Available Online: 9 October 2026 DOI: 10.54254/3049-5458/2026.37538

Single-cell transcriptomics links HPV transcript abundance to malignant epithelial heterogeneity and redistribution of cell-cell communication across cervical lesions

***Han Zhang*[1], *Wei Chen*[1], *Yaping Li*[2], *Wei Wang*[2,3,4,5,*]**

[1]School of Engineering Medicine, Beihang University, Beijing, China
[2]Department of Obstetrics and Gynecology, The Fifth People's Hospital of Ningxia, Shizuishan, China
[3]Department of Obstetrics and Gynecology, Peking Union Medical College Hospital, Beijing, China
[4]Chinese Academy of Medical Sciences, Beijing, China
[5]National Clinical Research Center for Obstetric and Gynecologic Diseases, Peking Union Medical College, Beijing, China

*Corresponding Author. Email: vivianwang_pumch@sina.com

Abstract. Persistent high-risk Human Papillomavirus (HPV) infection is a prerequisite for most cervical cancers, but only a subset of infections progress to invasive disease. This study investigated associations between HPV transcript abundance, malignant epithelial heterogeneity, inferred cell-state trajectories, and microenvironmental organization at single-cell resolution. Single-cell RNA sequencing data from GSE208653 and a supplementary cervical squamous cell carcinoma dataset were mapped to a combined human-HPV reference. The analytical framework integrated inferCNV-based copy-number analysis, GeneNMF metaprogram analysis, pseudotime inference, CellChat, network-based in silico knockout, survival analysis, and assessment in an independent single-cell dataset. After quality control, 105,799 cells were retained, and transcripts from HPV16, HPV18, and HPV59 were detected. Putative malignant epithelial cells separated into HPV^{high}/MP6-high and HPV^{dim}/MP6-low states. Among the 10 metaprograms, MP6 showed the strongest association with HPV transcript abundance and shared 56 genes with the intermediate-peaking C2 dynamic module (adjusted $p = 9.58 \times 10^{-22}$). The two states had similar aggregate interaction numbers and strengths but differed in the distribution of inferred pathways. Compared with HPV-positive normal cervical samples, high-grade squamous intraepithelial lesion samples showed fewer and weaker inferred interactions, together with a shift from immune-centered networks toward epithelial and epithelial-stromal communication. In silico perturbation prioritized COL6A1, ABCC3, STARD10, and AOC1, with stronger external support for the first three than for AOC1. Together, these analyses support associations between HPV transcript abundance, malignant epithelial states, and microenvironmental organization but do not establish causality. Experimental validation and larger multi-patient cohorts are required.

Keywords: human papillomavirus, cervical squamous cell carcinoma, single-cell RNA sequencing, cellular heterogeneity, cell-cell communication

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cervical cancer is one of the most common malignancies in women, and nearly all cases are associated with high-risk Human Papillomavirus (HPV) infection. Cervical carcinogenesis spans HPV-positive non-neoplastic epithelium, precancerous lesions, and invasive cancer, although only a small proportion of infections persist and progress [1-3]. High-risk HPV E6 promotes p53 degradation [4], whereas E7 interacts with the retinoblastoma protein and disrupts cell-cycle regulation [5-7]. These viral oncoproteins affect multiple host pathways involved in HPV-driven oncogenesis [8-11]. Nevertheless, how HPV transcript abundance varies across histopathological lesion categories and malignant epithelial states remains insufficiently resolved at single-cell resolution.

Conventional bulk transcriptomic profiling captures tissue-averaged signals and cannot readily distinguish individual cell populations or transitional states. Single-cell RNA sequencing (scRNA-seq) can resolve cellular composition, heterogeneity within malignant epithelial populations, continuous state trajectories, and cell-cell communication. It therefore provides a framework for investigating the relationships among HPV transcript abundance, host programs, epithelial architecture, and the local microenvironment [12-15].

Existing scRNA-seq studies of cervical cancer have focused mainly on constructing cell atlases, characterizing the tumor microenvironment, and examining changes in treatment response. Single-cell transcriptomic studies have revealed cellular heterogeneity and distinct epithelial subpopulations in cervical cancer [12]. Reviews of this field have summarized advances in understanding the cervical cancer ecosystem during disease progression and treatment [13, 14]. Building on these findings, this study examined associations between HPV transcript abundance, malignant epithelial programs, inferred cell-state trajectories, and cell–cell communication across cervical lesion categories. This study therefore addressed four questions. First, HPV transcript distributions were quantitatively compared across lesion categories and cell subsets. Second, malignant epithelial states and expression programs associated with HPV transcript abundance were identified. Third, variation in these programs was examined along inferred pseudotime trajectories and across cell-cell communication networks. Finally, prognostic associations and external data were integrated to prioritize candidate genes.

The analytical framework had two main features. First, a combined human-HPV reference containing GRCh38 and the genomes of 14 high-risk HPV types was constructed. This reference enabled simultaneous quantification of host genes and viral transcripts at single-cell resolution. Second, inferCNV, GeneNMF, Monocle3, CellChat, and scTenifoldKnk were integrated to relate HPV transcript abundance to malignant epithelial heterogeneity, inferred cell-state variation, and epithelial-stromal interactions. GEPIA and GSE168652 were used to evaluate prognostic associations and candidate-gene expression patterns, respectively.

2. Materials and methods

2.1. Data sources and sample information

Single-cell RNA sequencing data comprising two HPV-negative normal cervical tissue samples, two HPV-positive normal cervical tissue samples (N), and two HPV-positive High-grade Squamous Intraepithelial Lesion samples (HSIL; H) were obtained from the Gene Expression Omnibus (GEO) under accession number GSE208653. Fresh tissue from one patient diagnosed with cervical cancer was collected at Peking Union Medical College Hospital. This study was conducted in accordance with the principles of the Declaration of Helsinki and was approved by the ethics committees of the Peking Union Medical College Hospital. The

single-cell library construction and sequencing were performed according to the standard Chromium Single Cell 3′ GEM, Library & Gel Bead Kit workflow.

2.2. Single-cell data preprocessing

All samples were realigned to a combined human-HPV reference genome constructed from GRCh38 and sequences from 14 high-risk HPV types (including HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68). Host-gene and HPV-transcript expression matrices were then generated using a uniform workflow.

Quality control and preprocessing were performed using Seurat v5. Cells were retained if they met all of the following criteria: 300 < nFeature_RNA < 5,000; nCount_RNA > 1,000 but below the sample-specific 97th percentile; a mitochondrial-transcript proportion < 10%; and a hemoglobin-related transcript proportion < 1%. The retained data were normalized using NormalizeData, and the top 2,000 highly variable genes were selected before scaling with ScaleData. Batch effects were corrected using Harmony based on the Principal Component Analysis (PCA) results. A neighbor graph was constructed using the first 50 principal components, clustering was performed at resolution = 0.1, and the data were visualized using UMAP and t-SNE.

2.3. Cell-type annotation

Seurat was used to identify marker genes for each cluster. Automated annotation was performed using SingleR with the HumanPrimaryCellAtlasData reference set, followed by manual review based on established markers, including EPCAM, KRT8, KRT18, CD3D, MS4A1, NKG7, LYZ, COL1A1, and PECAM1. Differentially expressed genes were identified using the Wilcoxon rank-sum test, with an adjusted p value < 0.05 considered significant.

2.4. Identification of malignant cells

inferCNV was used to infer large-scale Copy-Number Variations (CNVs) at single-cell resolution. Immune cell populations served as the reference. Gene-expression signals were ordered and smoothed according to chromosomal position, and continuous expression shifts in the cells of interest were compared with those in the reference cells. Cell populations with clear amplification or deletion patterns were identified by integrating the heatmap and inferred CNV scores. These populations were designated as putatively malignant for downstream analyses.

2.5. Epithelial cell reclassification and stratification by HPV transcript abundance

After the malignant cells were determined to be derived mainly from the epithelial lineage, the epithelial subset was extracted and cells originating from the normal cervix were removed. Normalization, highly variable gene selection, scaling, PCA, and Harmony integration were then repeated. A neighbor graph was constructed in the corrected low-dimensional space, followed by clustering and UMAP visualization to resolve heterogeneity within the malignant epithelial population.

2.6. Identification of molecular metaprograms

Non-negative matrix factorization of the malignant epithelial expression matrix was performed using GeneNMF. Candidate programs were repeatedly identified with multiNMF across factorization ranks of k = 2:10. The assay was set to RNA, and min.exp = 0.05 was used to remove genes with low detection. Programs were subsequently integrated according to their reproducibility and gene composition across factorization ranks and cell subsets. This procedure yielded 10 stable metaprograms, designated MP1–MP10.

