ONCOLOGY

Nanovaccine targeting in colorectal cancer: a multi-dataset analysis of CEA expression, cytokine profiles, and co-expressed genes

Razvan-Septimiu Zdrehus[1], Cristina Mitrea[2], Lucian Mocan[1]

1) Nanomedicine Department, Regional Institute of Gastroenterology and Hepatology, Iuliu Hatieganu University of Medicine and Pharmacy, Cluj Napoca, Romania

2) Department of Computational Medicine and Bioinformatics, University of Michigan, USA

Abstract

Background. Carcinoembryonic antigen (CEA/CEACAM5) is a well-established tumor-associated antigen overexpressed in epithelial malignancies, including colorectal cancer (CRC). While its diagnostic and therapeutic relevance is recognized, its immunological context and potential as a nanovaccine target remain underexplored.

Aim. This study aims to enable the rational design and refinement of CEA-based nanovaccines by integrating transcriptomic and spatial data to identify immunologically relevant co-expressed biomarkers and potential therapeutic targets.

Methods. We conducted an integrative bioinformatics analysis using transcriptomic data from TCGA-COAD, GEO, and spatial datasets (GSE207843, GSE226997), complemented by differential gene expression analysis (GSE245218). CEACAM5 expression was correlated with cytokine profiles (IL10, IFNG, TNF, IL1B, IL12A, IL4), immune cell infiltration (via xCell), and co-expression networks. Genes with Spearman $\rho > 0.75$ were prioritized as vaccine candidates and evaluated through oncofetal expression and literature curation.

Results. CEACAM5 expression was inversely correlated with IFNG, IL10, TNF, and IL1B, suggesting a potential immunosuppressive phenotype. xCell analysis revealed negative trends between CEACAM5 and effector immune populations including CD8^{+} T cells and NK cells. Spatial transcriptomics confirmed CEACAM5 compartmentalization in tumor epithelium with minimal cytokine overlap. Co-expression analysis identified EPCAM and ATP10B as high-confidence candidates. Embryonic vs. adult differential analysis (GSE245218) confirmed their oncofetal expression patterns. Gene ontology analysis revealed downregulation of antibacterial humoral immune pathways.

Conclusion. CEACAM5 defines a distinct immune-silent tumor phenotype and co-localizes with other vaccine-relevant genes such as EPCAM. This study provides a comprehensive immunogenomic rationale for CEACAM5-directed nanovaccine development and proposes EPCAM and ATP10B as co-targets based on tumor-specific and developmental expression profiles.

Keywords: CEACAM5, colorectal cancer, EpCAM, tumor-associated antigen, nanovaccine

DOI: 10.15386/mpr-2917

Manuscript received: 14.07.2025
Received in revised form: 02.07.2025
Accepted: 18.07.2025

Address for correspondence:
Razvan-Septimiu Zdrehus
razvan.zdrehus@umfcluj.ro

Background and aim

Carcinoembryonic antigen (CEA, also known as CEACAM5) is a glycosylated cell adhesion molecule that is frequently overexpressed in a variety of epithelial malignancies, with colorectal cancer being one of the most prominent examples [1]. This restricted expression in normal adult tissues, contrasted with its robust upregulation in tumor cells, has established CEACAM5 as both a clinically relevant biomarker and a prime target for immunotherapeutic strategies—ranging from monoclonal antibodies to innovative nanovaccines [2].

In addition to its role as a tumor-associated antigen, CEACAM5 belongs to a family of molecules with known immunomodulatory activity, including CEACAM1, which is involved in T cell inhibition and myeloid regulation [3]. Previous studies have suggested that CEACAM5 may contribute to immune evasion by interfering with cytokine signaling or immune cell recruitment in the tumor microenvironment [4]. Therefore, investigating the spatial and transcriptomic relationships between CEACAM5 and key cytokines such as IL10, IFNG, TNF, and IL1B may provide important insights into its immunological context and help define tumor phenotypes characterized by immune suppression. Understanding this relationship is particularly relevant for optimizing nanovaccine targeting strategies and identifying immune-silent tumor regions that may benefit from combination immunotherapies.

Furthermore, CEACAM5 has been shown to actively promote colorectal cancer progression via epithelial–mesenchymal transition and activation of MAPK (mitogen-activated protein kinase) signaling, reinforcing its functional role in tumor aggressiveness and its confinement to epithelial compartments [5].

To address this, we leveraged the rich resource of The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) cohort, which provides comprehensive RNA-seq data, clinical annotation, and immune context for hundreds of colorectal cancer samples. By correlating CEACAM5 expression with cytokine signatures, as well as with key clinical parameters like pathological tumor stage, we sought to clarify how CEACAM5 may interact with and potentially modulate the immune milieu in colorectal tumors. The scale and granularity of TCGA-COAD allow for robust statistical analyses and offer unique insights into both tumor biology and the interplay between malignant and immune cells.

To validate the tissue-level compartmentalization of CEACAM5, we further incorporated spatial transcriptomic data from human colorectal tumors (Gene Expression Omnibus(GEO), GSE226997), enabling visualization of its expression in situ relative to immune and stromal architecture [6,7].

At the same time, we recognized a pressing need to look beyond CEACAM5 itself. Modern cancer immunotherapy increasingly relies on multi-antigen targeting and comprehensive biomarker panels [8]. To this end, we developed a computational pipeline designed to identify novel genes that are co-expressed with CEACAM5 in colorectal tumors. By characterizing these candidates in terms of developmental, functional, and immunological profiles, we aimed to reveal new biomarkers or vaccine targets with similar tumor specificity and translational promise. This dual approach—rooted in both in-depth analysis of CEACAM5 and systematic candidate discovery—provides a framework for understanding the molecular networks underlying tumor immune evasion and for guiding the rational design of next-generation nanovaccines.

Recent in vivo experiments, such as those by Zdrehus et al. [9], have demonstrated that CEA-functionalized gold nanoparticles (CEA-AuNPs) can modulate cytokine profiles and induce systemic immunomodulation in murine models. These studies provide strong evidence that CEA-AuNP administration results in measurable changes in both pro-inflammatory and anti-inflammatory cytokines, supporting the immunogenic and translational potential of such nanovaccine platforms. However, the precise mechanisms underlying these effects, and their relevance to human tumor immunology, remain to be fully elucidated. By integrating transcriptomic analyses from TCGA-COAD with targeted candidate discovery and spatial immune profiling, our study seeks to bridge the gap between preclinical experimental findings and clinical translation. This integrative, exploratory approach provides a comprehensive immunological rationale for the rational design and refinement of CEA-based nanovaccines and for identifying new biomarkers or therapeutic targets that reflect the real-world complexity of the tumor microenvironment.

Methods

Dataset selection

To investigate the immunological context of CEA expression, we curated publicly available transcriptomic datasets from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO). Datasets were selected based on relevance to CEA-overexpressing tumors and inclusion of immune-related gene expression data. RNA-seq data from TCGA Colon Adenocarcinoma (TCGA-COAD) were used for bulk transcriptomic analysis. In addition, spatial transcriptomic data from murine spleen and liver (GSE207843) and embryonic vs. adult differential gene expression data (GSE245218) were analyzed (sourced from GEO). Clinical annotations, including pathological stage, were obtained from the corresponding TCGA-COAD clinical matrix.

