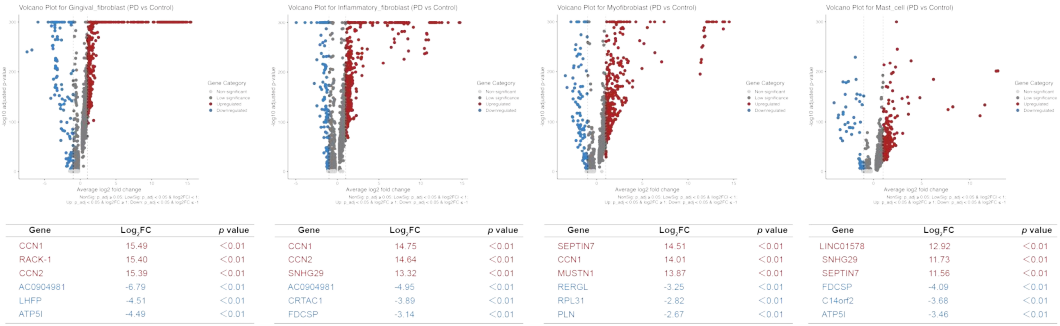


Gene	Log_2FC	*p* value
LINC01578	14.45	<0.01
SNHG29	13.88	<0.01
SNHG5	13.44	<0.01
S100A8	-2.22	<0.01
TXNIP	-2.07	<0.01
SEPT1	-1.95	<0.01

Gene	Log_2FC	*p* value
LINC01578	12.85	<0.01
EEF1G	6.00	<0.01
HSPA1B	5.35	<0.01
LINC01871	-3.32	<0.01
FDCSP	-3.09	<0.01
S100A8	-2.38	<0.01

Gene	Log_2FC	*p* value
LINC01578	12.85	<0.01
SNHG29	6.00	<0.01
SNHG5	5.35	<0.01
MTRNR2L10	-3.64	<0.01
SELK	-3.17	<0.01
C14orf2	-2.86	<0.01

Gene	Log_2FC	*p* value
HSPA1A	3.53	<0.01
SELENOS	3.42	<0.01
SELENOK	3.16	<0.01
S100A8	-4.01	<0.01
S100A9	-3.87	<0.01
FDCSP	-3.86	<0.01



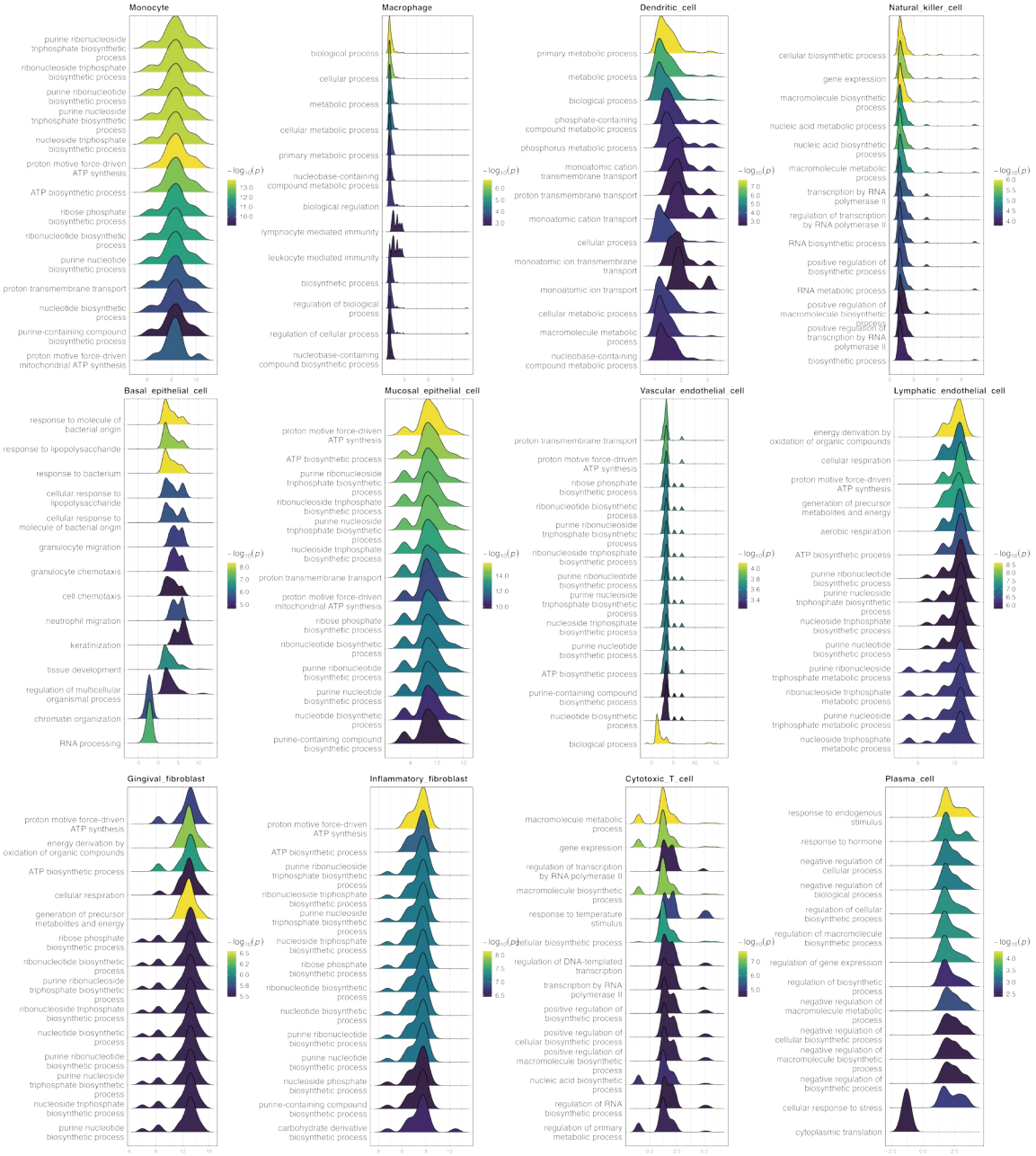
Gene	Log_2FC	*p* value
CCN1	15.49	<0.01
RACK1	15.40	<0.01
CCN2	15.39	<0.01
AC090498.1	-6.79	<0.01
LHFP	-4.51	<0.01
ATP5I	-4.49	<0.01

