Transcriptome and microbiome-immune changes across preinvasive and invasive anal cancer lesions

Ezequiel Lacunza, …, Juan C. Ramos, Martin C. Abba

JCI Insight. 2024. https://doi.org/10.1172/jci.insight.180907.

Anal squamous cell carcinoma (ASCC) is a rare gastrointestinal malignancy linked to high-risk Human papillomavirus (HPV) infection, which develops from precursor lesions like Low-Grade Squamous Intraepithelial Lesions (LGSIL) and High-Grade Squamous Intraepithelial Lesions (HGSIL). ASCC incidence varies across populations, posing increased risk for People Living with HIV (PLWH). Our investigation focused on transcriptomic and metatranscriptomic changes from Squamous Intraepithelial Lesions (SILs) to ASCC. Metatranscriptomic analysis highlighted specific bacterial species (e.g., Fusobacterium nucleatum, Bacteroides fragilis) more prevalent in ASCC than precancerous lesions. These species correlated with gene encoding enzymes (Acca, glyQ, eno, pgk, por) and oncoproteins (FadA, dnaK), presenting potential diagnostic or treatment markers. Unsupervised transcriptome analysis identified distinct sample clusters reflecting histological diagnosis, immune infiltrate, HIV/HPV status, and pathway activities, recapitulating anal cancer progression's natural history. Our study unveiled molecular mechanisms in anal cancer progression, aiding in stratifying HGSIL cases based on low- or high-risk progression to malignancy.

Find the latest version:

https://jci.me/180907/pdf
Transcriptome and microbiome-immune changes across preinvasive and invasive anal cancer lesions.

Ezequiel Lacunza 1,8,* Valeria Fink 2,8 María E. Salas 1,8 Ana M. Guñ 2,8 Jorge A. Basiletti 3 María A. Picconi 3 Mariano Golubicki 4 Juan Robbio 4 Mirta Kujaruk 4 Soledad Iseas 5 Sion Williams 6,8 María I. Figueroa 2,8 Omar Coso 7,8 Pedro Cahn 2,8 Juan C. Ramos 6,8 Martín C. Abba 1,8.*

(1) Centro de Investigaciones Inmunológicas Básicas y Aplicadas (CINIBA), Facultad de Ciencias Médicas, Universidad Nacional de La Plata, La Plata, Argentina.
(2) Dirección de Investigaciones, Fundación Huésped, Buenos Aires, Argentina.
(3) Laboratorio Nacional y Regional de Referencia de Virus Papiloma Humano. Instituto Nacional de Enfermedades Infecciosas - ANLIS "Dr. Malbrán", Buenos Aires, Argentina.
(4) Unidad de Oncología, Hospital de Gastroenterología “Dr. Carlos Bonorino Udaondo”, Buenos Aires, Argentina.
(5) Medical Oncology Department, Paris-St Joseph Hospital, Paris, France
(6) University of Miami - Center for AIDS Research (UM-CFAR) / Sylvester Comprehensive Cancer Center (SCCC), University of Miami Miller School of Medicine, Miami, FL, USA.
(7) Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE-CONICET), Universidad de Buenos Aires, Buenos Aires, Argentina.
(8) University of Miami - Centre for AIDS Research/Sylvester Cancer Comprehensive Center Argentina Consortium for Research and Training in Virally Induced AIDS-Malignancies, University of Miami Miller School of Medicine, Miami, Florida, USA.

*Corresponding authors: Correspondence to E Lacunza (ez.lacunza@gmail.com) or MC Abba (mcabba@gmail.com); phone number +54 221 423 6711 (342); mailing address: Facultad de Ciencias Médicas Universidad Nacional de La Plata Calle 60 y Av. 120 B1900 La Plata, Buenos Aires, Argentina.