2.7. Pseudotime analysis

Monocle3 was used to infer putative trajectories among malignant epithelial cell states. For the overall analysis, a cell_data_set was generated from the raw count matrix. The parameters were set to num_dim = 50 in preprocess_cds, an initial resolution = 1e-7 in cluster_cells, and minimal_branch_len = 80 in learn_graph. Pseudotime was calculated using order_cells.

The two main left and right partitions were then analyzed independently. For both partitions, num_dim was set to 30, whereas minimal_branch_len was set to 5 and 10 for the left and right partitions, respectively. Genes associated with pseudotime were identified using graph_test.

2.8. Cell-cell communication analysis

CellChat was used to infer ligand-receptor-mediated communication networks for the HPV^{high}/MP6-high and HPV^{dim}/MP6-low states and for the N and H groups. Incoming and outgoing signaling patterns, interaction numbers, aggregate interaction strengths, and pathway-level differences were compared between each pair of groups.

2.9. In silico knockout and functional enrichment analyses

scTenifoldKnk was used to perform individual in silico knockouts of HPV E1, E2, E4, E5, E6, E7, L1, and L2. The main parameters were qc_mtThreshold = 0.2, qc_minLSize = 200, nc_nNet = 3, nc_nCells = 100, and nc_nComp = 3. Significantly perturbed genes were selected using an adjusted p value (p.adj) < 0.05 and $|Z| \geq 2$.

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) over-representation analyses were performed using clusterProfiler. Ranked gene lists were constructed from the diffRegulation Z values obtained after the in silico knockouts. Gene Set Enrichment Analysis (GSEA) was then performed using the Molecular Signatures Database (MSigDB) Hallmark gene sets. Multiple-testing correction was performed using the Benjamini–Hochberg method.

2.10. Candidate-gene prioritization and prognostic analysis

Candidate genes were prioritized by integrating four evidence layers: in silico knockout results; overlap with MP6, C2, and the genes shared by MP6 and C2; pseudotime expression patterns; and GEPIA disease-free survival associations.

2.11. Assessment in an independent single-cell dataset

GSE168652 was selected for an external single-cell assessment. This dataset contains cervical cancer tissue (GSM5155196) and paired adjacent-normal tissue (GSM5155197) from the same patient. Sequencing was performed on an Illumina NovaSeq 6000 platform.

Cells were assigned to the Adjacent.normal and Cancer groups according to sample origin. Group-level expression differences were compared for COL6A1, ABCC3, STARD10, and AOC1. Pseudotime analysis was also performed across all cells and mapped onto the UMAP space to examine changes in the expression of these four candidate genes across cell states. In addition, the GSE120691 peripheral-blood bulk RNA-seq dataset, comprising 19 healthy controls, 21 Cervical Intraepithelial Neoplasia (CIN) samples, and 11 cervical cancer samples, was examined in an auxiliary analysis.

3. Results

3.1. Cell atlas construction and cell-type annotation

The original GSE208653 cohort comprised two normal cervical tissue samples, two HPV-positive normal cervical tissue samples, two HPV-positive HSIL samples, and three HPV-positive cervical cancer samples. Because HPV transcript signals were weak in the original C group, these samples were replaced with a supplementary cervical squamous cell carcinoma scRNA-seq dataset previously generated by the research team. Transcripts from HPV16, HPV18, and HPV59 were detected in the integrated single-cell dataset. After uniform quality control, 105,799 high-quality cells and 32,065 genes were retained, and the cells were assigned to the normal, N, H, and C groups according to sample category (Figure 1A and B).

Harmony integration, dimensionality reduction, and clustering identified 19 cell clusters. By combining SingleR annotation with established marker genes, cells accounting for 99% of the dataset were assigned to seven major cell types: T cells, B cells, NK cells, monocytes, fibroblasts, endothelial cells, and epithelial cells (Figure 1C).

Epithelial cells were the most abundant major cell type, accounting for 35.15% of all cells, and were therefore selected as the principal population for subsequent malignant-cell identification and cell-state analysis.

3.2. CNV-based inference of putative malignant epithelial cells

inferCNV-based analysis was performed using immune cell populations as the reference. The reference cells showed near-baseline profiles, whereas subsets of the cells of interest displayed continuous amplification and deletion bands across multiple chromosomal regions, consistent with large-scale copy-number abnormalities (Figure 1D).

Among the 19 clusters, clusters 1, 4, 5, 7, 10, and 14 showed higher inferred CNV burden and more extensive segmental amplification and deletion patterns. These clusters were therefore designated as putatively malignant epithelial populations. Cells with high inferred CNV burden were derived mainly from the epithelial lineage (Figure 1E).

On this basis, the epithelial subset was extracted, and cells originating from the normal cervix were removed. Repeated dimensionality reduction and clustering retained 24,417 cells and resolved 12 subclusters (Figure 2A and B).

The putative malignant epithelial cells formed two major compartments in UMAP space. The left compartment showed higher HPV transcript signals, whereas the right compartment showed markedly weaker signals, with viral transcripts being nearly undetectable in some cells. This distribution delineated malignant epithelial states associated with HPV transcript abundance.

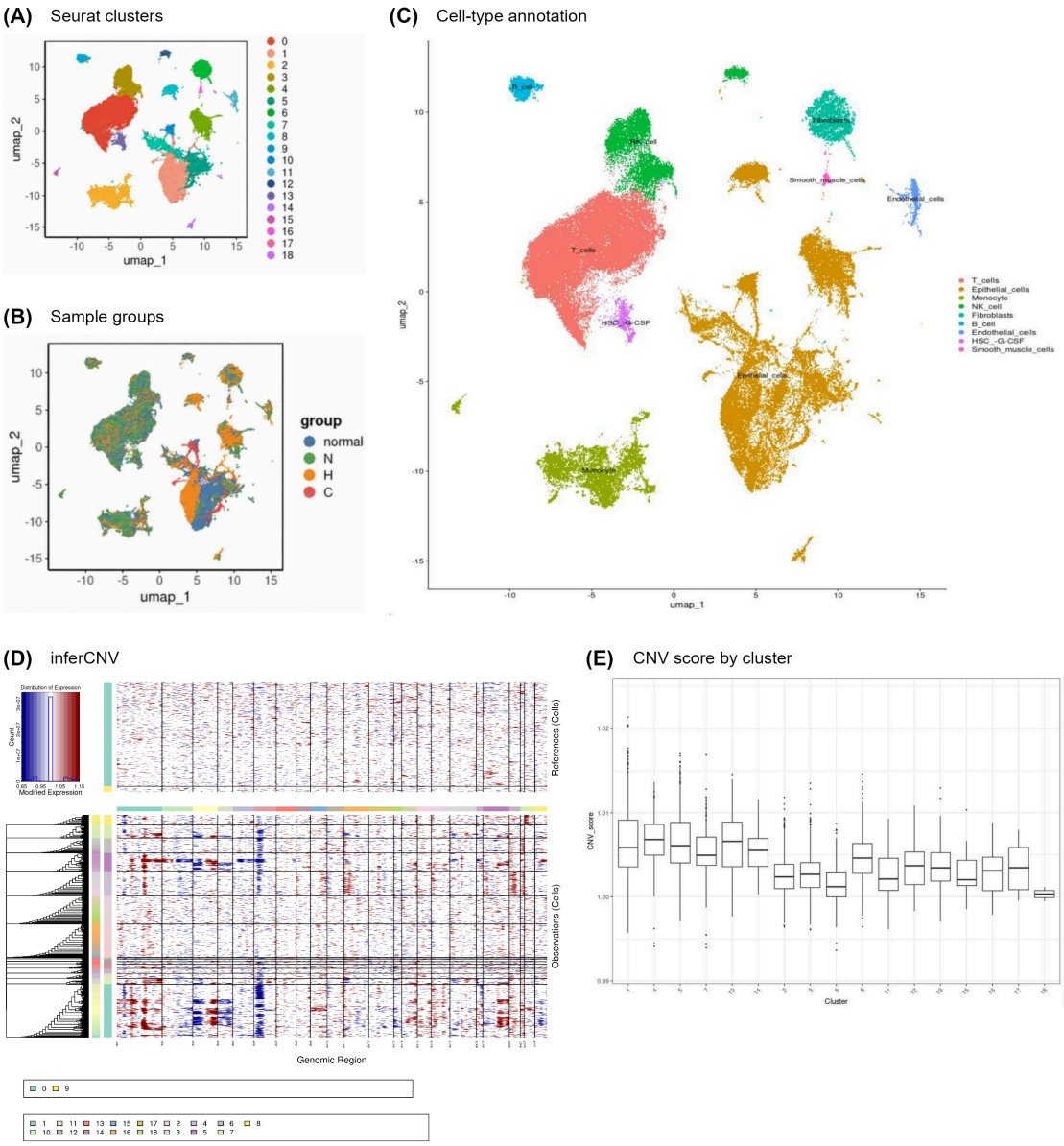


Figure 1. Single-cell atlas and copy-number variation-based identification of putative malignant epithelial cells. (A) UMAP representation of 105,799 cells colored by the 19 Seurat clusters. (B) The same UMAP colored by sample group: normal, HPV-negative normal cervix; N, HPV-positive normal cervix; H, HPV-positive high-grade squamous intraepithelial lesion; and C, cervical squamous cell carcinoma. (C) UMAP colored by cell-type annotation, highlighting T cells, B cells, NK cells, monocytes, fibroblasts, endothelial cells, and epithelial cells. (D) inferCNV heatmap with immune reference cells shown above and cells of interest below. Columns follow genomic position, whereas red and blue bands represent relative amplification-like and deletion-like signals, respectively. (E) Distribution of inferred CNV scores across epithelial cell clusters. Clusters 1, 4, 5, 7, 10, and 14 showed higher inferred CNV burden and were designated as putatively malignant epithelial populations