Gene	Log_2FC	*p* value
CCN1	14.75	<0.01
CCN2	14.64	<0.01
SNHG29	13.32	<0.01
AC090498.1	-4.95	<0.01
CRTAC1	-3.89	<0.01
FDCSP	-3.14	<0.01

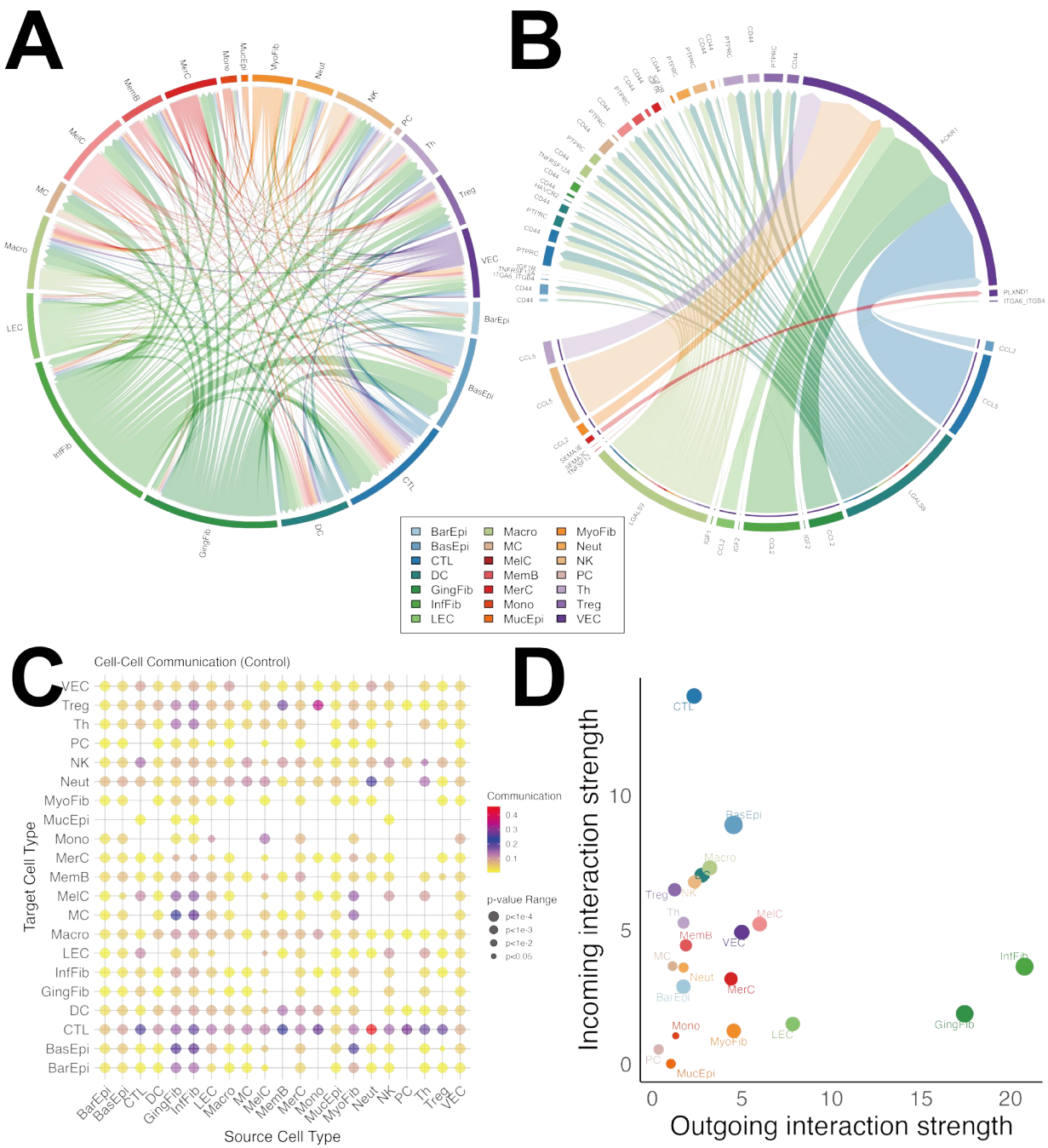
Gene	Log_2FC	*p* value
SEPTIN7	14.51	<0.01
CCN1	14.01	<0.01
MUSTN1	13.87	<0.01
RERGL	-3.25	<0.01
RPL31	-2.82	<0.01
PLN	-2.67	<0.01

Gene	Log_2FC	*p* value
LINC01578	12.92	<0.01
SNHG29	11.73	<0.01
SEPTIN7	11.56	<0.01
FDCSP	-4.09	<0.01
C14orf2	-3.68	<0.01
ATP5I	-3.46	<0.01

Supplementary Figure S1: Cell-type-specific transcriptional rewiring in PD.
Volcano plots depict DEGs across 16 gingival cell populations, with log_2 fold change on the x-axis and $-log_{10}$ *p*-value on the y-axis. Vertical dashed lines mark ±1 log_2 FC, and the horizontal dashed line marks $p = 0.01$. Points exceeding these thresholds are highlighted (upregulated in red, downregulated in blue), and inset tables list the three most significantly up- and downregulated genes (red and blue text, respectively) for each cell type.