Abstract

Anal squamous cell carcinoma (ASCC) is a rare gastrointestinal malignancy linked to high-risk Human papillomavirus (HPV) infection, which develops from precursor lesions like Low-Grade Squamous Intraepithelial Lesions (LGSIL) and High-Grade Squamous Intraepithelial Lesions (HGSIL). ASCC incidence varies across populations, posing increased risk for People Living with HIV (PLWH). Our investigation focused on transcriptomic and metatranscriptomic changes from Squamous Intraepithelial Lesions (SILs) to ASCC. Metatranscriptomic analysis highlighted specific bacterial species (e.g., *Fusobacterium nucleatum*, *Bacteroides fragilis*) more prevalent in ASCC than precancerous lesions. These species correlated with gene encoding enzymes (Acca, glyQ, eno, pgk, por) and oncoproteins (FadA, dnAK), presenting potential diagnostic or treatment markers. Unsupervised transcriptome analysis identified distinct sample clusters reflecting histological diagnosis, immune infiltrate, HIV/HPV status, and pathway activities, recapitulating anal cancer progression’s natural history. Our study unveiled molecular mechanisms in anal cancer progression, aiding in stratifying HGSIL cases based on low- or high-risk progression to malignancy.

Keywords: ASCC, Microbiome, Transcriptome, Metatranscriptome.
Introduction

Anal Squamous Cell Carcinoma (ASCC) is a rare gastrointestinal neoplasia that involves the formation of malignant tumors in the anal region. Over the past thirty years, the incidence of ASCC has been on the rise globally, particularly in men who have sex with men (MSM) and people living with HIV (PLWH) (1).

Squamous Intraepithelial Lesions (SILs), categorized into Low-Grade (LGSIL), analogous to anal intraepithelial neoplasia I, and High-Grade (HGSIL), analogous to anal intraepithelial neoplasia II and III, often precede the progression to ASCC (2, 3) Similar to cervical cancer, ASCC development is driven by the infection with oncogenic human papillomaviruses (HPV) (4, 5). The risk of anal cancer varies significantly across different population groups, with the highest risk observed in PLWH (1). This increased susceptibility is primarily attributed to a weakened immune system, which makes it more challenging to control infections, including HPV infections (6). Beyond the potential impact of oncogenic viruses, the microbiome may also play a significant role in the development of precancerous anal lesions and ASCC, as the influence of microbes is increasingly recognized in cancer development (7,8). The microbiome can influence the balance of host cell proliferation and apoptosis, disrupt anti-tumoral immunity, and affect the metabolism of host-produced factors, ingested food components, and drugs. (9). In a recent study, we defined the microbiome composition of the anal mucosa of HIV-exposed individuals. Metagenomic sequencing enabled us to identify viral and bacterial taxa linked to the development of anal lesions. Our results confirmed the occurrence of oncogenic viromes in this population and identified *Prevotella bivia* and *Fusobacterium gonidiaformans* as two relevant bacterial species predisposing to SILs. Moreover, gene family analysis identified bacterial gene signatures associated with SILs that may have potential as prognostic and predictive biomarkers for HIV-associated malignancies (10). Other reports using 16S rRNA gene sequencing to analyze the ASCC demonstrated the role of the anal microbiota in anal cancer response to therapy and toxicity, as well as changes in taxonomic compositions among normal, dysplasia, and anal cancer samples (11,12).

The molecular biology of ASCC is complex and not completely understood (13). However, several studies have identified potential molecular targets for ASCC therapy, including regulators of apoptosis (14), agents targeting the PI3K/AKT pathway (15), antibody therapy targeting EGFR (16) or PD-L1 expression to stratify good versus poor responders to chemoradiotherapy (17). Despite advancements in understanding ASCC from various perspectives, thus far, no prognostic or predictive markers have been identified that are useful in clinical practice. Furthermore, a notable gap in existing information is the paucity of
studies employing anal cancer biopsies for gene expression profiling, particularly utilizing advanced techniques like next-generation sequencing (NGS).