Gene expression processing and co-expression analysis

Raw gene-level counts for the TCGA-COAD cohort (n = 524 tumors) were retrieved using the TCGAbiolinks (v2.28.3) package in R (v4.5.1), specifying workflow.type = “STAR - Counts”. Ensembl transcript IDs were mapped to HGNC gene symbols with biomaRt (v2.48.3). Count matrices were normalized by applying a $\log_2$(counts + 1) transformation. From these normalized data, we extracted expression values for CEACAM5 and a panel of key cytokine genes (IL10, IL12A, IL4, IFNG, TNF, IL1B), retaining only protein-coding genes for downstream analysis.

To quantify the relationship between CEACAM5 and immune signaling, we computed Spearman correlation coefficients between CEACAM5 and each cytokine gene using base R's cor() and assessed significance with cor.test(), applying the Benjamini–Hochberg method to control the false discovery rate. To support interpretation of the CEACAM5–cytokine associations, we visualized correlation patterns using heatmaps and group-wise expression differences using boxplots. Heatmaps of Spearman correlation coefficients between CEACAM5 and selected cytokines (IFNG, IL10, IL1B, IL12A, IL4, TNF) were created using ggplot2 and reshape2, with color gradients representing the strength and direction of correlation. Additionally, expression values ($\log_2$ read counts) for each cytokine were compared between CEACAM5-high and CEACAM5-low quartiles using boxplots. These plots included median lines and interquartile ranges, and Wilcoxon rank-sum test p-values were annotated to highlight group differences.

Immune cell infiltration estimation

To profile the immune context of CEACAM5 expression, we applied the xCell2 algorithm (v1.0.6) [10], using two curated reference atlases—ImmuneCompendium and BlueprintEncode—on the TCGA-COAD expression matrix, which provides enrichment scores for all immune and stromal cell types in the reference sets. Samples were stratified by CEACAM5 expression quartiles as described in section 2.2.

Spatial transcriptomics data processing

Spatially resolved transcriptomic data (GSE207843) from the Gene Expression Omnibus (GEO) comprising 1,218 spots from mouse spleen and liver were obtained as a digital gene expression (DGE) matrix (DGE_matrix_min100.txt.gz). The counts matrix (35,388 genes × 1,218 spots) was loaded into R and converted into a Seurat object (CreateSeuratObject)[11]. Standard preprocessing steps included log-normalization (NormalizeData), identification of 2,000 highly variable features (FindVariableFeatures), and data scaling (ScaleData). Dimensionality reduction was performed via principal component analysis (RunPCA) and Uniform Manifold Approximation and Projection (UMAP) for visualization (RunUMAP, dims 1:20). Feature expression of CEACAM5, IL10, IFNG, TNF, and IL1B was visualized on the UMAP embedding using FeaturePlot. Spot-wise Spearman correlations between CEACAM5 and each cytokine were calculated on the normalized data matrix (GetAssayData(slot=”data”)) to quantify spatial co-expression patterns.

Spatial transcriptomic analysis of human colorectal cancer (GSE226997)

To validate the spatial distribution of CEACAM5 expression in human colorectal cancer tissue, we analyzed publicly available Visium spatial transcriptomics data from the GSE226997 dataset (human CRC, GEO). Raw data were processed using Seurat (v4.3.0). The filtered feature-barcode matrix, tissue image, scalefactors, and spatial coordinates were used to construct a spatial Seurat object via Load10X_Spatial(). Gene expression was log-normalized, and spatial feature plots (SpatialFeaturePlot()) were generated for CEACAM5. The spatial distribution was visualized relative to histological tumor architecture to assess compartmentalization and co-localization with immune-related regions.

Pathologic stage analysis using TCGA-COAD

Clinical annotations, including detailed pathologic stage, were obtained from the associated TCGA-COAD clinical matrix (TCGA.COAD.sampleMap-COAD_clinicalMatrix). Expression values for CEACAM5 were extracted and merged with the corresponding clinical data by TCGA barcode. Pathologic stage annotations were harmonized and categorized to reflect all available detailed stages (e.g., Stage I, II, IIA, IIB, III, IIIA, IIIB, IIIC, IV, IVA). For statistical visualization, CEACAM5 expression distributions were compared across pathologic stages using boxplots generated in R using the base graphics package. Only samples with available stage annotation were included in the analysis.

Automated association and validation pipeline

The association analysis was performed as follows. For unbiased candidate discovery, raw gene-level counts for the TCGA-COAD cohort (n = 524 tumors) were downloaded via TCGAbiolinks in R (v2.28.3), using STAR-aligned read counts, as preprocessed in section 2.2 [12,13]. We then calculated Spearman correlations between CEACAM5 and each of the ~20,000 protein-coding genes using base R's cor(), selecting those with $\rho > 0.75$ as top co-expression candidates (e.g., EPCAM, STK38, CDH1, ATP10B). To validate these findings, samples were stratified by CEACAM5 expression quartiles as described in Section 2.2 and expression differences for both the key cytokines and the four candidate genes were assessed via two-sided Wilcoxon rank-sum tests. All visualizations—heatmaps, boxplots, and sample-wise line plots—were produced with ggplot2 (v3.3.5) and reshape2 (v1.4.4) to ensure consistency and reproducibility.

Differential expression and functional enrichment (GSE245218)

To perform an extra validation, we did a differential gene expression analysis on a related dataset from the Gene Expression Omnibus (GEO)[14]. This specific dataset (GSE245218) [15] contains data from a study that systematically analyzed gene expression in mouse embryos at days E11 and E18, compared to adult mice, and examined whether these embryonic genes are overexpressed in isolated tumor endothelial cells from mice [16]. Certain embryonic genes, like carcinoembryonic antigens (CEA) and oncofetal antigens, are reactivated during cancer development. We hypothesized, as the authors in the study by Huijbers et al., that tumor endothelial cells might show a similar pattern.

Therefore, we pooled the data from the two days for the embryos and compared that data with the data from the adult mice. We performed differential gene expression analysis using the DESeq2 package in R [17]. DESeq2 (v1.34.0) was used for DGE (design: E vs. A). Shrunk $\log_2$FC values were obtained with lfcShrink(type=”apeglm”). Genes with FDR < 0.01 and $|\log_2\text{FC}| > 2$ were considered significant. We used the results to evaluate our 4 candidates from the previous association analysis (dge_analysis_results.tsv). We used a threshold of 0.01 for the p-values from this analysis and a 4-fold change threshold for the change in expression between groups (embryo vs. adult) to select the best gene candidate as these thresholds show a non-random change, due to the low p-value, and a high change, due to the high ratio for the change.

We further performed functional analysis with all the results of the differential expression analysis, to see what biological processes are significantly disrupted in the transition from the embryo state to the adult state and which of those might be relevant for cancer progression. The full ranked list of all genes ($\log_2$FC) was input into clusterProfiler’s gseGO() for GO term enrichment in our top up- and down- regulated genes [18]. We report terms with adjusted $p < 0.05$.

Statistical analysis and software reproducibility

All analyses were conducted using R (v4.5.1). Statistical comparisons between CEACAM5-high and CEACAM5-low expression groups were performed using the Wilcoxon rank-sum test, appropriate for non-parametric data. Correlation analysis was conducted using Spearman coefficients, and p-values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) method. Statistical significance was defined as FDR-adjusted $p < 0.05$.

All code was executed using an automated and fully documented script (analysis_CEA.R), provided as supplementary material to enable reproducibility and custom adaptation. Key packages included: TCGAbiolinks, SummarizedExperiment, biomaRt, DESeq2, ggplot2, ComplexHeatmap, Seurat, reshape2, clusterProfiler, and enrichplot.