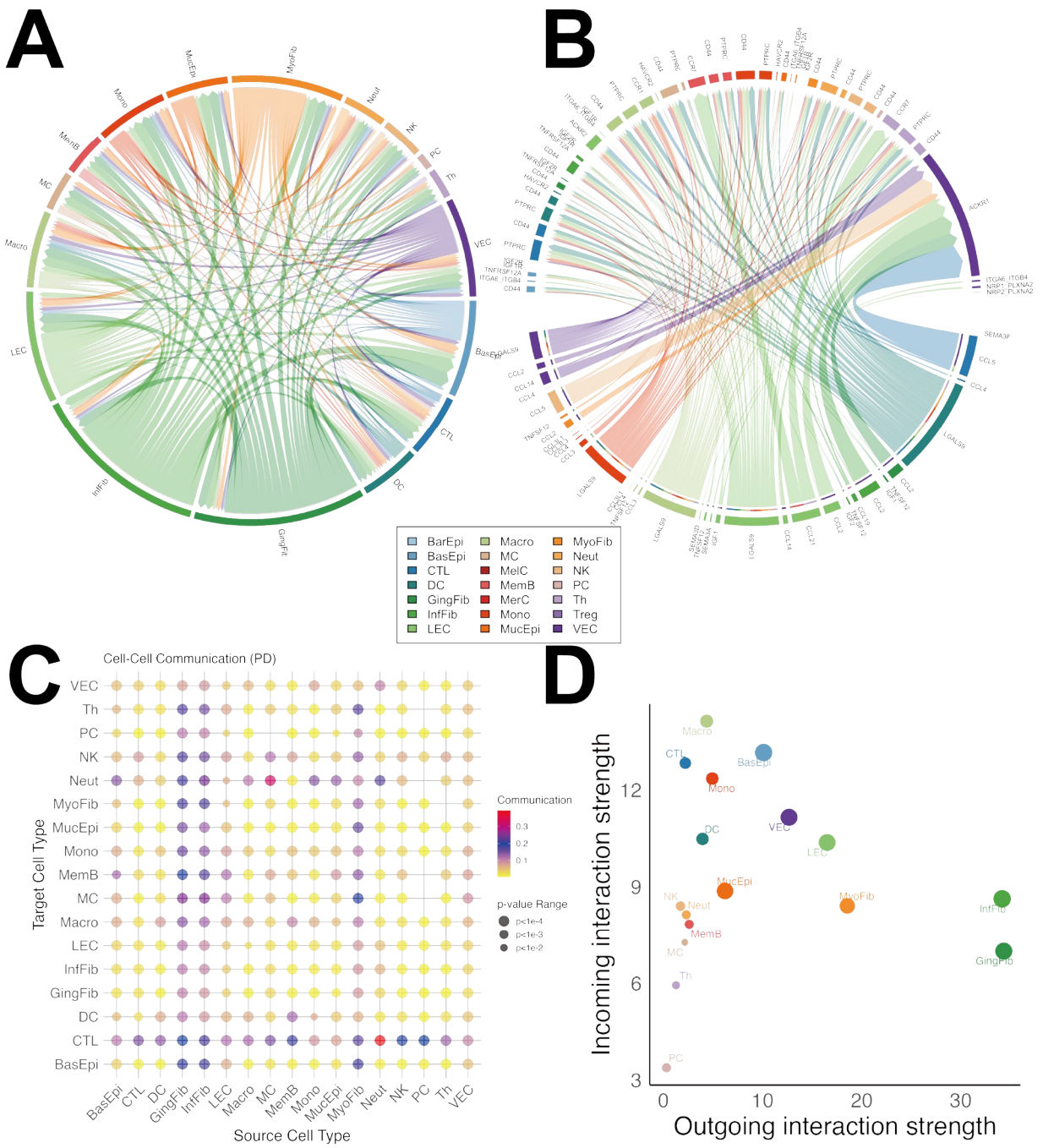


Supplementary Figure S2: GSEA GO-BP enrichment ridges across 12 cell populations. Each ridge plot displays the distribution of enrichment significance (-$\log_{10}$ *p*-value from gseGO, BP ontology) for the top ten GO terms per cell type; warmer hues denote more substantial enrichment. The Y axis displays the top 10 most significantly enriched GO:BP as determined by the GSEA analysis. Terms are typically ordered by significance from high to low. The X axis represents the distribution density of a gene-level statistic. Positive values on the x-axis indicate gene upregulation, while negative values indicate gene downregulation, based on the pre-ranked gene list used for GSEA. The position of the peak within each ridge is indicative of the overall trend of avg_ $\log_2$ FC for the genes in that pathway.