Transcriptomics and metatranscriptomics profiling are powerful NGS-based tools for the functional genomics characterization of complex diseases. In this sense, bulk RNA sequencing (RNA-seq) in neoplastic disease enables the simultaneous study of the host tumor transcriptome and its microenvironment, including the tumor immune infiltrate and the associated tumor microbiome. Transcriptomic profiling provides a thorough examination of gene expression patterns, uncovering crucial insights into the molecular mechanisms driving cancer development and progression. Metatranscriptomic profiling enables researchers to analyze gene expression levels of various organisms within a microbial community, providing insights into their metabolic processes and functional activities in cancer and immune-related diseases (18). In this sense, metatranscriptomics approaches enable the analysis of the active microbiota instead of more frequent studies based on 16S rRNA sequencing, which analyzes the “total” microbiota, including active and inactive bacteria.

The aim of this study was to analyze the transcriptomic and metatranscriptomic changes that occur during the progression from LGSIL to HGSIL and ultimately to ASCC.

We collected biopsies identified as SILs and ASCC from a cohort of 70 participants, encompassing individuals both with and without HIV, all of whom provided informed consent. Biopsies were subjected to bulk RNA-seq. Our goal was to gain insights into the molecular mechanisms underlying the development and progression of anal lesions, which could potentially lead to the identification of novel biomarkers and therapeutic targets for improved diagnostic and treatment strategies in patients with ASCC.

Results

Clinical characteristics of patients and microbial community variations in SILs and ASCC cases.

Seventy patients were included in the present study. All participants underwent anal cytology and high-resolution anoscopy with biopsies. Based on cytology and histology analysis, samples were classified into LGSIL, n=23, HGSIL, n=16, and ASCC, n=23. Demographic and clinical data were collected, including age, sex at birth (male or female), gender (cisgender men (CGM), transgender women (TGW), and cisgender women (CGW)), HPV DNA status, HIV status, and antiretroviral therapy (ART). This information is summarized in Table 1.
We first conducted a compositional analysis of the three distinct groups—LGSIL, HGSIL, and ASCC—by performing permutational multivariate ANOVA (PERMANOVA) with Euclidean distance. The Principal Coordinate Analysis (PCoA) defined two distinct clusters based on component I (p<0.001). Cluster I was enriched in LGSILs, comprising 24 out of 31 samples (77%), while Cluster II predominantly featured ASCC samples with 19 out of 23 (83%) (Fig. 1A). HGSIL demonstrated an almost equal distribution between the two clusters, with 9 out of 16 in Cluster I (56%) and 7 out of 16 in Cluster II (44%) (Fig. 1A). In addition, we considered covariates such as age, gender, HIV status, and high-risk HPV DNA genotyping (HR-HPV) to evaluate the factors influencing cluster formation based on diagnostic groups. Employing PERMANOVA, our analysis of beta diversity revealed distinctions primarily in samples positive for HR-HPV types compared to samples in which these HPV types were undetected (Figure 1B; Supp Figure 1).

The ASCC microbial community, assessed through Observed and Chao 1 indices based on metatranscriptome species composition, exhibited a significantly higher richness compared to LGSIL (Observed, p = 0.033; Chao 1, p = 0.035) and HGSIL (Observed, p = 0.029; Chao 1, p = 0.034). This trend persisted when merging LGSIL and HGSIL into the group termed SILs (Observed, p = 0.012; Chao 1, p = 0.018), suggesting that the ASCC environment may provide a more favorable habitat for a specific range of microorganisms, resulting in increased community richness (Fig. 1B; Supplementary Data 1). Richness indices were also augmented in the HR-HPV group compared to the negative cohort for HR-HPV types. In addition, a significant association between HIV-positive status and decreased alpha diversity was observed, in agreement with previous studies (10) (Fig. 1B).

Analysis of diversity indices (Shannon and Simpson) revealed a significant increase in ASCC compared to HGSIL (Shannon, p = 0.0082; Simpson, p = 0.0134), while no differences were observed between LGSIL and ASCC (Supplementary Data 1). These findings align with a recent study that reported similar alpha diversity indices between anal dysplasia and anal cancer but highlighted an elevated abundance of specific taxa in the latter (12). Consistent with our prior research we further observed a negative influence of aging on microbiome diversity (10) (Supplementary Data 1).