3.3. Molecular metaprogram analysis of malignant epithelial cells

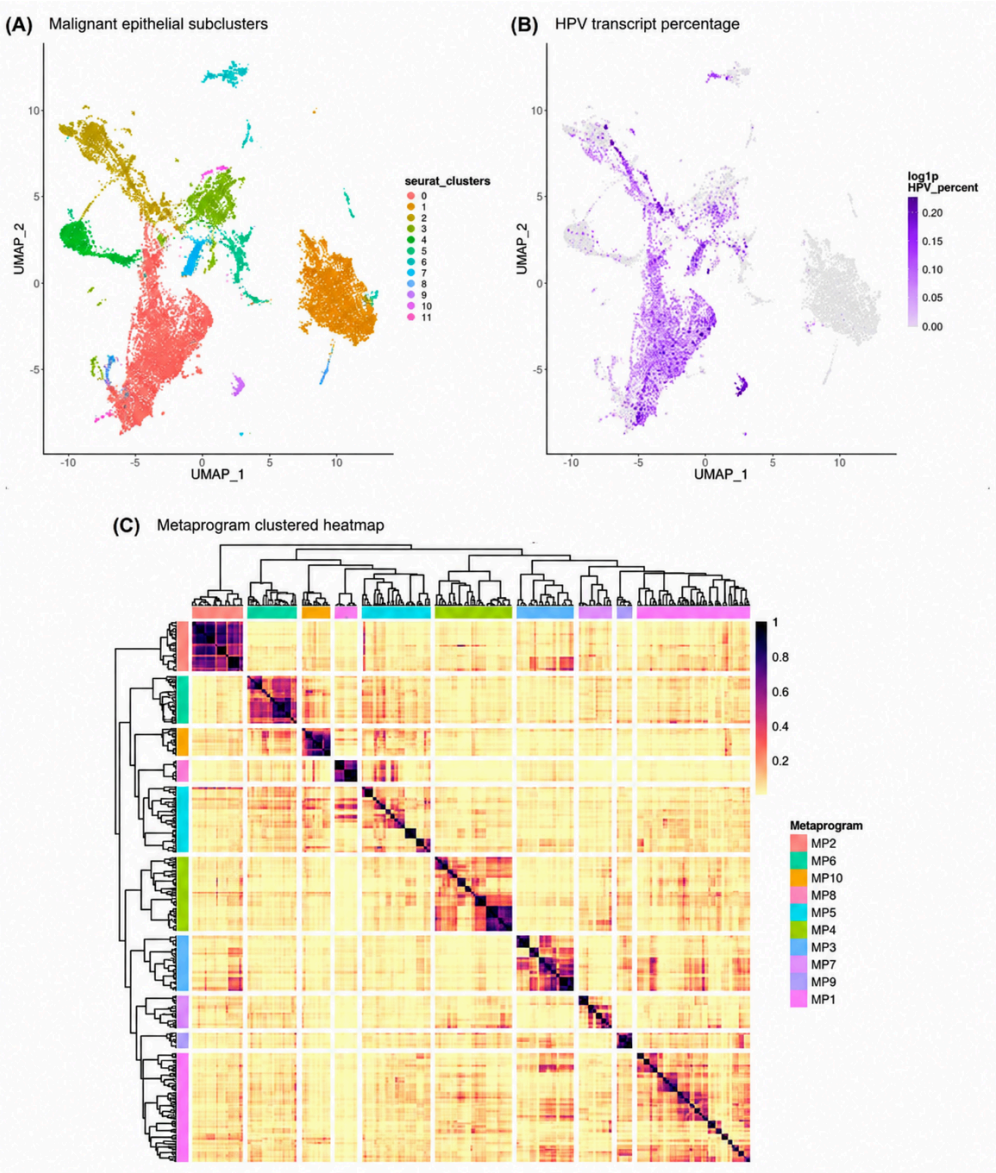


Figure 2. Malignant epithelial heterogeneity and the HPV-associated MP6 metaprogram. (A) UMAP reclustering of 24,417 putative malignant epithelial cells into 12 subclusters after exclusion of cells derived from normal cervical tissue. (B) HPV transcript percentage mapped onto the same UMAP using a log1p-transformed color scale. Higher viral transcript signals are concentrated in the left compartment, whereas the right compartment shows weak or nearly undetectable signals. (C) GeneNMF clustered heatmap of the 10 metaprograms, MP1-MP10. MP6 showed the clearest association with higher HPV transcript abundance and was selected for subsequent analyses

GeneNMF analysis identified 10 metaprograms, designated MP1–MP10. Among them, MP6 showed the clearest HPV-associated profile. HPV E1, E6, and E7 were directly detected among its representative genes, indicating that MP6 was a malignant epithelial gene-expression program associated with higher HPV transcript abundance (Figure 2C).

MP6 was significantly enriched for TNF-α signaling via NF-κB, KRAS signaling, mTORC1 signaling, IL-2/STAT5 signaling, glycolysis, hypoxia, apoptosis, reactive oxygen species, and epithelial-mesenchymal transition pathways. Complement/coagulation, xenobiotic metabolism, and early and late estrogen-response pathways were also enriched. This enrichment profile encompassed multiple host states related to inflammation, proliferation, metabolic alterations, and stress adaptation.

Together, these results showed that MP6 did not merely reflect whether HPV transcripts were detected. Instead, MP6 captured coordinated variation in HPV transcripts and host programs related to inflammation, proliferation, metabolism, and epithelial-mesenchymal transition.

3.4. Cell-state variation along pseudotime trajectories

The overall Monocle3 trajectory separated malignant epithelial cells into two major compartments. The left compartment contained 17,396 cells and spanned a broader range of states, whereas the right compartment contained 5,932 cells and showed a more compact structure. When analyzed separately, both compartments exhibited continuous pseudotime gradients, consistent with distinct computational trajectories of cell-state variation (Figure 3A).

Mapping HPV E1, E6, and E7 transcripts and MP6 activity onto the trajectories showed that cells with high viral transcript signals and high MP6 activity were concentrated mainly in the left compartment, whereas signals were generally weaker in the right compartment. Accordingly, the left and right compartments were designated the HPV^{high}/MP6-high and HPV^{dim}/MP6-low states, respectively.

Separate graph_test analyses identified 4,417 genes significantly associated with pseudotime in the HPV^{high}/MP6-high compartment, of which 3,381 were specific to that compartment, and 1,298 significant genes in the HPV^{dim}/MP6-low compartment. Cells in the left compartment were divided into 30 pseudotime intervals and subjected to k-means clustering, yielding five pseudotime-dependent dynamic modules, C1–C5 (Figure 3B).

MP6 showed the most significant overlap with the intermediate-peaking C2 dynamic module, with 56 shared genes (BH-adjusted $p = 9.58 \times 10^{-22}$). C2 peaked in the middle of pseudotime and contained HPV E1, E5, and E7. It was enriched mainly for early and late estrogen-response pathways and apical junctions, consistent with an intermediate dynamic window associated with higher HPV transcript abundance, increased MP6 activity, and epithelial structural remodeling.

Functional comparison showed that genes upregulated in the high-HPV-transcript left compartment were involved mainly in the cell cycle, DNA replication, RNA processing, and cancer-related signaling, whereas the low- or undetectable-HPV-transcript right compartment was characterized more strongly by antigen processing and presentation, complement/coagulation, cell adhesion, and immune-related processes (Figure 3C).