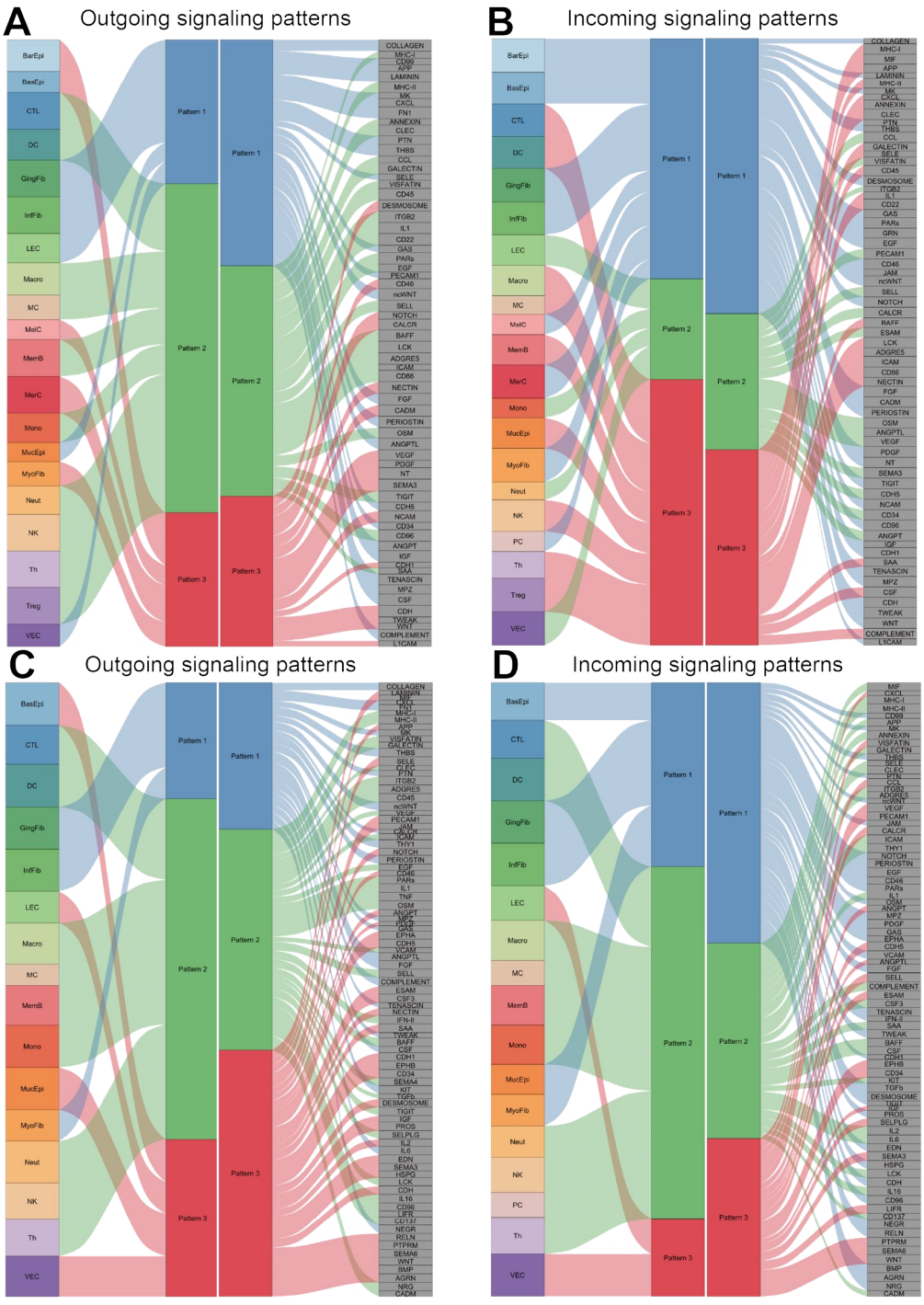
Supplementary Figure S3: Intercellular communication network in healthy gingiva.

A. Chord diagram showing net changes in outgoing communication strength between all cell-type pairs. Sector widths represent total outgoing signal; chord thickness reflects pairwise Δweight. **B.** Gene-level Subnetworks: Five exemplar pathways in healthy gingiva, highlighting the origin and target of key ligand-receptor interactions. **C.** Dot plots of average communication probability for each source-target pair. Dot size encodes the $-\log_{10}$ p-value; color scale indicates mean interaction strength. **D.** Signaling-role scatter plots of total outgoing versus incoming interaction strength per cell type. Each point represents a cell-type sender; color denotes source identity. Deviation above or below the diagonal indicates gain or loss of net signaling influence.



Supplementary Figure S4: Intercellular communication network in PD gingiva.

A. Chord diagram showing net changes in outgoing communication strength between all cell-type pairs. Sector widths represent total outgoing signal; chord thickness reflects pairwise Δweight. **B.** Gene-level Subnetworks: Five exemplar pathways in PD gingiva, highlighting the origin and target of key ligand-receptor interactions. **C.** Dot plots of average communication probability for each source-target pair. Dot size encodes the -$\log_{10}$ p-value; color scale indicates mean interaction strength. **D.** Signaling-role scatter plots of total outgoing versus incoming interaction strength per cell type. Each point represents a cell-type sender; color denotes source identity. Deviation above or below the diagonal indicates gain or loss of net signaling influence.



Supplementary Figure S5: Individual signaling patterns in the healthy and PD gingiva. **A, B.** Riverplots of global signaling patterns in healthy gingiva: Outgoing (A, left) and incoming (B, right) pathways clustered into three coherent modules. **C, D.** Riverplots of global signaling patterns PD gingiva: Outgoing (C, left) and incoming (D, right) riverplots recapitulate the same three feed-forward loops now amplified in disease. Patterns are colored consistently (blue = immune recruitment, green = matrix remodeling, red = vascular niche activation).

MATERIALS & METHODS

1. Single-cell data integration, quality control, and harmonization

1.1 Raw data sources

All the scRNA-seq datasets of gingival biopsies from chronic PD patients and periodontally healthy controls published before Jan 2025 were obtained as feature-barcode count matrices (Covid-10 Cell Atlas, GSE164241, GSE152042, GSE207502)[11-13,65], yielding 60,972 cells from PD samples and 54,209 cells from control samples.

1.2 Initial quality control

Each library was converted to a Seurat object (Seurat v5.2.1) and subjected to two rounds of stringent filtering before and after merging the control and PD cohorts separately. The second round filtering was performed using the following thresholds: Features per cell >200 & < 5,000; Mitochondrial transcript fraction < 10 %; Ribosomal transcript fraction < 40 %. Violin and jitter plots of these metrics before versus after filtering confirmed effective elimination of low-quality cells.

1.3 Data Harmonisation and Batch Integration

To ensure consistency between raw counts and normalized data layers across different libraries, the RNA assay layers (counts and data) were merged via JoinLayers in Seurat. Optional regression of cell-cycle effects was performed using G_2/M and S phase gene sets. Batch correction was then carried out with Harmony (Harmony v1.2.3) on the first 20 principal components (PCs), the axes of greatest variance identified via principal component analysis (PCA). Thereby, donor identity, library chemistry, and sequencing depth were specified as covariates and regressed out. Successful mitigation of batch effects was assessed by plotting uniform manifold approximation and projection (UMAP) embeddings colored by original batch, by computing silhouette widths.