We analyzed bacterial abundance at the phylum level between SILs and ASCC groups. Fusobacteriota, Bacteroidota, and Bacillota, among the most abundant phyla, were significantly more enriched in ASCC compared to SILs (Fig. 1D). Additionally, Pseudomonadota showed enrichment within the ASCC group compared to precancerous lesions (Fig. 1D).
At the species level, we identified a total of 25 taxa, each exhibiting a relative abundance exceeding 20% of the overall composition in at least one of the samples (Fig. 1E). Among these taxa, *Fusobacterium nucleatum, Fusobacterium necrophorum, Bacteroides fragilis, and Prevotella intermedia* are well-established gut-associated bacteria with previous associations with colorectal cancer (CRC) (19). Conversely, other taxa such as *Mycoplasma hominis, Prevotella bivia, Fusobacterium gonidiaformans, Sneathia amnii, Campylobacter ureolyticus* or *Bacteroides fragilis* have been linked to HPV-related precancerous and cancerous genital lesions (10, 12, 20, 21, 22).

To identify bacterial species associated with ASCC compared to SILs, we used MaAsLin2 analysis. To account for potential confounders, we refined the model by incorporating additional covariates, including HIV status, HR-HPV DNA status, sex at birth, and age. Significant enrichment was observed for *Fusobacterium nucleatum* (p = 0.001), *Fusobacterium gonidiaformans* (p = 0.001), *Bacteroides fragilis* (p = 0.01), *Campylobacter ureolyticus* (p = 0.003), and *Criibacterium bergeronii* (p = 0.006) (Fig. 1E; Supplementary Data 2). Moreover, *C. ureolyticus* (p=0.002), *F. gonidiaformans* (p=0.01), and *C. bergeronii* (p=0.02) were associated with male sex (Supplementary Data 2). Additionally, *C. ureolyticus* correlated with HIV-negative cases (p=0.03) (Supplementary Data 2).

*F. nucleatum* and *B. fragilis* have established roles in CRC progression, highlighting their importance in ASCC development and progression (19). While knowledge about *F. gonidiaformans, C. ureolyticus,* and *C. bergeronii* is limited, prior associations exist between *F. gonidiaformans* and *C. ureolyticus* with HPV presence and the development of precancerous lesions in anal and cervical cancers (10, 21, 23). These findings suggest a potential contribution of specific bacteria to ASCC progression.

Exploring Viral Signatures in Anal Lesions Progression: Alpha Papillomavirus and Non-HPV Species.

In terms of viral composition analysis, among the 40 species identified at the transcript level in all samples, eight were the most prevalent, with abundances greater than 30% of the total abundance in any sample and detected more than three times. Notably, seven of these species belonged to the *Alpha Papillomavirus* (Alpha-PV) genus, along with the *Human endogenous retrovirus K* (HERV-K), with evident variations in their relative abundances across distinct diagnostic groups (Fig. 2A). MaAsLin2 analysis revealed a higher abundance of Alpha-PV-10, which includes low-risk genotypes like HPV6 and HPV11, in both LGSIL and HGSIL compared to ASCC. (Fig. 2A, B, Supplementary Data 2). Conversely, Alpha-PV-9 (HPV16, 31, 33, 52, 58) and Alpha-PV-7 (HPV18, 39, 59, 68, 45, 70) were significantly associated with HGSIL and ASCC.
This trend persisted when considering the number of positive cases for these species independent of their relative abundance (Fig. 2C). Although the significance was not established for Alpha-PV-10, it remained significant for Alpha-PV-7 and Alpha-PV-9 (Fig. 2C).