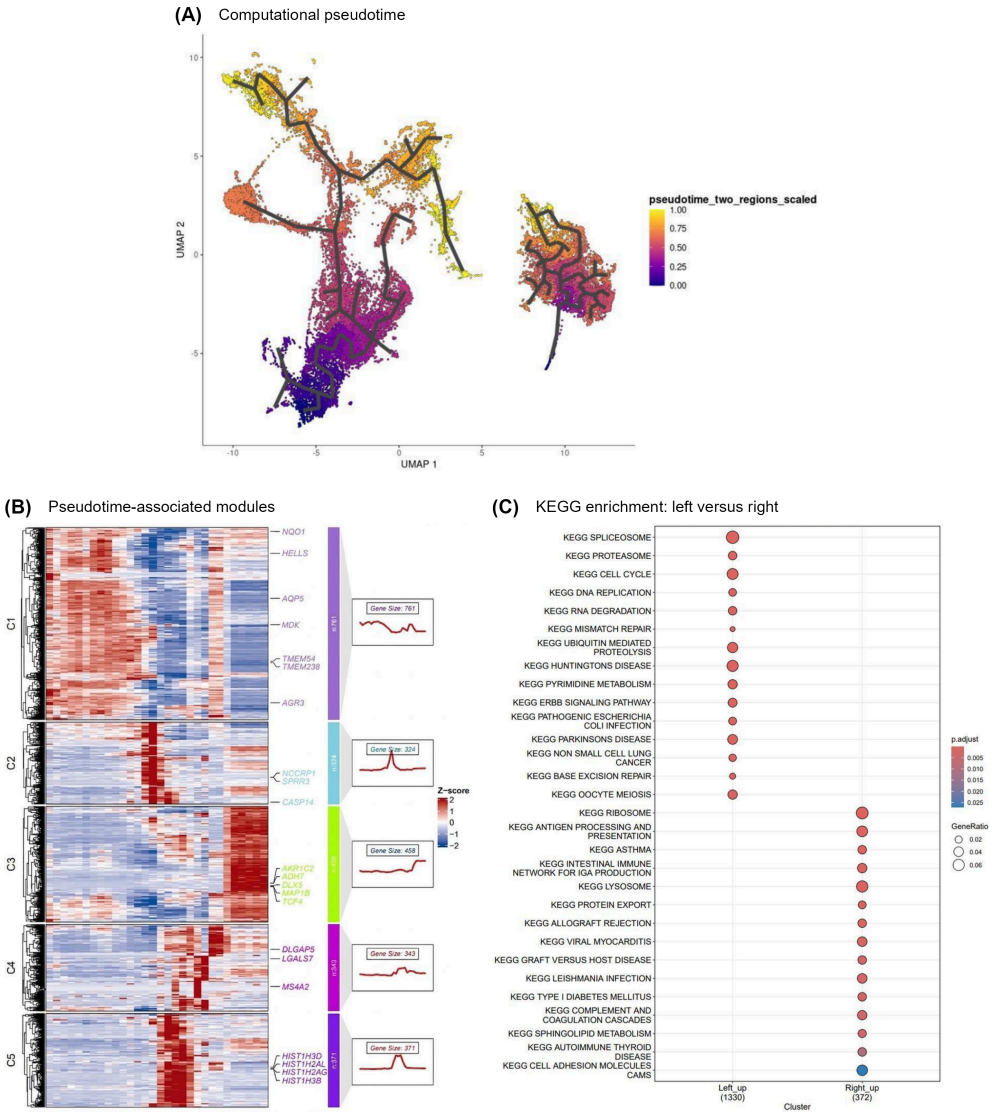


Figure 3. Computational trajectories and dynamic programs of malignant epithelial cell states. (A) Monocle3 trajectory of malignant epithelial cells, colored by separately scaled pseudotime across the left and right compartments. (B) Heatmap of genes significantly associated with pseudotime in the HPVhigh/MP6-high compartment, grouped into five dynamic modules, C1-C5; the adjacent profiles summarize the expression trajectory of each module. C2 contains 324 genes and shows a transient intermediate-pseudotime peak. It is also the module with the most significant overlap with MP6, comprising 56 shared genes (BH-adjusted $p = 9.58 \times 10^{-22}$). (C) KEGG enrichment of genes upregulated in the left high-HPV-transcript compartment and the right low- or undetectable-HPV-transcript compartment. Dot size denotes GeneRatio, and color denotes the adjusted p value. The left compartment is characterized primarily by cell-cycle, DNA-replication, RNA-processing, and cancer-related pathways, whereas the right compartment shows stronger enrichment of antigen-presentation, complement/coagulation, adhesion, and immune-related pathways

3.5. Cell-cell communication and microenvironmental interactions

The preceding analyses located MP6 predominantly within the HPV^{high}/MP6-high state and identified a significant association between MP6 and the intermediate-peaking C2 dynamic module. On this basis,

CellChat was used to infer cell-cell communication at two levels. First, the HPV^{high}/MP6-high and HPV^{dim}/MP6-low malignant epithelial states were compared within the complete microenvironment to characterize local interaction patterns associated with each state. Second, the N and H groups were compared to examine differences in inferred microenvironmental networks between HPV-positive normal cervical tissue and HSIL samples.

CellChat inferred 989 and 976 intercellular interactions in the HPV^{high}/MP6-high and HPV^{dim}/MP6-low states, respectively, with aggregate interaction strengths of 21.954 and 21.75. The comparable interaction numbers and strengths indicated that the principal difference involved redistribution among specific cell populations and pathways rather than a uniform increase in communication (Figure 4A and B).

Comparison across lesion stages identified 1,013 and 780 inferred intercellular interactions in the N and H groups, respectively, with aggregate interaction strengths of 0.534 and 0.35. Thus, both the inferred interaction number and aggregate strength were lower in H than in N (Figure 4C and D).

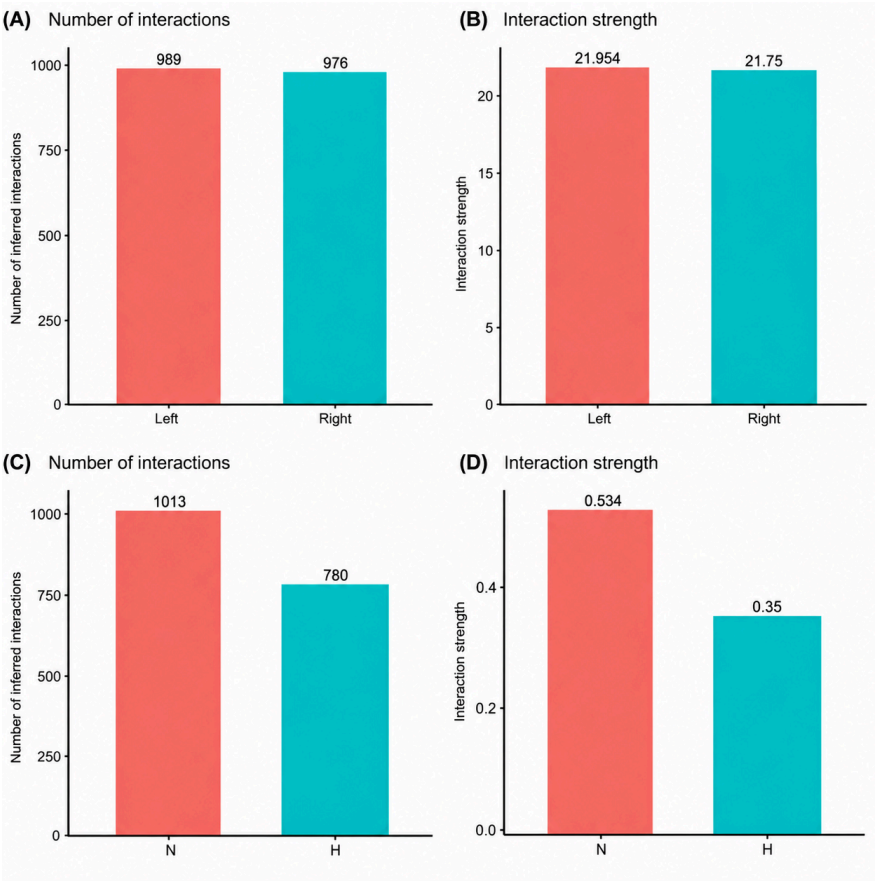


Figure 4. Aggregate CellChat-inferred communication across malignant epithelial states and lesion groups. (A, B) Number of inferred interactions (A) and aggregate interaction strength (B) in the HPVhigh/MP6-high and HPVdim/MP6-low states. The two states contain 989 and 976 inferred interactions, with aggregate strengths of 21.954 and 21.75, respectively. (C, D) Number of inferred interactions (C) and aggregate interaction strength (D) in HPV-positive normal cervical tissue (N) and HPV-positive high-grade squamous intraepithelial lesion (H). N and H contain 1,013 and 780 inferred interactions, with aggregate strengths of 0.534 and 0.35, respectively. The similar aggregate values between the two malignant epithelial states and the lower values in H relative to N motivated further cell-type- and pathway-level comparisons