Results

CEACAM5 expression correlation with immune cytokine profiles

To investigate the immunological context of CEACAM5 expression in colorectal cancer, we analyzed transcriptomic data from TCGA-COAD. Association analysis using Spearman correlation revealed inverse relationships between CEACAM5 expression and several key cytokines involved in immune regulation, notably IFNG ($\rho = -0.233$, $p = 0.11$), IL10 ($\rho = -0.200$, $p = 0.08$), and TNF ($\rho = -0.110$, $p > 0.05$) (Figure 1). These correlations did not reach statistical significance on this dataset, however, the observed trends suggest a potential role for CEACAM5 expression in modulating cytokine activity within the tumor microenvironment, warranting further investigation in larger studies or using meta-analyses.

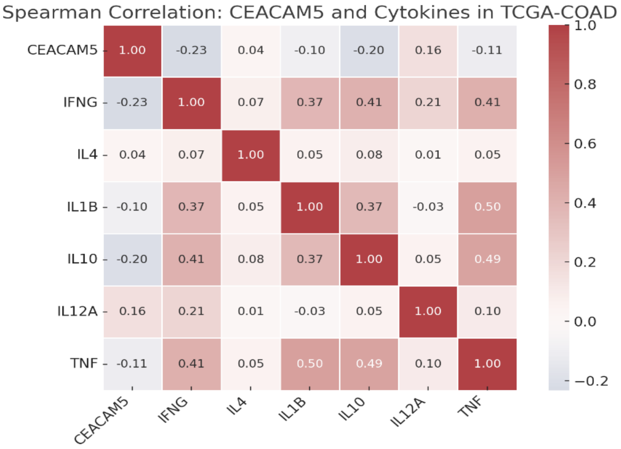

Figure 1. Heatmap depicting Spearman correlation coefficients between CEACAM5 expression and selected cytokines (IFNG, IL10, TNF, IL12A, IL4, IL1B) in TCGA colorectal adenocarcinoma (COAD) samples. Negative correlations are highlighted, indicating reduced cytokine expression associated with high CEACAM5 levels, suggesting a role in suppressing immune-related signaling pathways.

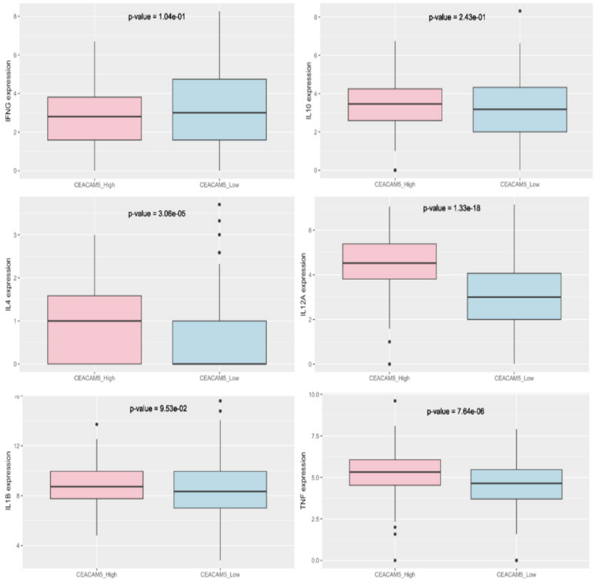

Figure 2. Cytokine expression in CEACAM5-high versus CEACAM5-low TCGA-COAD samples. Boxplot comparison of cytokine gene expression (IL10, IL12A, IL4, IFNG, IL1B, TNF) between samples stratified by CEACAM5 expression quartiles. Expression values are log2-transformed read counts. Median and interquartile ranges shown.

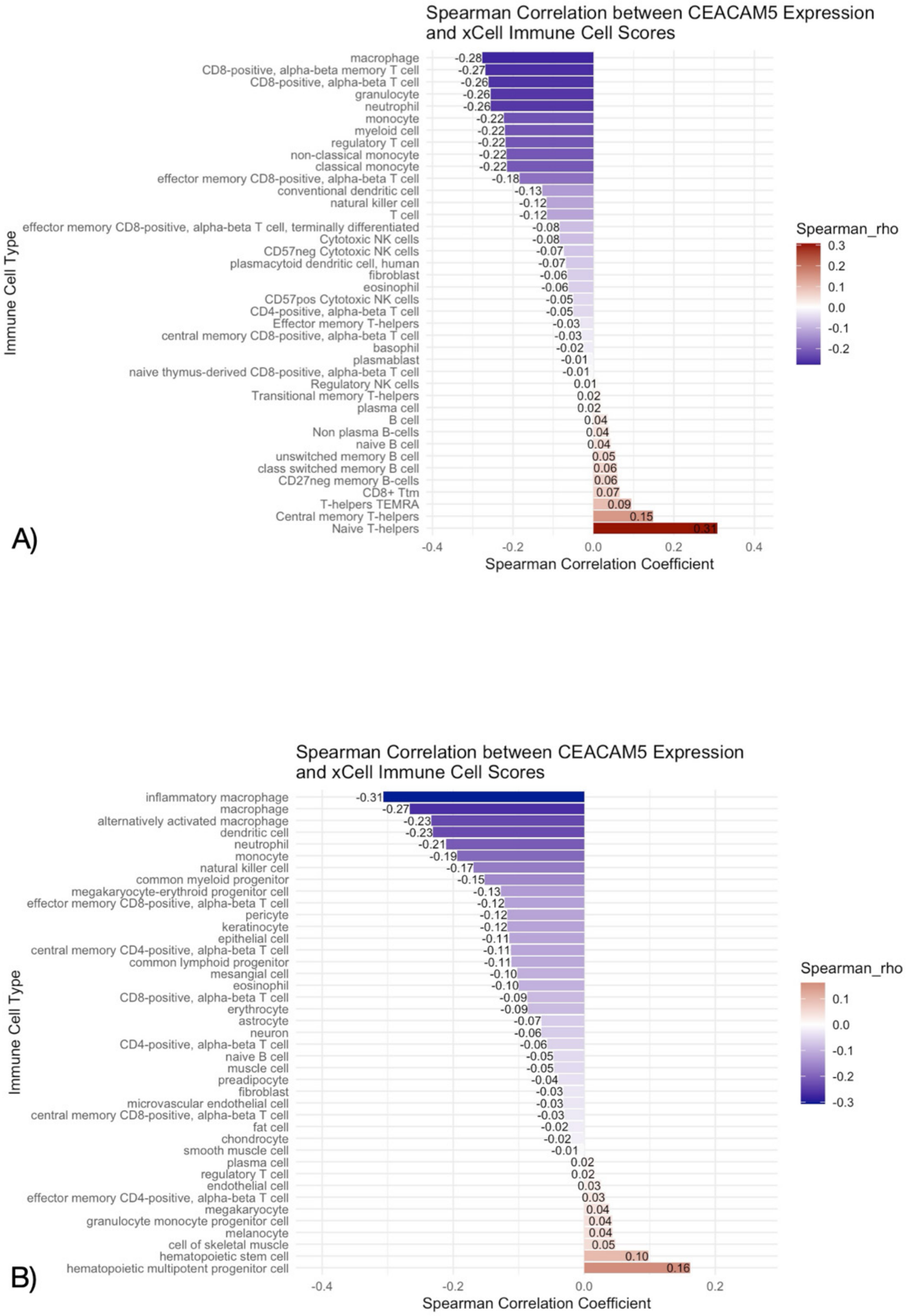

Figure 3. Spearman correlation between CEACAM5 expression and immune cell enrichment scores in colorectal cancer via xCell2. A) Correlation profile using the ImmuneCompendium reference signature matrix. B) Correlation profile using the BlueprintEncode reference matrix. Blue bars represent negative correlations, and red bars indicate positive correlations. Cell types are ordered by descending ρ. Only the top 40 cell types by absolute correlation are shown for each panel.