The HPV DNA genotyping data highlighted a robust association between HPV16 and both HGSIL and ASCC, correlating with the pattern observed with Alpha-PV-9 (Fig. 2C). However, HPV18 was detected in only one case of ASCC, contrasting with Alpha-PV-7 detected at the RNA level in over 20% of participants (Fig. 2A). This discrepancy could be due to Alpha-PV-7 containing other HPV genotypes (24). HPV6 and HPV11 were predominantly linked to LGSIL (Fig. 2D). Analyzing positive and negative cases for all low-risk (LR) HPV types and high-risk (HR) HPV types identified within the cohort revealed negative (p<0.001) and positive associations (p<0.05), respectively, with the diagnostic groups (Fig. 2E, F). These results confirm the prominence of HR and LR HPV types, particularly HPV6 and HPV16, in delineating the diagnostic groups (25).

Among the non-HPV species, it is noteworthy to highlight a significant increase in the relative abundance of the endogenous HERV-K in ASCC compared with HGSIL (p<0.01; Supplementary Data 2). HERV-K overexpression is widely associated with malignant phenotypes and is upregulated in various cancers such as breast lymphoma, germ-line tumors, and melanoma (26). Additionally, Human betaherpesvirus 5 (HCMV), although with low relative abundance, demonstrated significant enrichment in ASCC compared with SILs (p<0.05; Supplementary Data 2). HCMV is linked to several cancer types, including lymphoma, cervical cancer, Kaposi’s sarcoma, CRC, prostate cancer, skin cancer, and glioblastomas (27). However, it remains unclear whether HCMV actively contributes to malignant tumor progression or is reactivated under conditions leading to chronic inflammation or immunosuppression (27).

Overall, these findings confirm the significance of specific viral Alpha Papillomavirus species and their association with SILs toward ASCC progression. Furthermore, our data reveals a potential involvement of HERV-K and HCMV in ASCC tumorigenesis. Additionally, the use of metatranscriptomics demonstrates remarkable reliability, sensitivity, and specificity in detecting the presence of HPV types, even in cases where DNA genotyping results were negative.

**Metabolic Pathways in ASCC Progression**

To understand the functional implications of microbial community changes between SILs and ASCC, we conducted metatranscriptomics analysis, revealing 20 MetaCyc modules as significantly enriched pathways in ASCC compared to SILs (Table 2). These modules encompassed Nucleotide, Amino Acid, and Lipid
Biosynthesis pathways. This finding aligns with our prior observations, where pathways related to amino acid and de novo nucleotide biosynthesis were enriched in HIV individuals with anal precancerous lesions (10). These pathways are vital for cell growth and proliferation, as cells require energy and nutrients from their environment to support these processes. Similarly, cancer cells exhibit metabolic adaptations essential for their growth (28). Hence, our data suggest that certain bacteria within the evolving microenvironment during malignancy may exploit these pathways to thrive and proliferate, like cancer cells.

Microbial Contributions to Anal Lesions: Enriched Proteins and Taxonomic Associations.

To go further, we next explored the gene proteins contributed by the microbial organisms in the comparison of SILs versus ASCC. MaAsLin2 analysis yielded a total of 2523 UniRef90 sequence proteins differentially expressed (Supplementary Data 3). We further employed the KEGG database to annotate 387 proteins of which 349 were significantly enriched in ASCC and 37 in SILs (Supplementary Data 3). Functional annotation using KEGG Mapper revealed metabolic pathways such as glycolysis, lipid, amino acid, and nucleotide biosynthesis, contributed by 60 bacterial proteins enriched in ASCC (Fig. 3A). Proteins like Acca (acetyl-CoA carboxylase carboxyl transferase subunit alpha), glyA (glycine hydroxymethyltransferase), glyQ (glycyl-tRNA synthetase alpha chain), eno (enolase), pgk (phosphoglycerate kinase) and por (pyruvate-ferredoxin/flavodoxin oxidoreductase), previously identified in anal samples from individuals with precancerous anal lesions (10), underline their potential roles as metabolic markers in anal cancer progression. In addition, among these 60 proteins, we identified the enrichment of the oncogenic FadA adhesion protein from F. nucleatum in ASCC, a factor widely associated with CRC; and dnaK, a protein kinase with a known involvement in carcinogenesis and cancer progression (29,30). These findings align with the taxonomic abundance analysis, highlighting the significant role of bacteria like B. fragilis, F. nucleatum, and C. ureolyticus, alongside other relevant and distinct gut microbiota taxa, in orchestrating these processes (Fig. 3A). Furthermore, four proteins linked to the oncogene E6 from Human Papillomavirus 16 were enriched in ASCC (Fig. 3B). E6 oncoprotein promotes p53 degradation, contributing to keratinocyte immortalization. In SILs, 37 enriched proteins were detected, all predominantly associated with genes from the LR Human Papillomavirus genomes HPV6 and HPV11, underscoring their potential role as drivers or sustainers of precancerous anal lesions (31) (Fig. 3B).
Transcriptomic profiling and functional insights across anal lesion progression.