1.4 Doublet Detection and Removal

DoubletFinder (v2.0.4) was applied in batches of 5,000 cells to both control and PD objects. Within each batch, parameter sweeping (paramSweep) and sweep summarization (summarizeSweep) were used to identify the optimal pK (the neighborhood size parameter for artificial doublet generation). Homotypic doublet rate correction was applied, and an expected doublet rate of 7.5 % was assumed based on cell loading densities. Cells classified as singlets were retained, and final singlet counts were recorded for downstream analysis.

2. Dimensionality reduction, Iterative clustering and multi-round annotation

Singlet-only matrices were normalized by log-transformation (LogNormalize, scale factor = 10,000), and highly variable genes (HVGs) were selected using the ‘vst’ method (nFeatures = 2,000). Data were scaled and subjected to PCA (20 PCs retained). Harmony-corrected PCs were used to compute UMAP projections.

Clustering was performed in Seurat at three nested resolutions to capture hierarchical organization: (i) global landscape, resolution = 0.2, yielding broad cell-lineage clusters; (ii) lineage-specific sub-clustering, resolution = 0.6, yielding transcriptionally coherent subclusters; and (iii) resolution = 1.0, resolving fine-grained immunophenotypes within lymphoid and myeloid branches.

For each clustering round, a tri-modal annotation strategy was employed: (a) canonical marker genes manually curated from CellMarker 2.0, (b) label transfer via Azimuth constrained to an oral mucosa reference, and (c) automated classification with SingleR (v2.8.0) against Human Primary Cell Atlas Data, Blueprint Encode Data, Monaco Immune Data, and Novershtern Hematopoietic

Data. Cellular identity labels were finalized by integrating automated classification outputs with expert curation of marker gene expression.

3. Global and Lineage-Specific Compositional Profiling

3.1 Multi-Dimensional Visualization

UMAP and t-distributed stochastic neighbor emulation (t-SNE) were computed on Harmony-corrected principal components to visualize the overall cellular landscape. Embeddings were colored by tissue source (healthy versus PD), donor, and final cell-type annotation, illustrating seamless batch mixing and preservation of biological heterogeneity. Additionally, for each group, the proportion of cells assigned to every high-resolution cell type was computed and summarized as stacked bar plots and donut charts. These visualizations provide an intuitive overview of the gingival microenvironment in health and disease.

3.2 Statistical Inference of Cellular Shifts

To identify statistically significant shifts in cell-type composition between PD and control cohorts, two complementary approaches were employed: (i) Pearson's chi-square tests on contingency tables of cell counts per type (Bonferroni-corrected $p < 0.05$), and (ii) individual Fisher's exact tests on each lineage with a false discovery rate (FDR) < 0.05. Effect sizes (odds ratios) were computed to quantify the expansion or contraction of specific populations.

4. High-Resolution Cell Type-Specific Differential Gene Expression

4.1 Differentially expressed genes (DEGs) within Annotated Cell Types

For each manually annotated cell type, single-cell DEGs between PD and control groups were identified using the Wilcoxon rank-sum test (FindMarkers, min.pct = 0.25, $\log_2$folder-changes(FC) > 0.25). Raw p-values were adjusted by the Benjamini-Hochberg correction, and genes with adjusted p-value < 0.05 were considered significant.

4.2 Volcano Plot Classification and Heatmap of Top DEGs

For each manually annotated cell type, DEGs were classified into four categories: Non-significant: adjusted $p \geq 0.05$; Low-significance: adjusted $p < 0.05$ and $|\text{avg\_log}_2\text{FC}| < 1$; Up-regulated: adjusted $p < 0.05$ and $\text{avg\_log}_2\text{FC} \geq 1$; Down-regulated: adjusted $p < 0.05$ and $\text{avg\_log}_2\text{FC} \leq -1$. Volcano plots were rendered in ggplot2 (v3.5.2) with $\log_2$(fold-change) on the x-axis and $-\log_{10}$(adjusted p-value) on the y-axis. Dashed vertical lines at ±1 and a horizontal line at $-\log_{10}(0.05)$ denote fold-change and significance thresholds. In each plot, the top 25 up-regulated and top 25 down-regulated genes (ranked by adjusted p-value) were highlighted.

4.3 Global Top50 Heatmap

To provide a global overview, a control vs. PD comparison was performed across all cells using Seurat's FindMarkers. Genes were ranked by the smallest adjusted p-value, and the top 50 were selected. The mean normalized expression of these top 50 genes in each cohort was computed via AggregateExpression and visualized as a heatmap using ComplexHeatmap.

4.4 Dot-Plot Summary and Data Export

The 50 most significant DEGs were displayed as dot plots, both by group and by cell type, respectively. They encode both average expression and the percentage of cells expressing each gene. Complete DEG tables show gene name, $\text{avg\_log}_2\text{FC}$, p-value, adjusted p, category label, and a composite volcano_score ($|\text{avg\_log}_2\text{FC}| \times -\log_{10}$(adjusted p)). They were exported as CSV files in a dedicated DEG_Lists directory for downstream pathway and machine-learning analyses.

5. Multi-Faceted Functional and Compositional Profiling

5.1 Immune Microenvironment Composition

Cell-type proportions were computed from the contingency table of group × cell type and converted to percentages. Stacked bar plots and complementary donut charts were drawn in ggplot2.

Immune activation markers (IFNG, TNF, IL6, PDCD1) were retrieved via Seurat's FetchData and compared between the control and PD cohorts by two-sided Wilcoxon rank-sum tests. Violin plots with overlaid jittered points were generated, and exact p-values were computed for each marker. Immune checkpoint genes (PDCD1, CTLA4, LAG3) were visualized on uniform manifold approximation and projection (UMAP) embeddings via FeaturePlot.