Boxplot analyses further supported these observations, demonstrating higher levels of IFNG, IL10, and IL1B expression in tumors with low CEACAM5 expression compared to those with high CEACAM5 expression (Figure 2). However, these differences were not statistically significant, reinforcing the need for additional research to clarify these relationships.

Immune cell infiltration analysis via xCell

Immune cell infiltration was assessed using the xCell algorithm, by providing enrichment scores for various immune populations, sourced from two curated reference atlases, ImmuneCompendium and BlueprintEncode, on the TCGA-COADexpresion matrix. Using both reference sets, we observed consistent negative correlations between CEACAM5 and a range of immune cell types.

From the ImmuneCompendium reference (Figure 3A), inflammatory macrophages (ρ = −0.31), macrophages (ρ = −0.27), and alternatively activated macrophages (ρ = −0.23) were among the most negatively associated with CEACAM5 expression, followed by neutrophils, dendritic cells, and NK cells (ρ < −0.2). BlueprintEncode results (Figure 3B) corroborated these findings, highlighting strong negative correlations with macrophages (ρ = −0.28), memory and effector CD8$^+$ T cells (ρ ≈ −0.26), neutrophils, and regulatory T cells.

Despite variability across reference sets, a consistent trend emerged: CEACAM5-high tumors exhibit reduced scores for multiple immune subsets, particularly myeloid and cytotoxic compartments. These findings suggest that CEACAM5 expression is associated with an immune-excluded or suppressed tumor microenvironment, aligning with its hypothesized role in immune evasion.

Spatial transcriptomics of spleen and liver in mice

To explore CEACAM5's regional immuno-modulatory role, we analyzed mouse spatial transcriptomics data from spleen and liver tissues (GSE207843). CEACAM5 expression was visualized alongside four key cytokines—IL10, IFNG, TNF, and IL1B—on UMAP embeddings. Ceacam5 was detected in distinct spatial clusters, which showed minimal overlap with regions expressing pro- or anti-inflammatory cytokines. Quantitatively, Spearman correlations between Ceacam5 and these cytokines were negligible (ρ = –0.002 to 0.012), suggesting that CEACAM5 expression in homeostatic spleen and liver tissue in mice is spatially decoupled from inflammatory transcriptional activity (Figure 4).

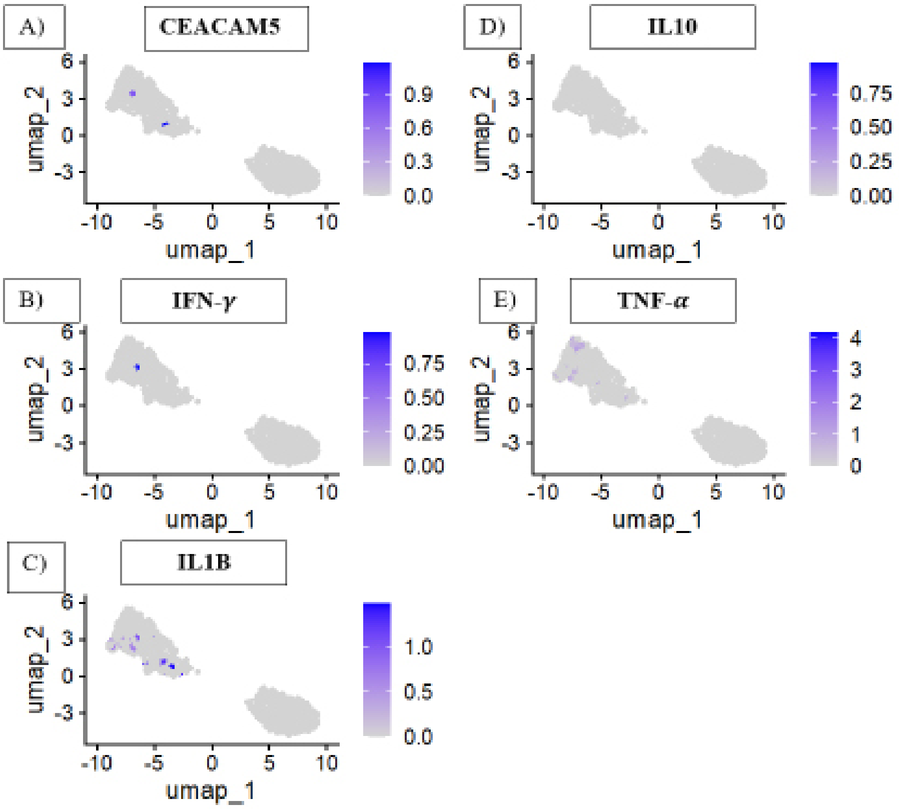

Figure 4. UMAP FeaturePlots of Ceacam5 and cytokines in mouse spleen/liver (GSE207843). UMAP embedding of 1,218 spatial spots colored by (A) CEACAM5, (B) IFN-, (C) IL1B, (D) IL10, and (E) TNF- expression (log-normalized). Color scales indicate relative expression intensity across spots. Spearman correlations between Ceacam5 and each cytokine are provided in the main text.

Visually, Ceacam5 expression (panel A) appears in discrete cell clusters that do not colocalize strongly with hotspots of IFN-gamma (panel B), IL1B (panel C), IL10 (panel D), or TNF-alfa (panel E). Quantitatively, spot-wise Spearman correlation between Ceacam5 and IFN-gamma is essentially null (ρ = –0.002), with similarly low correlations for IL10 (ρ ≈ –0.005), IL1B (ρ ≈ 0.012), and TNF-alfa (ρ ≈ 0.009). These minimal correlations indicate that, in homeostatic murine spleen and liver tissue, CEACAM5 expression is not directly coupled to regional pro- or anti-inflammatory cytokine transcription, suggesting that CEACAM5's immunomodulatory effects may require additional context or co-stimuli.

Spatial transcriptomic analysis of human colorectal cancer (GSE226997)

To validate the spatial context of CEACAM5 expression, we analyzed spatial transcriptomics data from human primary colorectal cancer (GSE226997, GEO) using the Seurat pipeline [11]. SpatialFeaturePlot visualization revealed that CEACAM5 is predominantly expressed in epithelial compartments of the tumor, with minimal signal in surrounding stromal or immune-rich zones (Figure 5). This localized expression pattern reinforces CEACAM5's relevance as a tumor-restricted marker and highlights its utility for targeted delivery platforms such as nanovaccines. The spatial restriction further supports prior findings from TCGA-COAD indicating strong tumor-specific overexpression of CEACAM5 and minimal co-expression with key inflammatory cytokines.

Pathologic stage analysis using TCGA-COAD

To explore the clinical correlates of CEACAM5 expression, we stratified tumor samples from the TCGA-COAD cohort according to detailed American Joint Committee on Cancer (AJCC) pathological stages (I, IIa, IIb, IIIa, IIIb, IIIc, IV, IVa). As illustrated in Figure 6, CEACAM5 expression remained consistently elevated across all stages, with median log2(read counts) values exceeding 15 in nearly all subgroups. While a slight decrease was observed in certain stage II subcategories (e.g., stage IIB), the overall expression profile showed no statistically significant stage-dependent downregulation. These findings support the notion that CEACAM5 overexpression is a conserved molecular hallmark across colorectal tumor progression, further reinforcing its value as a stable biomarker and potential target for stage-independent nanovaccine strategies.