We then explored the host transcriptome of LGSIL, HGSIL, and ASCC. Like the metatranscriptomes analysis, the unsupervised clustering of samples revealed two primary clusters (Fig. 4A). Cluster I is predominantly composed of LGSILs, with the inclusion of some HGSILs. In contrast, Cluster II comprises most anal cancer samples, alongside a subgroup of SILs. One plausible interpretation for this distribution is that precancerous lesions may be at varying stages of progression, with some nearing malignant transformation and others in a regressive or early stage (32).

Next, we applied supervised comparative analysis between LGSIL and HGSIL as well as HGSIL and ASCC. The analysis revealed a higher number of differentially expressed genes (DEG; FC>2, FDR < 0.05) in the transition from HGSIL to ASCC (544 DEG) than in the comparison among the two SIL groups (121 DEG) (Fig. 4B, C; Supplementary Data 4). Among the most significant genes, a decrease in keratins in HGSIL compared to LGSIL stands out (Fig. 4B) as well as the overexpression of members of the MAGE gene family of cancer/testis antigens in ASCC compared with HGSIL, like MAGEA4, MAGEA3, and MAGEA1 (Fig. 4C). The MAGE family has gained attention as a potential cancer biomarker and immunotherapy (33). Notably, a phase I trial for autologous T-cell therapy targeting MAGEA4-positive solid cancers is currently underway (34).

To comprehend the functional significance of DEG, we employed Gene Set Enrichment Analysis (GSEA) on Gene Ontology (GO), Cancer Hallmarks, and Disease Ontology (DO) terms. GSEA revealed activated processes such as nuclear division, chromatin modification, and cell proliferation, along with suppressed pathways like keratinocyte differentiation and leukocyte-mediated immunity in HGSIL compared to LGSIL (Fig. 4D; Supplementary Data 5). These processes align with the histopathological features of HGSIL, including a higher nuclear-to-cytoplasmic ratio, decreased organization of cell layers, a greater degree of nuclear pleomorphism, and increased mitotic index (35). Furthermore, analysis of Cancer Hallmarks indicated the activation of pathway terms associated with sustaining proliferative signaling, such as MYC targets, E2F targets, G2M checkpoint, or mitotic spindle (Fig. 4E). Notably, there was a decrease in genes related to IFN-alpha and IFN-gamma levels, potentially compromising the ability of the immune system to mount an effective defense against viral infections and favoring persistent infection and progression to HGSIL (36) (Fig. 4E). The activation of DNA repair genes may be a response to potential damage caused by viral oncoproteins E6 or E7, which
aim to integrate the host genome through DNA double breakpoints (35) (Fig. 4E).

The network representation resulting from GSEA with GO comparing LGSIL to HGSIL provided valuable insights into the molecular landscape (Supplementary Fig. 2A). Three distinct clusters emerged, each revealing specific functional themes: a DNA and chromosome organization cluster, characterized by a dense interconnection of genes primarily related to histones and chromatin modifiers, suggesting a potential role in the epigenetic regulation and structural integrity of the genome; a chromosome segregation cluster with genes predominantly linked to processes such as the mitotic spindle and cell division; and a skin development cluster, offering insights into the gene network governing epidermal differentiation (Supplementary Fig. 2A). These findings suggest a complex interplay of molecular events involving DNA organization, chromosome segregation, and skin differentiation in the transition from LGSIL to HGSIL. Some of these events may be attributed to HPV E6 oncoprotein. The expression of viral E6 enhances cell cycle progression and induces mitotic defects leading to centrosome amplification observed in keratinocytes, contributing to chromosomal instability through aberrant chromosome segregation (37).