5.2 Transcription Factor Regulatory Inference (DoRothEA + VIPER)
Human transcription factor-target interactions were first filtered for confidence levels A and B from the DoRothEA database and converted into a regulon object. Per-cell transcription factor activity scores were then inferred via the VIPER algorithm on the log-normalized RNA assay. Cohort-average activity scores were computed for each factor, and the 20 most variable transcription factors (ranked by variance) were visualized in a heatmap using ComplexHeatmap. Pairwise Pearson correlations among these top factors were thresholded at $|r| \geq 0.5$ and displayed as an undirected network via igraph (v2.1.4), with node colors encoding the difference in mean activity between PD and control. The five transcription factors exhibiting the highest variance were further examined by violin and jitter plots, and their activity differences were tested by two-sided Wilcoxon rank-sum tests. Finally, the inferred activity scores and network topology metrics were incorporated as features in downstream machine-learning models (Random Forest, XGBoost) to enhance disease-state prediction.

5.3 Metabolic Pathway Scoring
Single-gene violin plots were generated for glycolysis markers (HK1, HK2, PFKL, ALDOA, GAPDH, PGK1, ENO1, PKM), tricarboxylic acid (TCA) cycle markers (CS, ACLY, IDH3A, OGDH, SUCLG1, SDHA), and the top 10 mitochondrial transcripts (genes beginning with MT-), each tested by Wilcoxon rank-sum. Module scores for Glycolysis, TCA, and Mitochondrial signatures were calculated with AddModuleScore. Violin plots of each module score were drawn, and FeaturePlot was used to map module scores onto the UMAP embedding. Cohort-average module scores were aggregated via FetchData and visualized as a heatmap with ComplexHeatmap.

5.4 Cell Death Pathway Profiling
Violin and jitter plots were produced for apoptosis markers (CASP3, CASP8, BAX, BCL2), pyroptosis markers (GSDMD, CASP1, IL1B), and ferroptosis markers (GPX4, ACSL4, SLC7A11), with Wilcoxon p-values computed for each. A heatmap of average expression for all cell-death-related genes was generated via AverageExpression and ComplexHeatmap to survey cohort differences in cell-death programs.

5.5 Extracellular Matrix Remodeling Analysis
Single-gene violin plots were generated for ECM remodeling markers (MMP2, MMP9, COL1A1, FN1, ITGB1) and tested by Wilcoxon rank-sum. An ECM remodeling module score was computed via AddModuleScore, visualized both on UMAP embeddings (FeaturePlot) and as violin plots. Mean ECM scores per cohort were summarized in a heatmap to quantify matrix remodeling in PD.

6. Cell-Cell Communication Network Analysis
To dissect the intercellular signaling architecture underpinning gingival homeostasis and PD, we leveraged the CellChat framework (CellChat v1.6.1; CellChatDB.human) for comprehensive ligand-receptor inference and network topology analysis.

6.1 Construction of Group-Specific CellChat Objects

The final integrated Seurat object was stratified by cohort and subsetted into control (n = 54,209 cells) and PD (n = 60,972 cells) metadata and normalized RNA assays. The human ligand-receptor database was assigned, and low-abundance cell types and interactions involving fewer than five cells were removed by 'filterCommunication' to ensure robust inference.

6.2 Inference of Overexpressed Genes and Interactions

Within each CellChat object, overexpressed ligands and receptors were identified, and putative interactions were enumerated. Communication probabilities were computed and aggregated at the pathway level, yielding quantitative matrices of interaction strength (probability) and statistical significance (p-value). Only interactions with $p < 0.05$ were retained for downstream analysis.

6.3 Network Aggregation and Topological Characterization

Aggregated signaling networks were constructed, summarizing inter-cluster communication by pathway. Global network topology was interrogated through centrality metrics, quantifying each cell type's incoming and outgoing signaling importance. Signaling-role bar charts, scatter plots, and heatmaps of outgoing/incoming roles were generated to compare cohort-specific shifts in intercellular crosstalk.

6.4 Multi-Scale Visualization of Signaling Patterns

- Chord Diagrams: Aggregate chord plots illustrated the overall connectivity among cell types for the top 20 most active pathways, with layout optimized to minimize overlap and node/edge labels suppressed for clarity. Ligand-receptor-level chord diagrams highlighted the top 15 receptor-ligand pairs per cohort, emphasizing pathway composition and relative interaction strength.
- Alluvial River Plots: Communication patterns (outgoing vs. incoming) were classified into $k = 3$ modules, and alluvial diagrams traced how cell types distribute their signaling outputs and inputs across these modules, revealing conserved versus disease-specific communication modes.
- Bubble Plots: The top 30 pathway-specific cell-type interactions were visualized as bubble charts, encoding mean communication probability by dot size and merging large families to simplify interpretation.
- Cell-Cell Scatter Plots: Pairwise communication between every source-target cell-type pair was summarized in ggplot2 scatter plots: dot color scaled to mean probability and dot size binned by p-value range, enabling simultaneous appraisal of strength and statistical support.
- Sankey Diagrams: Alluvial flows were further distilled into three-axis Sankey plots (Group-Source-Target) using ggalluvial, with node fills drawn from our custom color palette to emphasize shifts in dominant signaling routes between cohorts.
- Word Clouds: To capture pathway and cell-pair prominence at a glance, word clouds of pathway names and of Source-Target pairs were generated from summed communication probabilities, highlighting the most active signaling axes in Control and PD.

All visualizations were exported at 300 dpi with transparent backgrounds to facilitate seamless integration into multi-panel figures. This multi-angled interrogation of intercellular signaling delineates both conserved homeostatic circuits and their reconfiguration in PD, yielding a systems-level blueprint of gingival immunobiology.