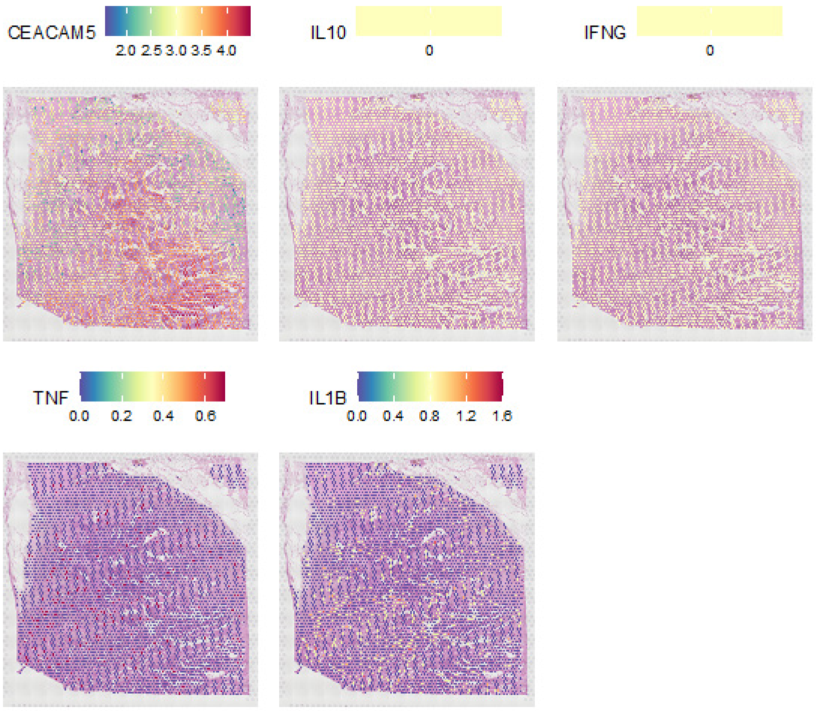

Figure 5. Spatial expression of CEACAM5 in human colorectal tumor tissue. SpatialFeaturePlot showing the log-normalized expression of CEACAM5 across tissue spots from a Visium slide (GSE226997). Expression is concentrated in epithelial tumor regions, with reduced signal in stromal or immune-dense zones. The spatial distribution highlights CEACAM5's compartmentalization in tumor tissue and supports its relevance as a tumor-localized antigen for targeted immunotherapy.

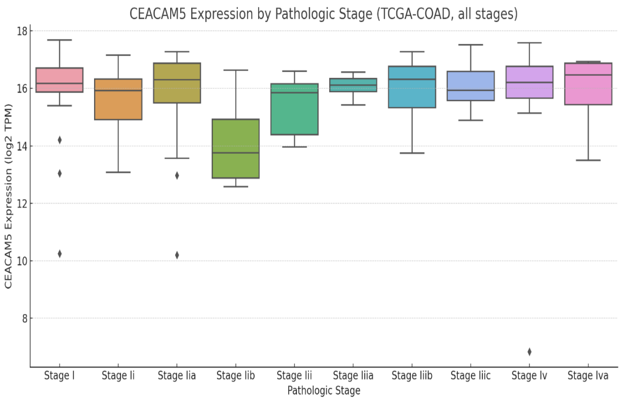

Figure 6. CEACAM5 expression across colorectal cancer stages (TCGA-COAD). Boxplot showing the distribution of $\log_2$-transformed CEACAM5 expression (read counts) across detailed AJCC pathological stages (I–IVa) in TCGA-COAD samples (n = 524). Expression remains consistently elevated in all tumor stages, with minimal variation between subgroups. This conserved expression pattern supports the role of CEACAM5 as a robust and stage-independent biomarker.

Integrated immunological interpretation

In summary, the current analysis indicates trends linking elevated CEACAM5 expression with reduced cytokine activity and immune cell infiltration. While these results are preliminary and not statistically significant, they provide a foundation for future research aimed at understanding the role of CEACAM5 in immune regulation and tumor biology in colorectal cancer.

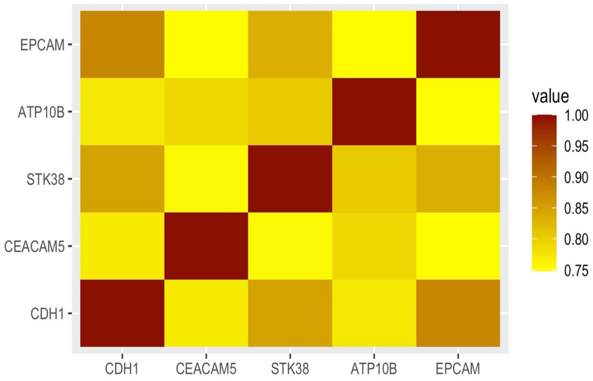

Figure 7. Spearman correlation heatmap of CEACAM5 and top co-expressed genes in TCGA-COAD. Heatmap showing pairwise Spearman correlation coefficients between CEACAM5 and four highly co-expressed genes: CDH1, STK38, ATP10B, and EPCAM. All gene pairs display strong positive correlations ($\rho > 0.75$), supporting the selection of these candidates for further immunological and biomarker analysis.

CEACAM5 co-expression and candidate discovery

We found four genes—CDH1, STK38, ATP10B, and EPCAM—with strong positive correlation to CEACAM5 (Spearman $\rho > 0.75$), supporting a shared expression profile (Figure 7).

To better visualize expression concordance, we plotted the sample-wise $\log_2$ read counts expression of each gene across all tumor samples, sorted by CEACAM5 levels. As shown in Figure 8, the expression profiles of CDH1, STK38, ATP10B, and EPCAM closely track with CEACAM5, forming near-parallel trends across the tumor cohort. This concordant behavior supports their inclusion as co-expression candidates for downstream functional and immunological characterization.

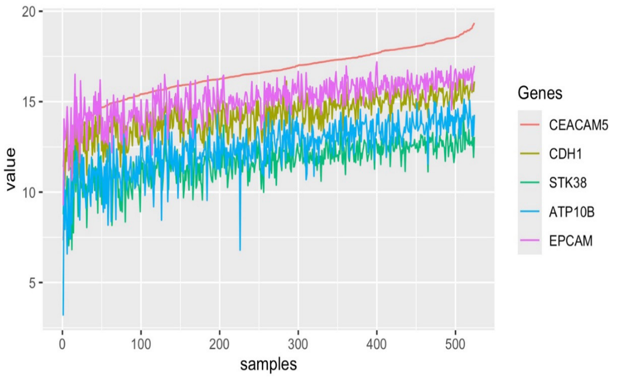

Figure 8. Sample-wise expression trends for CEACAM5 and its top four co-expressed genes in TCGA-COAD. Each line represents the normalized expression ($\log_2$ read counts) of one gene across 524 tumor samples, sorted by CEACAM5 expression (red line). The overlapping expression profiles of CDH1, STK38, ATP10B, and EPCAM highlight their strong positive co-expression with CEACAM5, supporting their selection for further immunogenomic investigation.