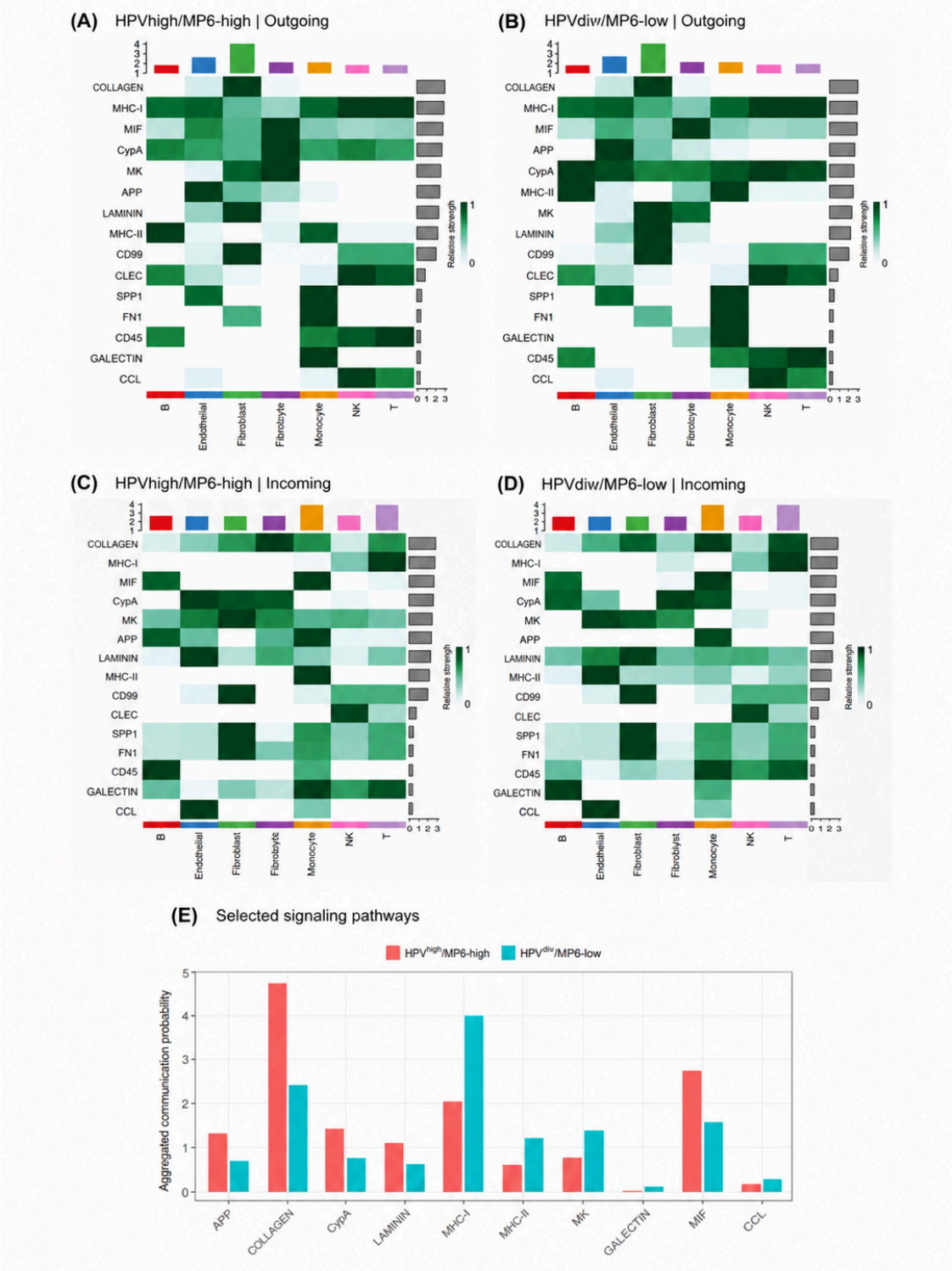


Figure 5. Redistribution of predicted signaling between HPVhigh/MP6-high and HPVdim/MP6-low malignant epithelial states. (A, B) Outgoing signaling patterns inferred for the HPVhigh/MP6-high (A) and HPVdim/MP6-low (B) states. (C, D) Incoming signaling patterns inferred for the HPVhigh/MP6-high (C) and HPVdim/MP6-low (D) states. Rows represent signaling pathways, columns represent cell populations, and green intensity denotes relative signaling strength; marginal bars summarize pathway- and cell-level contributions. (E) Aggregate communication probabilities for selected signaling pathways. Fibroblasts are the

principal predicted signal senders in both states, whereas monocytes and T cells are major predicted recipients. The HPVhigh/MP6-high state shows greater contributions from COLLAGEN, LAMININ, MK, CypA, and MIF signaling, whereas the HPVdim/MP6-low state shows greater contributions from MHC-I, MHC-II, APP, COMPLEMENT, and GALECTIN signaling. These patterns reflect redistribution of inferred communication rather than uniform communication enhancement

Analysis of outgoing signaling patterns showed that fibroblasts were the principal predicted signal senders in both malignant epithelial states, followed by endothelial cells. The predicted contribution of malignant epithelial cells to outgoing signaling was greater in the HPV^{dim}/MP6-low state (Figure 5A and B).

The two states shared a broadly similar pattern of outgoing signaling, although the predicted contribution of malignant epithelial cells differed. Monocytes and T cells were the principal predicted signal recipients in both states, whereas the HPV^{high}/MP6-high state received greater predicted input from the microenvironment (Figure 5C and D).

Comparison of key signaling pathways showed that the HPV^{high}/MP6-high state was characterized mainly by COLLAGEN, LAMININ, DESMOSOME, MK, and CypA interactions related to epithelial-stromal adhesion, extracellular-matrix support, and local structural stability. In contrast, MHC-I, MHC-II, COMPLEMENT, and GALECTIN pathways associated with immune recognition were more prominent in the HPV^{dim}/MP6-low state (Figure 5E).

In the incoming signaling patterns of the N and H groups, T cells and monocytes were the principal predicted signal recipients in N, whereas epithelial cells became the predominant predicted signal receivers in H (Figure 6A and B).

The dominant predicted signal recipients therefore shifted from immune cells in N toward epithelial cells in H. Outgoing signaling changed in the same direction: T cells and monocytes were the principal predicted signal senders in N, whereas epithelial cells made the largest predicted contribution to outgoing signaling in H, accompanied by fibroblast participation (Figure 6C and D).

Further comparison of key signaling pathways showed that the H group was characterized mainly by MK, COLLAGEN, CypA, LAMININ, CLDN, and APP networks related to local epithelial signaling, extracellular-matrix dependence, and epithelial-stromal adhesion. In contrast, the N group was characterized mainly by MHC-I, CLEC, GALECTIN, and CD99 networks associated with immune recognition and immune coordination (Figure 6E).

In summary, the two malignant epithelial states had similar aggregate interaction numbers and strengths, but the distribution of inferred interactions differed. Compared with N, H showed fewer inferred interactions and lower aggregate interaction strength, together with a shift in network focus from immune-related connections toward local networks centered on epithelial cells and epithelial-stromal interactions. Together, these comparisons were consistent with redistribution and reorganization of inferred cell-cell communication networks rather than a uniform increase in communication.