7. Single-Cell Trajectory Inference and Dynamic Gene Regulation

To delineate continuous cell-state transitions and uncover lineage-specific transcriptional dynamics in health and disease, we applied Monocle 3 (v1.2.9) pseudotime trajectory analysis to four key lineages: B cells, epithelial cells, myeloid cells, and T cells.

7.1 Lineage Subsetting and CDS Construction

For each lineage, cells with relevant annotations were subsetted. The raw UMI count matrix, per-cell metadata (including cohort and final cell-type labels), and gene annotations were encapsulated in a Monocle 3 cell data set (CDS), ensuring consistency across all four analyses.

7.2 Dimensionality Reduction and Principal Graph Learning

Each CDS underwent preprocessing with 20 principal components, followed by UMAP embedding to capture major axes of transcriptional variance. Cells were clustered using default Leiden parameters, and principal graphs were learned via reversed graph embedding, yielding a continuous manifold of cell states.

7.3 Root-Cell Specification and Pseudotime Ordering

To anchor developmental trajectories, root populations were defined for each lineage as previously described: Memory B cell for B cells[66], Basal epithelial cell for epithelial cells[67], Monocyte for myeloid cells[68], and Naive T cell for T cells[69]. Designating these well-characterized progenitor or naive states ensured that pseudotime ordering reflected bona fide developmental hierarchies.

7.4 Trajectory Visualization

Trajectories were overlaid on UMAP embeddings colored by (i) manual cell-type annotation, (ii) pseudotime value, and (iii) experimental cohort (control vs. PD), providing an intuitive depiction of both branching structure and disease-associated shifts.

7.5 Dynamic Gene Expression along Pseudotime

To capture transcriptional programs varying along each trajectory:

- Binning: Cells were partitioned into 200 equal-sized bins according to pseudotime.
- Aggregation: Within each bin and cohort, mean $\log_2$-normalized expression was calculated for all genes.
- Standardization: The combined bin gen matrix was row-scaled (z-score), and values were clipped at ±3 to enhance visualization.
- Heatmap Rendering: ComplexHeatmap was used to display cohort-specific and difference heatmaps, with genes ordered by variance or maximal fold-change.

7.6 Identification of Trajectory-Associated Genes

We applied Moran's spatial autocorrelation to pinpoint genes whose expression exhibits significant dependence on trajectory position (adjusted p-value < 0.05). Top candidates were visualized in smoothed trend plots, highlighting both lineage-restricted and shared regulatory programs.

7.7 Branchpoint Analysis

Key branchpoints, where trajectories diverge into alternative fates, were detected with Monocle 3's topology metrics. Differential expression between branches was assessed via generalized additive models (GAMs), revealing genes driving fate bifurcation. GO-term enrichment of branch-specific genes elucidated functional modules underlying lineage decisions.

Together, this comprehensive trajectory framework revealed continuous differentiation hierarchies, identified dynamic gene modules sculpting cell-state transitions, and highlighted perturbations in PD at unprecedented resolution.

8. Multi-Scale Functional and Pathway Enrichment (GO, KEGG & GSEA)

To interrogate the biological programs underpinning cell-type specialization and dynamic lineage progression, complementary over-representation and gene-set enrichment analyses on two orthogonal gene collections were performed: (i) cell-type-specific DEGs and (ii) pseudotime-associated genes along each inferred trajectory. All analyses employed the clusterProfiler framework (v4.14.6) with the Org.Hs.eg.db human reference (v3.20.0) and DOSE (v4.0.1).

8.1 Over-Representation Analysis of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG)

For each cell type and for each pseudotime lineage, up-regulated and down-regulated gene lists ($|\log_2 FC| > 0.2$, adjusted $p < 0.05$) were compiled against the background of all expressed genes. Gene identifiers were mapped to Entrez IDs, and separate enrichGO and enrichKEGG tests were run with Benjamini-Hochberg adjustment (p-adjust cutoff = 0.05). To maintain consistency and depth of interpretation, results from all cell types were aggregated, combining up-gene GO/KEGG tables into unified heatmap or network representations, instead of reporting each cell type separately. Likewise, pathway enrichments from all four pseudotime trajectories were pooled to reveal common versus lineage-specific functional modules.

8.2 Gene-Set Enrichment Analysis (GSEA)

Beyond discrete DEG sets, we leveraged full ranked gene lists ordered by $\log_2 FC$ for each cell type, and by pseudotime correlation scores for each trajectory to perform preranked GSEA. Using gseGO (Biological Process, BP ontology) and gseKEGG, we identified pathways whose component genes exhibit gradual up- or down-regulation across cell-state axes. For cell types, one unified GSEA was run per cell-type ranking, then collectively summarized to highlight shared regulatory circuits across the gingival ecosystem. For trajectories, one GSEA per lineage captured the temporal activation of developmental or inflammatory programs as cells progress from progenitor to terminal states.

8.3 Visualization and Comparative Interpretation

Top enriched GO terms and KEGG pathways (up to 15 per category) were visualized with barplots, dotplots, and category-gene network diagrams (cnetplot) using a transparent, publication-ready ggplot2 theme. GSEA results were displayed as ridgeplots of $-\log_{10}$(p.adjust) and individual enrichment curves (gseaplot2), with color scales encoding statistical significance. By juxtaposing cell-type and pseudotime GSEA landscapes, we distilled both static (cell-type identity) and dynamic (state transition) functional architectures, uncovering pathway modules, such as cytokine signaling, extracellular matrix remodeling, and metabolic reprogramming, consistently altered in PD and mobilized during cell-state evolution.

9. Integrated Predictive Modeling, Pathway Activity Scoring & Causal Inference

Building on our differential-expression and trajectory analyses, we implemented a unified R pipeline to select high-confidence biomarkers, assess their functional context, predict disease status, and probe causal links at the patient level. Key packages included randomForest (v4.7.1.2), caret (v7.0.1), pROC (v1.18.5), enrichplot (v1.26.1), and mediation (v4.5.0).