Expression comparisons between CEACAM5-high and CEACAM5-low quartile groups further validated these associations, with each candidate gene showing significantly elevated expression in the CEACAM5-high group (Wilcoxon rank-sum tests, $p < 0.001$ for each; Figure 9.A-D).

Literature-annotated vaccine candidate table

To better contextualize these candidate genes for potential immunotherapeutic application, we conducted a structured literature review. Each candidate was annotated according to existing evidence supporting their immunotherapeutic relevance, such as vaccine development, circulating tumor cell marker status, or neoantigen potential (Table I). Notably, EPCAM emerged as particularly promising due to its established utility in cancer vaccines, and roles as a marker for circulating tumor cells and as a candidate in neoantigen identification.

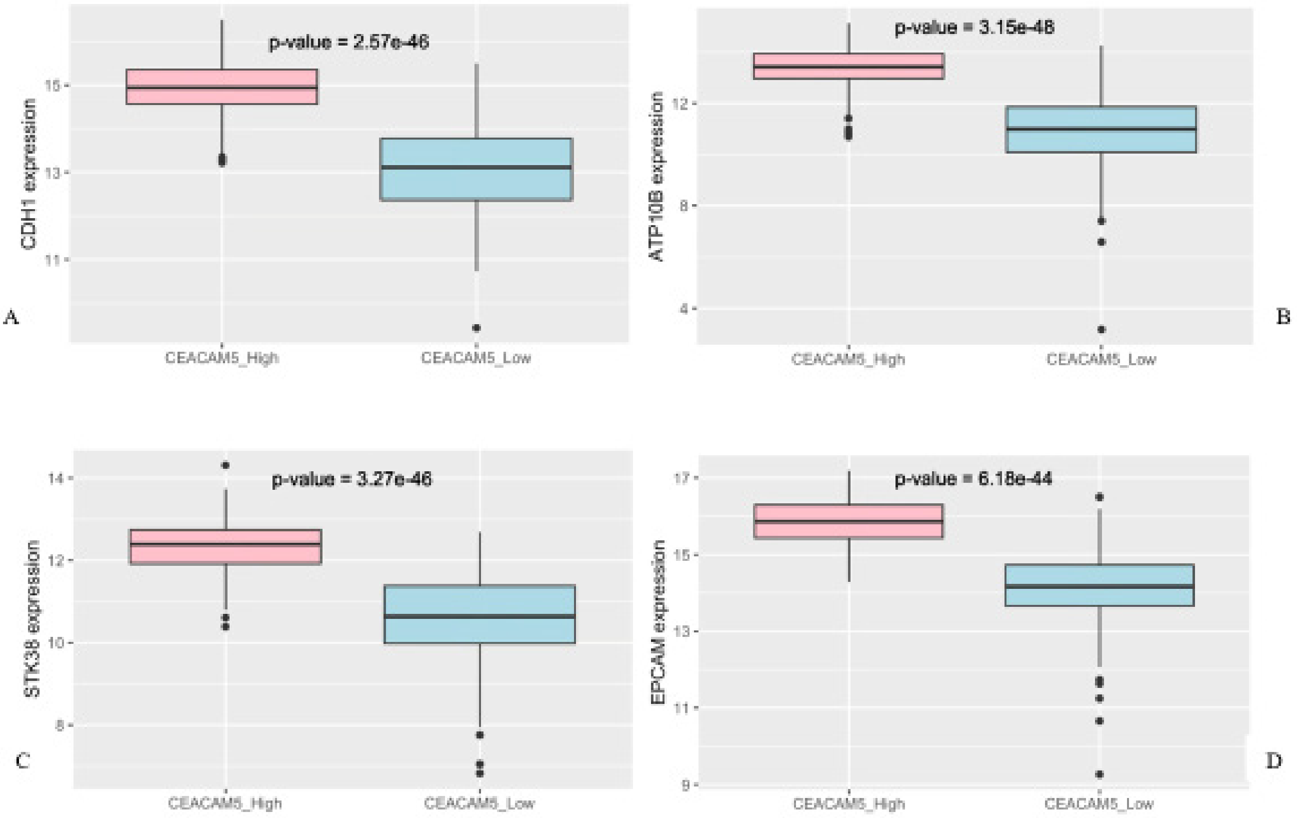

Figure 9. (A–D) Quartile comparisons of CDH1, ATP10B, STK38, and EPCAM expression in CEACAM5-high vs. CEACAM5-low TCGA-COAD tumors (n = 524). Boxplots show the values of the respective genes from the samples in the highest and lowest quartile of the CEACAM5 expression. We observe that all the candidate genes also have the higher values in the CEACAM5-high group versus the CEACAM5-low group.

Table I. Immunotherapeutic relevance of CEACAM5 co-expression candidates.

Gene	Mechanism Similarity to CEACAM5	Vaccine Candidate Potential	Current Evidence & Notes
CDH1	Moderate (cell adhesion, tumor progression)	Low	No direct vaccine development; risk of autoimmunity; iPSC vaccines induce broader T cell responses [19].
STK38	Low (intracellular kinase, signaling)	Moderate (as part of multi-antigen vaccines)	Included in broad tumor cell-based vaccines showing efficacy in mice [12].
ATP10B	None	None	Overexpressed in embryonic tissues (GSE245218); not yet validated in CRC or vaccines [20].
EPCAM	High (cell adhesion, overexpressed in CRC)	Moderate	Studied in immunotherapy; limited vaccine data; potential inclusion in multi-epitope vaccines [2].
CEACAM5	High (cell adhesion, immune evasion)	High	Validated vaccine and antibody target; peptides tested in clinical trials; monoclonal antibodies in trials [21].

Extra-validation study

To investigate the biological processes most affected during the transition from embryonic to adult tissue states, we conducted Gene Ontology (GO) enrichment analysis using the full list of differentially expressed genes derived from GSE245218. The differential gene expression anlaysis was performed using the DESeq2 package in R. As illustrated in Figure 10, the results of the differential expression analysis for the candidate genes highlights EPCAM and ATP10B as the only CEACAM5-correlated candidates showing significant overexpression in embryonic tissues (|$\log_2$FC| > 2, FDR < 0.01), reinforcing their relevance as oncofetal antigens. This developmental profiling was intended to prioritize CEACAM5 co-expression candidates with oncofetal characteristics, thereby identifying genes like ATP10B and EPCAM that exhibit embryonic overexpression and potential translational relevance as tumor-specific vaccine antigens.

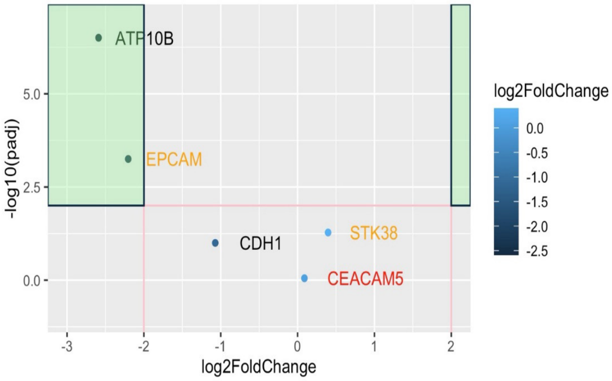

Figure 10. Volcano plot of differential gene expression between embryonic and adult murine tissues (GSE245218). Log_2 fold change is plotted against $-\log_{10}$ adjusted p-value for genes co-expressed with CEACAM5. Shaded regions denote significance thresholds (FDR < 0.01 and |$\log_2$FC| > 2). ATP10B and EpCAM are significantly upregulated in embryonic tissues, consistent with their role as oncofetal markers.