Moreover, Disease Ontology (DO) revealed additional clusters of genes related to gut inflammatory processes, HIV disease, and B cell immunodeficiency (Supplementary Fig. 2B). Together, these data unveil the impact on the anal transcriptome caused during the transition from LGSIL to HGSIL, defining distinct driver processes, including several genes that can be new avenues for further research.

Conversely, in comparing HGSIL and ASCC, GO analysis revealed a predominant activation of immune response in ASCC but a decrease in epidermal differentiation-related genes (Fig. 4F). Hallmarks analysis demonstrated activation of IFN pathways emphasizing immune activation. Remarkably, suppression of the p53 pathway may be linked to the overexpression of HPV16 E6 protein (Fig. 4G). The network representation of GO revealed clusters of genes mainly representing immune activation, leukocyte migration, cytokine and immunoglobulin production but also epidermal cell differentiation (Supplementary Fig. 2C). Additionally, DO yield terms related to inflammatory processes of colon, HIV, and skin disease (Supplementary Fig. 2D).

Therefore, unlike the comparison between LGSIL and HGSIL, the data suggest that the transition from HGSIL to ASCC is characterized by a predominance of immune response activation over processes related to cell proliferation or DNA modifications (38).
Host transcriptome reveals two intrinsic signatures with varied features and prognoses.

GSEA highlighted deregulated processes across anal lesion stages, emphasizing central roles for the cell cycle, immune response, viral infection, and epidermal differentiation. We focused on significant gene signatures obtained by GSEA related to these processes to visualize gene expression patterns including epidermal differentiation (30 genes – Fig. 5A, Supplementary Data 6), immune response (72 genes – Fig. 5B, Supplementary Data 6), and cell cycle (86 genes – Fig. 5C, Supplementary Data 6) Heatmaps revealed at least two subtypes within each diagnosis group, one with high expression of the gene signature and the other with low expression. To categorize samples, we introduced "high" and "low" scores based on the average expression of each gene signature, divided by the median value (Fig. 5 A-C).

Next, we incorporated these signatures along with LR and HR HPV and HIV status into the unsupervised clustering of samples. This allowed us to discern two primary clusters with distinct characteristics (Fig. 6A). Cluster I primarily comprised SILs (p< 0.01; 24 out of 26 in Cluster I) with a low immune signature (p<0.001), high epidermal differentiation (p<0.001), a low cell cycle signature (p<0.05), and a smaller number of samples infected with HR HPV types detected at both RNA (p<0.05) and DNA (p<0.05) levels compared to Cluster II. In contrast, Cluster II encompasses 91% of anal cancer cases (p< 0.01; 21 out of 23 ASCC) and 62 % of HGSIL (10 out of 16 HSGIL) It exhibits a higher immune signature score (p<0.001), low epidermal differentiation (p<0.001), a greater number of samples with a high cell cycle signature (p<0.05), and a higher prevalence of HR HPV infections (p<0.05; Fig. 6A). Of note, Cluster II included most of the subjects without HIV (92%; 11 out of 12 HIV-negative cases) compared with Cluster I (p<0.05) which was mainly integrated with PLWH (25 out of 26 cases in Cluster I).

Immune infiltration and cell composition analysis.