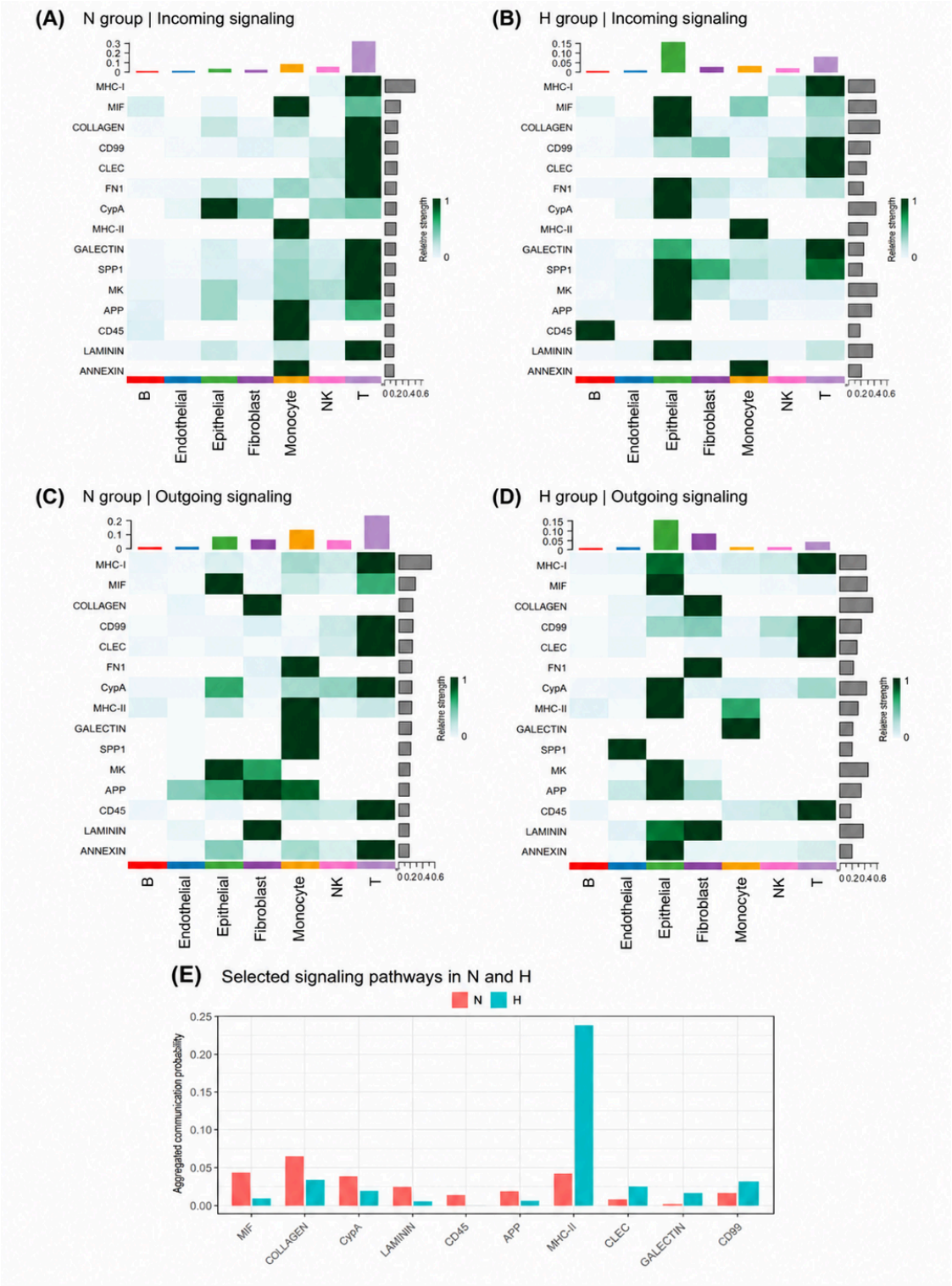


Figure 6. CellChat-inferred network organization in HPV-positive normal cervical tissue and high-grade squamous intraepithelial lesions. (A, B) Incoming signaling patterns in the N (A) and H (B) groups. (C, D) Outgoing signaling patterns in the N (C) and H (D) groups. Rows represent signaling pathways, columns represent cell populations, and green intensity denotes relative signaling strength. (E) Aggregate communication probabilities for selected pathways in N and H. T cells and monocytes are the principal

predicted signal recipients and senders in N, whereas epithelial cells make the predominant predicted contribution in H, accompanied by fibroblast participation. H shows greater contributions from MK, COLLAGEN, CypA, LAMININ, CLDN, and APP networks, whereas N shows greater contributions from MHC-I, CLEC, GALECTIN, and CD99 networks. These comparisons are consistent with redistribution from immune-centered communication in N toward epithelial and epithelial-stromal communication in H

3.6. Network-based in silico knockout identifies predicted gene-network perturbations

Simulated knockout of HPV E5, E2, and E7 identified 91, 89, and 86 significantly perturbed genes, respectively, whereas simulated knockout of E1 and E6 identified 15 and 8, respectively (Figure 7A). The significantly perturbed gene sets for E5, E2, and E7 overlapped the C2 dynamic module by 18, 17, and 14 genes, with fold-enrichment values of 6.97, 6.97, and 6.10, respectively. The perturbed gene sets for E7, E2, and E5 each overlapped MP6 by 11 genes, with fold-enrichment values of 14.92, 14.04, and 13.26, respectively.

Within the intersection of C2 and MP6, the E2, E5, and E7 perturbation-associated gene sets overlapped 11, 11, and 10 genes, respectively, and the corresponding fold-enrichment values increased to 26.08, 24.63, and 25.19. In addition to COL6A1, ABCC3, STARD10, and AOC1, recurrent genes included ABLIM1, B3GALT5, TMEM54, AQP5, and SORBS2. E1 showed only limited overlap, whereas E6 did not form a stable intersection under the stringent significance threshold.

Hallmark GSEA identified recurrent enrichment of early and late estrogen responses, oxidative phosphorylation, apoptosis, coagulation, IL-2/STAT5 signaling, hypoxia, the p53 pathway, and epithelial-mesenchymal transition across multiple knockout models. Gene-specific comparisons showed that E1 perturbation involved MYC targets, E2F targets, glycolysis, and interferon responses. E2 and E5 perturbations more prominently involved oxidative phosphorylation, KRAS signaling, IL-2/STAT5 signaling, hypoxia, and metabolic stress. E6 perturbation was characterized more strongly by inflammatory response, complement, damage response, and apical junction programs, whereas E7 perturbation was concentrated in E2F targets, oxidative phosphorylation, the p53 pathway, apoptosis, and epithelial-mesenchymal transition. These results reflect predicted perturbations of the expression network and can inform mechanistic hypotheses and candidate-gene prioritization, but they do not establish experimental knockout effects.

Integration of the in silico knockout results, the association between MP6 and C2, pseudotime expression patterns, and GEPIA prognostic data prioritized four candidate genes: COL6A1, ABCC3, STARD10, and AOC1.

GEPIA survival analysis showed that the expression of COL6A1, ABCC3, STARD10, and AOC1 was significantly associated with disease-free survival (Figure 7B). The convergence of viral-perturbation evidence, overlap between MP6 and C2, pseudotime expression patterns, and prognostic associations supported retaining these four genes for further evaluation.

The computational analyses associated simulated HPV-gene perturbation with predicted changes in estrogen-response, epithelial-junction, metabolic, and stress-response programs. At the candidate-gene level, COL6A1, ABCC3, STARD10, and AOC1 pointed to extracellular-matrix remodeling, membrane transport, lipid homeostasis, and amine metabolism, respectively.