This functional analysis, performed with the clusterProfiler package in R, identified pathways enriched in the differentially expressed genes from the comparison of embryo versus adult mice mentioned in the previous paragraph. Results, shown in Figure 11, identified the antibacterial humoral response pathway as significantly downregulated (adjusted p < 0.05). The humoral immune response, mediated by B lymphocyte–derived antibodies, is critical for neutralizing extracellular pathogens and orchestrating adaptive immunity. Although EPCAM—one of our CEACAM5 co-expression candidates—is not directly involved in antibody production, it can indirectly influence immune activity by modulating antigen-presenting cell function and epithelial barrier integrity. Notably, EPCAM is frequently overexpressed in epithelial tumors and has been implicated in promoting tumor growth and metastasis [22]. This immunological downregulation supports a broader mechanism of immune evasion associated with CEACAM5 and its co-expressed antigens.

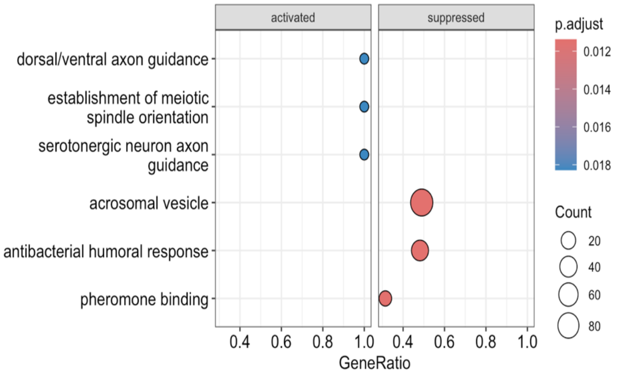

Figure 11. Gene ontology analysis of differentially expressed genes between embryonic and adult murine tissues (GSE245218). Panel shows top three significantly enriched biological processes upregulated ('activated') or downregulated ('suppressed') in embryonic tissues. The antibacterial humoral response, shown in red, is notably downregulated, consistent with immune-suppressive phenotypes seen in CEACAM5-high tumors.

Discussion

Our findings provide a multidimensional perspective on CEACAM5 as both a biomarker and an active modulator of the immune landscape in colorectal cancer. Although transcriptomic data from the TCGA-COAD cohort revealed a pattern of CEACAM5 overexpression in tumors and inverse correlations with several cytokines central to antitumor immunity (e.g., IFNG, IL10, TNF, IL1B), these associations did not reach statistical significance. As such, they should be interpreted as exploratory, suggesting that CEACAM5-high tumors may exhibit an immune-excluded or transcriptionally suppressed phenotype—hypotheses that warrant further validation in larger, stratified cohorts. [1,23,24]. In future work, we plan to incorporate immune cell–stratified or microsatellite instability (MSI)–annotated colorectal cancer datasets available through platforms such as cBioPortal or the GEO datasets GSE39582 and GSE14333, which include immune signatures and clinical subtypes. These resources will help refine and validate the observed trends.

xCell-based immune deconvolution supported this immune-silent signature, highlighting negative correlation trends between CEACAM5 expression and the presence of $CD8^+$ effector memory T cells, NK cells, monocytes, and Th2 cells. The depletion of cytotoxic and regulatory populations suggests potential impairment of both innate and adaptive immune surveillance. These findings align with prior evidence that CEACAM family members can regulate immune checkpoints and suppress inflammation through contact-dependent mechanisms [24-26].

Spatial transcriptomics provided additional insights into CEACAM5 localization and immune exclusion. In human colorectal tumor tissue (GSE226997), CEACAM5 expression was confined to epithelial compartments, while inflammatory cytokines were spatially restricted to stromal zones. This decoupling suggests that CEACAM5-positive regions may be insulated from immune activation. Importantly, CEACAM5 expression was consistently elevated across all pathologic stages, reinforcing its potential as a universal nanovaccine target. Similarly, in murine spleen and liver data (GSE207843), Ceacam5 expression showed negligible spatial correlation with IFN-G, IL10, IL1B, or TNF, reinforcing the idea that CEACAM5's immunomodulatory capacity may depend on additional signals such as inflammation, antigen presentation, or nanoparticle delivery.

These observations also correlate with in vivo studies using CEA-functionalized gold nanoparticles (CEA-AuNPs), which induced shifts in cytokine levels in murine models [19]. The concordance between our TCGA-based and spatial data and these experimental findings underscores CEACAM5's potential as both a biomarker of immune evasion and a target for nanovaccine-based immunotherapies.

To extend beyond CEACAM5, we employed a co-expression pipeline to identify immunologically relevant genes with similar expression patterns. Among these, EPCAM showed strong correlation with CEACAM5 and has known applications in CAR-T and cancer vaccine development [27,28]. STK38, although not a surface protein, has been incorporated into multi-antigen vaccine formulations [12,29]. In contrast, CDH1, despite moderate correlation, was excluded due to its tumor-suppressor function, frequent inactivation in CRC, and associated autoimmunity risk [30, 31]. Although ATP10B has not yet been explored as a vaccine antigen in CRC, its robust re-expression in embryonic tissues suggests an oncofetal profile worth further exploration [32].

The developmental expression profiling was conducted exclusively in murine embryonic and adult tissues. While it provides a rationale for considering ATP10B and EPCAM as potential oncofetal markers, direct extrapolation to human fetal or tumor contexts is limited. Future validation will be pursued using human fetal and adult tissue datasets—such as those available via the Human Developmental Biology Resource (HDBR), the EMBL-EBI Expression Atlas, or newer spatial fetal atlases (e.g., GSE171156). These efforts will enable translational refinement of oncofetal candidate selection beyond the current mouse embryo dataset (GSE245218).

Complementary gene ontology analysis further revealed downregulation of antibacterial humoral immune responses in CEACAM5-high tumors. While CEACAM5 itself is not directly involved in antibody signaling, this pathway-level suppression suggests reduced B-cell engagement and humoral immune evasion. This immunological footprint aligns with prior reports linking CEACAM1 and CEACAM5 to systemic immune dampening in epithelial cancers [1,4,26].

This study serves as a proof of concept for an automated bioinformatics pipeline to support vaccine candidate discovery and tumor antigen profiling. While the current analysis demonstrated feasibility across multiple datasets, there remain clear limitations and opportunities for refinement. For instance, improved normalization strategies could be applied to raw count data, and future differential expression analysis would benefit from incorporating human-derived embryonic datasets rather than murine proxies. Additionally, more robust functional annotation could be achieved by integrating pathway-level analysis or decomposing Gene Ontology enrichment results into its three primary subcategories: biological process, molecular function, and cellular component. These considerations highlight the potential for future customization and expansion of this pipeline for tailored immunogenomic studies.

Taken together, these results support the continued evaluation of CEACAM5 as a therapeutic target and stratification biomarker. Its integration into vaccine platforms may benefit from co-targeting epithelial co-expressors like EPCAM and the addition of immune-stimulatory adjuvants capable of overcoming cytokine suppression and immune cell exclusion. Further validation in functional assays and tumor-bearing models will be critical for advancing CEACAM5-guided nanovaccine strategies.

To support translational application of these findings, future work should explore the functional immunogenicity of CEACAM5 and its co-expression candidates *in vitro* (e.g., antigen presentation assays, dendritic cell activation) and *in vivo* (e.g., syngeneic tumor models or humanized mice). Such validations are essential for advancing from computational prediction to therapeutic development.