9.1 Data Cleaning & Feature Filtering

After loading the integrated Seurat object, all mitochondrial genes (identified via ^MT[-\\.]) were removed to avoid confounding by cellular respiration artifacts. A curated list of 34 housekeeping

and ribosomal/heat-shock genes was excluded, and the top 20 DEGs (ranked by adjusted p-value) were retained as candidate markers.

9.2 Predictive Modeling with Random Forest
Expression of these 20 genes (normalized RNA counts) formed the feature matrix for a Random Forest classifier (ntree = 1000). Cells were split 80 %/20 % into training and test sets via caret::createDataPartition, respectively. Performance was quantified by confusion matrices (sensitivity/specificity per group) and ROC AUC (pROC), demonstrating robust discrimination between PD and control.

9.3 Functional Annotation of Candidate Markers
The top 50 DEGs were mapped to Entrez IDs (org.Hs.eg.db::mapIds) and subjected to KEGG (clusterProfiler::enrichKEGG) and GO-BP (clusterProfiler::enrichGO) over-representation tests (BH-adjusted $p < 0.05$). Results were visualized with enrichplot::dotplot (top 10 categories), ensuring gene sets were rendered without list-column issues.

9.4 Single-Cell Pathway Activity Profiling (GSEA)
To capture pathway-level variation across all cells, enriched KEGG modules (gene sets ≥ 5 members) were extracted and used as input to GSEA. The resulting per-cell activity matrix was saved for downstream clustering or differential activity testing.

9.5 Patient-Level Mediation Analysis
Where patient identifiers were available in the meta.data, we aggregated the average expression of a key mediator gene per patient. Linking clinical intake to PD status via this mediator, we fit a linear model (lm) for the mediator and a logistic model for the PD flag. We then conducted causal mediation analysis (mediation::mediate, sims = 1,000) to estimate direct and indirect effects.

This end-to-end pipeline not only pinpoints top candidate biomarkers and their functional pathways but also integrates machine learning, pathway-activity scoring and causal inference, providing a comprehensive framework for the generation of translational hypotheses in PD.

10. Extended Population-Based Analysis Using NHANES 2011-2012

10.1 Data Source Selection
The NHANES 2011-2012 cycle was deliberately chosen because it represents one of the most recent survey waves containing a complete periodontal examination (full-mouth measurements of probing pocket depth and clinical attachment loss), which is essential for defining PD status. Later cycles either omitted or only partially assessed periodontal probing, precluding rigorous case ascertainment.

10.2 Data Harmonisation and Variable Recoding
All available XPT modules (demographics, oral health, biochemistry, complete blood counts, diet recalls, etc.) were imported and merged by participant SEQN. Within each module, numeric variables were aggregated by mean and categorical variables by first non-missing entry, yielding a single, de-duplicated row per subject. A comprehensive annotation dictionary was built from each variable's label attribute. Dichotomous survey items were recoded to 0/1 (no / yes) with a uniform missing-value scheme, and new bin indicators were generated to support binary-feature models. Continuous covariates were retained only if they met prespecified thresholds for sample size, variance, and unique values.

10.3 Definition of PD

PD status was defined in a fully reproducible manner: subjects exhibiting ≥2 sites with both probing pocket depth ≥5 mm and clinical attachment loss ≥4 mm were coded as cases. This dual criterion ensures specificity for moderate-to-severe disease and aligns with contemporary consensus definitions.

10.4 Univariate and Covariate-Adjusted Comparisons

Each candidate variable was first summarized by PD: counts and proportions for binary items; means, medians, interquartile ranges, and ranges for continuous measures. Covariate-adjusted differences were then estimated via linear regression models including age and sex. When residuals deviated from normality (Shapiro-Wilk $p < 0.05$), rank-based models on pseudo-ranks were employed. Categorical exposures were compared by chi-square or Fisher's exact tests, with all results reported with two-sided p-values and 95 % confidence intervals for PD effects.

10.5 LASSO for Feature Extraction

To distill the most informative subset of hundreds of potential predictors, an L_1-penalized logistic regression (LASSO) was applied to standardized continuous variables. Five-fold cross-validation identified the penalty parameter minimizing out-of-sample deviance, and features with nonzero coefficients at this optimum were retained. This approach leverages sparsity to guard against overfitting while highlighting the strongest biomarkers of PD.

10.6 Random Forest and XGBoost for Predictive Ranking

Ensemble tree-based methods were used to capture complex, potentially nonlinear interactions among predictors:

- Random Forest models provided variable importance rankings via mean decrease in Gini impurity, offering robustness to noise and missing data.
- XGBoost (gradient-boosted decision trees) further refined these rankings by optimizing a regularized log-loss objective and allowing fine-grained control over tree complexity and shrinkage.

Together, these machine-learning algorithms furnish complementary views of predictor relevance, emphasizing artificial-intelligence-driven discovery, and ensure that both global and local feature effects are captured.

10.7 Causal Forest for Heterogeneous Treatment Effects

To probe potential causal drivers (from exposures to PD) and consequences (from PD to downstream outcomes) within the observational data, causal forests from the generalized random forest framework were employed. This nonparametric, tree-based method estimates subject-level treatment effects under unconfoundedness assumptions, automatically adjusting for all covariates and detecting effect heterogeneity. Only variables whose 95 % confidence intervals for the estimated average treatment effect excluded zero were reported as putative causal factors or outcomes, highlighting novel mechanistic insights beyond mere prediction.

10.8 Visualization and Reporting

Top-ranked variables from each method were visualized using high-resolution gradient bar plots in ggplot2, with custom color ramps to underscore rank order. Volcano-style scatterplots and boxplots of key biomarkers were generated to contextualize statistical and practical significance. All intermediate and final results, including annotated master datasets, univariate tables, lists of LASSO-selected features, random-forest and XGBoost importance tables, and causal-forest effect estimates, were exported in CSV and RDS formats to ensure full reproducibility and transparency.