3.7. Assessment in an independent single-cell dataset

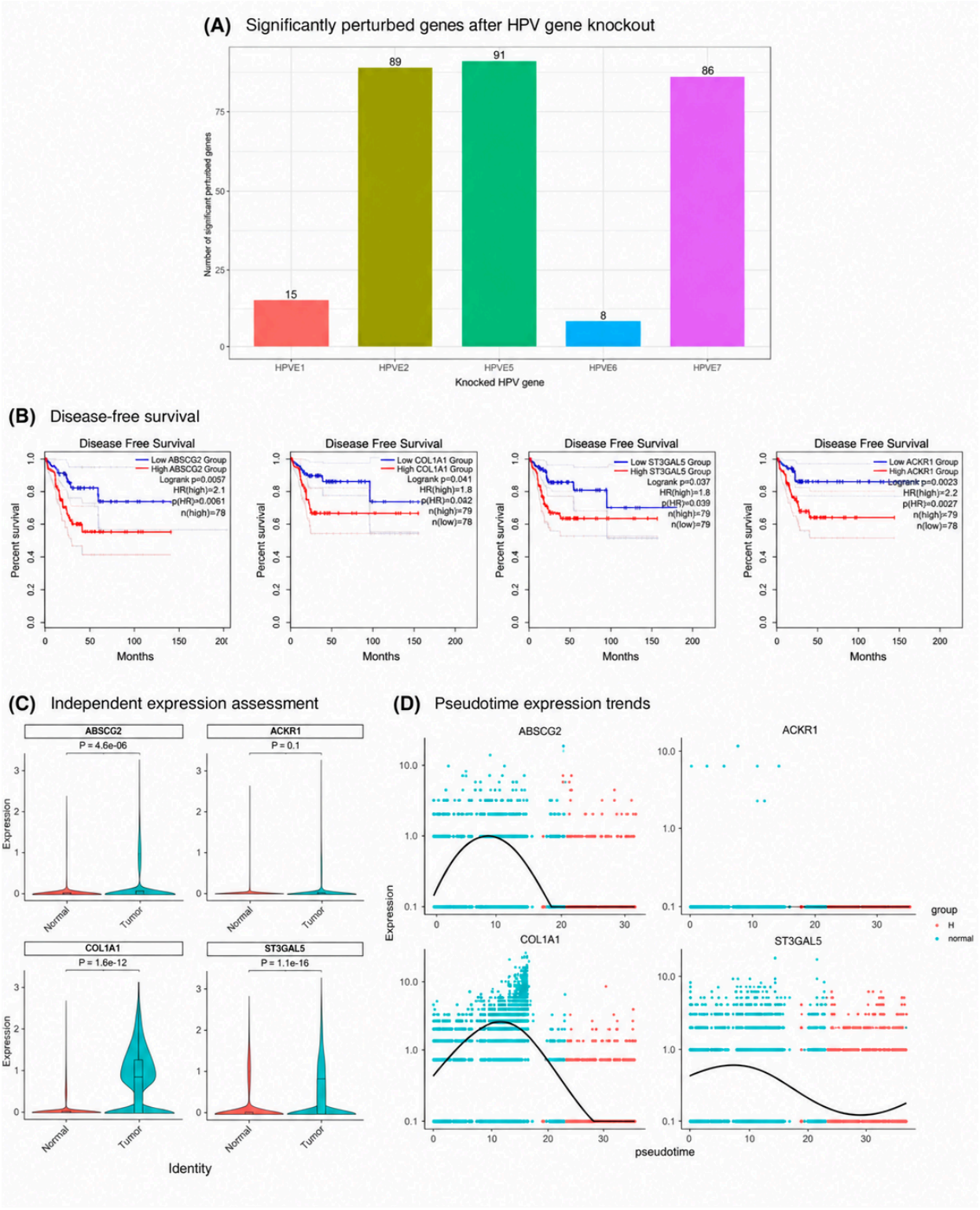


Figure 7. Candidate-gene prioritization through network perturbation, survival analysis, and independent single-cell assessment. (A) Number of significantly perturbed genes following network-based in silico knockout of HPV E1, E2, E5, E6, and E7, which yielded 15, 89, 91, 8, and 86 significantly perturbed genes,

respectively. (B) GEPIA Kaplan-Meier disease-free survival curves for ABCC3, COL6A1, STARD10, and AOC1, stratified by high and low expression; log-rank *p* values are displayed in the individual plots. (C) Violin plots comparing candidate-gene expression between cervical cancer and paired adjacent-normal cells in GSE168652. ABCC3, COL6A1, and STARD10 show higher expression in adjacent-normal cells, whereas AOC1 remains weakly expressed and shows no significant between-group difference. Adjusted *p* values are displayed above the comparisons. (D) Smoothed expression trends along computational pseudotime in GSE168652. ABCC3 and COL6A1 show pronounced intermediate-pseudotime peaks, STARD10 shows a weaker intermediate pattern, and AOC1 remains weakly expressed without a clear peak. Because GSE168652 contains paired tissues from a single patient, panels C and D provide supportive external evidence rather than population-level replication. Neither the in silico knockout nor pseudotime analysis establishes a causal relationship

After quality control and normalization, 11,280 cells and 22,100 genes were retained from GSE168652, which contains cervical cancer tissue and paired adjacent-normal tissue from a single patient. Dimensionality reduction showed a clear separation between the Adjacent.normal and Cancer groups in low-dimensional space, indicating differences in their overall transcriptional states. Group-level expression and pseudotime dynamics were then compared for the four candidate genes.

ABCC3, COL6A1, and STARD10 were expressed at significantly higher levels in the Adjacent.normal group than in the Cancer group, whereas AOC1 showed low overall expression in both groups and no significant between-group difference (Figure 7C).

Pseudotime analysis arranged the cells along a continuous distribution of computational states. ABCC3, COL6A1, and STARD10 showed relatively high expression in the middle of pseudotime followed by a decline at later positions, with more pronounced intermediate increases for ABCC3 and COL6A1. AOC1 remained weakly expressed and did not show a clear intermediate-peaking pattern (Figure 7D). This computational ordering describes variation across cell states and does not establish a chronological sequence or a causal progression.

Taken together, the group comparison and pseudotime results in GSE168652 provided additional support for higher adjacent-normal expression and intermediate-pseudotime peaks of COL6A1, ABCC3, and STARD10. Support for AOC1 was comparatively weaker. Because GSE168652 contains paired tissues from only one patient, these findings provide supportive external evidence rather than population-level replication. Auxiliary analysis of the peripheral-blood dataset GSE120691 showed trends toward higher ABCC3 and AOC1 expression in the CIN group; this result was considered an auxiliary observation outside the independent single-cell assessment.

4. Discussion

This study integrated HPV transcript abundance, malignant epithelial heterogeneity, inferred cell-state variation, and microenvironmental communication within a unified analytical framework encompassing 105,799 high-quality single cells. The results showed that HPV-related cervical lesions were not characterized by a single homogeneous malignant state. Instead, HPV^{high}/MP6-high and HPV^{dim}/MP6-low states coexisted within the putative malignant epithelial population.

HPV^{high}/MP6-high cells showed stronger HPV transcript signals and greater MP6 activity and were characterized by programs related to proliferation, metabolism, and stress adaptation [16]. In contrast, HPV^{dim}/MP6-low cells showed more prominent immune-recognition and microenvironmental-interaction features [17]. CellChat analysis predicted similar aggregate interaction numbers and strengths in the two states

but different distributions of inferred interactions. The HPV^{high}/MP6-high state was biased toward epithelial-stromal interactions such as COLLAGEN and LAMININ signaling, whereas the HPV^{dim}/MP6-low state was biased toward immune-recognition interactions such as MHC-I signaling. Compared with N, H showed fewer inferred interactions and lower aggregate interaction strength, and the network focus shifted from immune-related connections toward local networks centered on epithelial cells and epithelial-stromal interactions. These differences were more consistent with network contraction and functional reorganization than with a uniform increase in communication.

Among the 10 metaprograms, MP6 showed the strongest association with HPV transcript abundance and included HPV E1, E6, and E7. MP6 also showed the most significant overlap with the intermediate-peaking C2 dynamic module, which was enriched for early and late estrogen responses, apical junctions, and epithelial structural changes [18, 19]. A conservative interpretation is that the HPV-associated program coincides with an intermediate window of inferred cell-state variation rather than a single linear progression.

The network-based in silico knockout results were consistent with the observed association between MP6 and the C2 dynamic module. The significantly perturbed gene sets following E2, E5, and E7 knockout showed the clearest correspondence with MP6, C2, and their intersecting gene set. Estrogen responses, oxidative phosphorylation, hypoxia, apoptosis, the p53 pathway, IL-2/STAT5 signaling, TNF-α signaling via NF-κB, and epithelial-mesenchymal transition recurred across multiple knockout models. Because these findings were derived from network inference, their principal value lies in generating mechanistic hypotheses and prioritizing candidate genes; experimental functional studies are still required.

COL6A1, ABCC3, STARD10, and AOC1 were prioritized through convergent evidence from viral perturbation, the association between MP6 and the C2 dynamic module, pseudotime, and prognosis. These genes were linked, respectively, to extracellular-matrix organization and epithelial-stromal interactions, membrane transport and stress tolerance, lipid transport and membrane-lipid homeostasis, and amine metabolism and the local metabolic environment [20-24]. GSE168652 provided clearer support for COL6A1, ABCC3, and STARD10 in the Adjacent.normal group and at intermediate pseudotime positions, whereas support for AOC1 was limited.

Overall, the findings associated HPV transcript abundance with malignant epithelial heterogeneity, intermediate pseudotime dynamics, and redistribution of cell-cell communication networks at single-cell resolution. The communication analyses further showed that the two malignant epithelial states had similar aggregate interaction numbers and strengths but distinct interaction content. Compared with HPV-positive normal cervical tissue, HSIL samples showed fewer inferred interactions, lower aggregate interaction strength, and a network focus shifted toward local epithelial and epithelial-stromal interactions. Together with the significant association between MP6 and the C2 dynamic module, these observations associated higher HPV transcript abundance with variation in estrogen-response, epithelial-structure, and microenvironmental programs across the sampled lesion categories.

These interpretations remain bounded by the analytical design. Pseudotime is a computational ordering of cell states and cannot by itself establish chronology or causality. Network-based in silico knockout is a computational perturbation and does not replace experimental gene knockout. In addition, GSE168652 contains paired samples from only one patient and therefore does not constitute population-level replication; within this dataset, independent support for AOC1 was weaker than that for COL6A1, ABCC3, and STARD10. Subject to these limitations, the study provides single-cell evidence relevant to HPV-associated cervical lesion biology and supports subsequent functional validation in larger, multi-patient cohorts.

5. Conclusion

By integrating HPV transcript quantification with malignant epithelial metaprograms, pseudotime, cell-cell communication inference, network-based in silico knockout, prognostic analysis, and assessment in an independent dataset, this study characterized two HPV-associated malignant epithelial states and identified a significant association between MP6 and the intermediate-peaking C2 dynamic module. CellChat analysis predicted redistribution of communication networks across malignant states and lesion categories, whereas computational perturbation and multidimensional evidence prioritized COL6A1, ABCC3, STARD10, and AOC1 as candidate genes. External single-cell evidence was clearer for COL6A1, ABCC3, and STARD10 and weaker for AOC1. These analyses provide a hypothesis-generating framework for understanding HPV-associated epithelial-state heterogeneity and microenvironmental remodeling, but they do not establish causal relationships. Experimental validation and replication in larger, multi-patient cohorts are required.