Conclusions

This study provides an integrative immunogenomic characterization of CEACAM5 in colorectal cancer, highlighting its association with an immune-excluded tumor phenotype. Our multi-cohort transcriptomic analyses revealed negative correlations between CEACAM5 expression and key cytokines such as IFNG, IL10, and IL1B, alongside reduced immune cell infiltration scores in CEACAM5-high tumors. Spatial transcriptomics confirmed the compartmentalization of CEACAM5 expression to epithelial tumor regions, distinct from immune-enriched zones, further supporting its role in immune evasion.

Through co-expression analysis, we identified EPCAM and ATP10B as promising tumor-selective targets, with developmental regulation consistent with an oncofetal expression profile. Functional enrichment analyses suggested suppression of humoral immune pathways in CEACAM5-high contexts, adding another layer to its immunosuppressive potential.

Together, these findings position CEACAM5 as not only a diagnostic and prognostic marker but also as a rational immunotherapy target. The study establishes a proof-of-concept pipeline that can be expanded to prioritize antigens for nanovaccine development or multi-target immunotherapies based on tumor–immune spatial and transcriptional interplay.

Acknowledgments

The authors used artificial intelligence (AI)-based tools, including ChatGPT (OpenAI) and Perplexity AI, to assist with language editing, phrasing suggestions, and improving manuscript clarity. No AI tool was used for data analysis, scientific interpretation, or decision-making. All intellectual content, analysis, and conclusions are the original work of the authors, who assume full responsibility for the integrity and accuracy of the manuscript.

References

1. Beauchemin N, Arabzadeh A. Carcinoembryonic antigen-related cell adhesion molecules (CEACAMs) in cancer progression and metastasis. Cancer Metastasis Rev.;32:643-671.
2. Keller L, Werner S, Pantel K. Biology and clinical relevance of EpCAM. Cell Stress. 2019;3:165-180.
3. Fiori V, Magnani M, Cianfriglia M. The expression and modulation of CEACAM1 and tumor cell transformation. Ann Ist Super Sanita. 2012;48:161-171.
4. Thomas J, Klebanov A, John S, Miller LS, Vegesna A, Amdur RL, et al. CEACAMS 1, 5, and 6 in disease and cancer: interactions with pathogens. Genes Cancer. 2023;14:12-29.
5. Koveitypour Z, Panahi F, Vakilian M, Peymani M, Seyed Forootan F, Nasr Esfahani MH, et al. Signaling pathways involved in colorectal cancer progression. Cell Biosci. 2019;9:97.
6. Decalf J, Albert ML, Ziai J. New tools for pathology: a user's review of a highly multiplexed method for in situ analysis of protein and RNA expression in tissue. J Pathol. 2019;247:650-661.
7. Maynard KR, Jaffe AE, Martinowich K. Spatial transcriptomics: putting genome-wide expression on the map. Neuropsychopharmacology. 2020;45:232-233.
8. Hoffmann PR, Hoffmann FW, Premeaux TA, Fujita T, Soprana E, Panigada M, et al. Multi-antigen Vaccination With Simultaneous Engagement of the OX40 Receptor Delays Malignant Mesothelioma Growth and Increases Survival in Animal Models. Front Oncol. 2019;9:720.
9. Zdrehus RS, Mocan T, Sabau LI, Matea CT, Tabaran A-F, Pop T, et al. CEA-Functionalized Gold Nanoparticles for Oral Prophylaxis: An In Vivo Evaluation of Safety, Biodistribution, and Cytokine Expression in Healthy Mice. Journal of Nanotheranostics. 2025; 6:18.
10. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18:220.
11. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33:495-502.
12. Liu Y, Shi Z, Zheng Z, Li J, Yang K, Xu C, et al. Prognostic and Immunological Role of STK38 across Cancers: Friend or Foe? Int J Mol Sci. 2022;23:11590.
13. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15-21.
14. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207-210.
15. Gene Expression Omnibus (GEO). Series accession: GSE245218. NCBI.
16. Huijbers EJM, van Beijnum JR, van Loon K, Griffioen CJ, Volckmann R, Bassez A, et al. Embryonic reprogramming of the tumor vasculature reveals targets for cancer therapy. Proc Natl Acad Sci U S A. 2025;122:e2424730122.
17. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
18. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284-287.
19. Jwo SH, Ng SK, Li CT, Chen SP, Chen LY, Liu PJ, et al. Dual prophylactic and therapeutic potential of iPSC-based vaccines and neoantigen discovery in colorectal cancer. Theranostics. 2025;15:5890-5908.
20. Goto S, Cao F, Kono T, Ogawa H. Microarray analysis of differentially expressed genes in inner cell mass and trophectoderm of parthenogenetic embryos. Journal of Mammalian Ova Research. 2016;33:45-54.
21. Bhagat A, Lyerly HK, Morse MA, Hartman ZC. CEA vaccines. Hum Vaccin Immunother. 2023;19:2291857.
22. Huang L, Yang Y, Yang F, Liu S, Zhu Z, Lei Z, et al. Functions of EpCAM in physiological processes and diseases (Review). Int J Mol Med. 2018;42:1771-1785.
23. Wu X, Huang Q, Chen X, Zhang B, Liang J, Zhang B. B cells and tertiary lymphoid structures in tumors: immunity cycle, clinical impact, and therapeutic applications. Theranostics. 2025;15:605-631.
24. Saiz-Gonzalo G, Hanrahan N, Rossini V, Singh R, Ahern M, Kelleher M, et al. Regulation of CEACAM Family Members by IBD-Associated Triggers in Intestinal Epithelial Cells, Their Correlation to Inflammation and Relevance to IBD Pathogenesis. Front Immunol. 2021;12:655960.
25. Lee HS-W. Carcinoembryonic antigen-related cellular

adhesion molecule 1-dependent inhibition of T cell responses. PhD Thesis, University of Toronto, 2009.

26. Kelleher M, Singh R, O'Driscoll CM, Melgar S. Carcinoembryonic antigen (CEACAM) family members and Inflammatory Bowel Disease. Cytokine Growth Factor Rev. 2019;47:21-31.

27. Patriarca C, Macchi RM, Marschner AK, Mellstedt H. Epithelial cell adhesion molecule expression (CD326) in cancer: a short review. Cancer Treat Rev. 2012;38:68-75.

28. Münz M, Kieu C, Mack B, Schmitt B, Zeidler R, Gires O. The carcinoma-associated antigen EpCAM upregulates c-myc and induces cell proliferation. Oncogene. 2004;23:5748-5758.

29. Sotirov S, Dimitrov I. Tumor-Derived Antigenic Peptides as Potential Cancer Vaccines. Int J Mol Sci. 2024;25:4934.

30. Petrova YI, Schecterson L, Gumbiner BM. Roles for E-cadherin cell surface regulation in cancer. Mol Biol Cell. 2016;27:3233-3244.

31. Guilford P, Hopkins J, Harraway J, McLeod M, McLeod N, Harawira P, et al. E-cadherin germline mutations in familial gastric cancer. Nature. 1998;392:402-405.

32. Real R, Moore A, Blauwendraat C, Morris HR, Bandres-Ciga S; International Parkinson's Disease Genomics Consortium (IPDGC). ATP10B and the risk for Parkinson's disease. Acta Neuropathol. 2020;140:401-402.