Author contribution

H. Z. was responsible for data analysis and manuscript writing. W. C. provided guidance on data analysis. Y. L. was responsible for sample and data collection. W. C. and W. W. designed and supervised the entire research process. All authors critically reviewed the draft, provided important feedback, and collectively approved the final version of the manuscript.

Funding

The Key Research and Development Program of Ningxia Hui Autonomous Region (2022BEG02055)

Data availability

The public datasets analyzed in this study are available from the NCBI Gene Expression Omnibus under accession numbers GSE208653, GSE168652, and GSE120691. Raw sequencing data associated with GSE208653 are accessible through the GEO record and linked Sequence Read Archive accessions. The scRNA-seq data of the cervical cancer sample are available from the authors upon request.

Declaration of interests

The authors declare no competing financial or non-financial interests.

References

[1] Pimple, S., & Mishra, G. (2022). Cancer cervix: Epidemiology and disease burden. *CytoJournal, 19*, Article 21. https://doi.org/10.25259/CMAS_03_02_2021

[2] Muñoz, N., Bosch, F. X., de Sanjosé, S., Herrero, R., Castellsagué, X., Shah, K. V., Snijders, P. J. F., Meijer, C. J. L. M., & International Agency for Research on Cancer Multicenter Cervical Cancer Study Group. (2003). Epidemiologic classification of human papillomavirus types associated with cervical cancer. *New England Journal of Medicine, 348*(6), 518–527. https://doi.org/10.1056/NEJMoa021641

[3] Walboomers, J. M. M., Jacobs, M. V., Manos, M. M., Bosch, F. X., Kummer, J. A., Shah, K. V., Snijders, P. J. F., Peto, J., Meijer, C. J. L. M., & Muñoz, N. (1999). Human papillomavirus is a necessary cause of invasive

cervical cancer worldwide. *Journal of Pathology, 189*(1), 12–19. https://doi.org/10.1002/(SICI)1096-9896(199909)189:1<12::AID-PATH431>3.0.CO;2-F

[4] Scheffner, M., Werness, B. A., Huibregtse, J. M., Levine, A. J., & Howley, P. M. (1990). The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. *Cell, 63*(6), 1129–1136. https://doi.org/10.1016/0092-8674(90)90409-8

[5] Münger, K., Werness, B. A., Dyson, N., Phelps, W. C., Harlow, E., & Howley, P. M. (1989). Complex formation of human papillomavirus E7 proteins with the retinoblastoma tumor suppressor gene product. *EMBO Journal, 8*(13), 4099–4105. https://doi.org/10.1002/j.1460-2075.1989.tb08594.x

[6] Dyson, N., Howley, P. M., Münger, K., & Harlow, E. (1989). The human papillomavirus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product. *Science, 243*(4893), 934–937. https://doi.org/10.1126/science.2537532

[7] Giarrè, M., Caldeira, S., Malanchi, I., Ciccolini, F., Leão, M. J., & Tommasino, M. (2001). Induction of pRb degradation by the human papillomavirus type 16 E7 protein is essential to efficiently overcome p16INK4a-imposed G1 cell cycle arrest. *Journal of Virology, 75*(10), 4705–4712. https://doi.org/10.1128/JVI.75.10.4705-4712.2001

[8] Münger, K., Baldwin, A., Edwards, K. M., Hayakawa, H., Nguyen, C. L., Owens, M., Grace, M., & Huh, K. (2004). Mechanisms of human papillomavirus-induced oncogenesis. *Journal of Virology, 78*(21), 11451–11460. https://doi.org/10.1128/JVI.78.21.11451-11460.2004

[9] Moody, C. A., & Laimins, L. A. (2010). Human papillomavirus oncoproteins: Pathways to transformation. *Nature Reviews Cancer, 10*(8), 550–560. https://doi.org/10.1038/nrc2886

[10] Pal, A., & Kundu, R. (2020). Human papillomavirus E6 and E7: The cervical cancer hallmarks and targets for therapy. *Frontiers in Microbiology, 10*, Article 3116. https://doi.org/10.3389/fmicb.2019.03116

[11] Scarth, J. A., Patterson, M. R., Morgan, E. L., & Macdonald, A. (2021). The human papillomavirus oncoproteins: A review of the host pathways targeted on the road to transformation. *Journal of General Virology, 102*(3), Article 001540. https://doi.org/10.1099/jgv.0.001540

[12] Li, C., Wu, H., Guo, L., Liu, D., Yang, S., Li, S., & Hua, K. (2022). Single-cell transcriptomics reveals cellular heterogeneity and molecular stratification of cervical cancer. *Communications Biology, 5*(1), Article 1208. https://doi.org/10.1038/s42003-022-04142-w

[13] Yue, S., Wang, Q., Zhang, J., Hu, Q., & Liu, C. (2023). Understanding cervical cancer at single-cell resolution. *Cancer Letters, 576*, Article 216408. https://doi.org/10.1016/j.canlet.2023.216408

[14] Pu, C., Xing, B., Wang, S., Liu, Z., & Zhao, Y. (2026). Advancements in single-cell sequencing for cervical cancer research. *Molecular and Cellular Biochemistry, 481*(2), 615–637. https://doi.org/10.1007/s11010-025-05407-8

[15] Akagi, K., Symer, D. E., Mahmoud, M., Jiang, B., Goodwin, S., Wangsa, D., Li, Z., Xiao, W., Dunn, J. D., Ried, T., Coombes, K. R., Sedlazeck, F. J., & Gillison, M. L. (2023). Intratumoral heterogeneity and clonal evolution induced by HPV integration. *Cancer Discovery, 13*(4), 910–927. https://doi.org/10.1158/2159-8290.CD-22-0900

[16] An, J., Mo, D., Liu, H., Veena, M. S., Srivatsan, E. S., Massoumi, R., & Rettig, M. B. (2008). Inactivation of the CYLD deubiquitinase by HPV E6 mediates hypoxia-induced NF-κB activation. *Cancer Cell, 14*(5), 394–407. https://doi.org/10.1016/j.ccr.2008.10.007

[17] Manzo-Merino, J., Del-Toro-Arreola, S., Rocha-Zavaleta, L., Peralta-Zaragoza, Ó., Jiménez-Lima, R., & Madrid-Marina, V. (2020). Immunology of cervical cancer. *Revista de Investigación Clínica, 72*(4), 188–197. https://doi.org/10.24875/RIC.20000057

[18] James, C. D., Morgan, I. M., & Bristol, M. L. (2020). The relationship between estrogen-related signaling and human papillomavirus-positive cancers. *Pathogens, 9*(5), Article 403. https://doi.org/10.3390/pathogens9050403

[19] Popova, N. V., & Jücker, M. (2022). The functional role of extracellular matrix proteins in cancer. *Cancers, 14*(1), Article 238. https://doi.org/10.3390/cancers14010238

[20] Hou, T., Tong, C., Kazobinka, G., Zhang, W., Huang, X., Huang, Y., & Zhang, Y. (2016). Expression of COL6A1 predicts prognosis in cervical cancer patients. *American Journal of Translational Research, 8*(6), 2838–2844.

[21] Ghanem, C. I., & Manautou, J. E. (2019). Modulation of hepatic MRP3/ABCC3 by xenobiotics and pathophysiological conditions: Role in drug pharmacokinetics. *Current Medicinal Chemistry, 26*(7), 1185–1223. https://doi.org/10.2174/0929867325666180221142315

[22] Ramírez-Cosmes, A., Reyes-Jiménez, E., Zertuche-Martínez, C., Hernández-Hernández, C. A., García-Román, R., Romero-Díaz, R. I., Manuel-Martínez, A. E., Elizarrarás-Rivas, J., & Vásquez-Garzón, V. R. (2021). The implications of ABCC3 in cancer drug resistance: Can we use it as a therapeutic target? *American Journal of Cancer Research, 11*(9), 4127–4140.

[23] Olayioye, M. A., Vehring, S., Müller, P., Herrmann, A., Schiller, J., Thiele, C., Lindeman, G. J., Visvader, J. E., & Pomorski, T. (2005). StarD10, a START domain protein overexpressed in breast cancer, functions as a phospholipid transfer protein. *Journal of Biological Chemistry, 280*(29), 27436–27442. https://doi.org/10.1074/jbc.M413330200

[24] Maintz, L., & Novak, N. (2007). Histamine and histamine intolerance. *American Journal of Clinical Nutrition, 85*(5), 1185–1196. https://doi.org/10.1093/ajcn/85.5.1185