We utilized EPIC and ESTIMATE algorithms for predicting immune infiltration and cell fraction composition (Fig. 6B). Cluster II exhibited a higher level of immune infiltration, as determined by EPIC (p<0.001). The analysis of cell composition revealed a significant increase in B cells (p<0.001), CD4 T cells (p<0.001), CD8 T cells (p<0.05), and macrophages (p<0.001), aligning with the high immune signature assigned to this cluster (Supplementary Data 7). A possible explanation for these findings could be the higher prevalence of HIV-negative cases in Cluster II, suggesting a potentially less compromised immune system compared to individuals in Cluster I.
To explore this further, we conducted a comparison of the immune profile between HIV-positive and HIV-negative individuals, irrespective of their cluster assignment. Results revealed a significant reduction in B cells (p<0.01) and CD4+ T cells (p<0.001) among PLWH in our cohort (Supplementary Data 7). This aligns with the asymptomatic phase of HIV infection, characterized by ongoing viral replication leading to a gradual depletion of CD4+ T cells, which can be partially restored with ART. While the impact of HIV on B-cell numbers is less clear, studies indicate a reduction in B-cell counts in HIV-infected individuals (39). Dysregulation of B cells during HIV infection is also influenced by ART therapy. Of note, a significant portion of individuals in our HIV-infected cohort were on ART during recruitment, contributing to observed variations in B cell composition.

Furthermore, we explored whether there was an association between these immune profiling differences and HPV16 infection. Results indicated a significantly higher immune profile of macrophages in HPV16-infected cases (p<0.01; Supplementary Data 7). Previous studies have reported that M2-like macrophages infiltrate HPV16-associated tumors, suppressing antitumor T-cell response and facilitating tumor growth (40).

Overall, Cluster II is represented by ASCC tumors and precancerous lesions with a high immune infiltration. The significance of tumor-infiltrating lymphocytes (TILs) in influencing favorable outcomes across various tumor types, including ASCC, has been reported in the literature (41, 42, 43, 44). Our recent study demonstrated the crucial role of PD-L1 expression in influencing complete response rates and survival outcomes in non-metastatic ASCC patients undergoing standard definitive chemoradiotherapy (17). Motivated by the importance of immune factors in ASCC, we employed the T cell dysfunction and exclusion score (TIDE) in our current study to predict cancer immunotherapy response.

The results yielded a compelling connection between immune-related characteristics and treatment response. Cluster II, characterized by a higher immune signature and immune cell infiltration, exhibited a significantly higher number of responders (p<0.05; Fig. 6C). The TIDE analysis highlighted specific immune cell changes associated with responders, including an increase in CD4+ TILs (p<0.05) and macrophages (p<0.05), and a concurrent decrease in cancer-associated fibroblasts (CAFs, p < 0.01) and endothelial cells (p<0.01) (Supplementary Data 7). These findings underscore the potential predictive value of immune-related parameters in discerning responders and non-responders to cancer immunotherapy in the context of anal cancer progression.
Furthermore, we compared the gene expression profiles of two surrogate markers for HPV-related malignancy, Ki67 and p16. Results showed that both markers were higher in Cluster II (Fig. 6D). Additionally, Cluster I was linked to younger subjects and MSM, while Cluster II was associated with older patients, enriched in TGW and cis-gender women (Fig. 6E). In coincidence with the latter, high p16 expression has been shown to correlate with the female sex and with better outcomes following chemo-radiotherapy (45,46,47).

These findings might help to better understand the molecular landscape within and between different stages of anal lesions and reveal potential biomarkers and therapeutic pathways for further research.

Immune profiling of p16, CD3 / CD8 cells and PD-L1 expression among ASCC

The immunohistochemical (IHC) analysis of p16, CD3, CD8, and PD-L1 in anal cancer not only provides valuable insights into the tumor microenvironment, but also serves as a guide for treatment decisions and aids in predicting patient outcomes (17).

In our study, we explored these markers in 10 (for p16) and 14 (for CD3, CD8 and PD-L1) out of the 23 ASCC samples using IHC. Ninety percent of ASCC (9 out of 10) showed a diffusely positive pattern of p16 (Fig. 7A). The density of CD3 and CD8 TILs was moderate to high in 47% (6 out of 14) of ASCC samples (Fig. 7B). Of note, all these samples exhibited a high immune signature, correlating with increased immune infiltration as assessed by EPIC (Fig. 7C). In this context, tumors with moderate to high CD3 and CD8 expression were associated with lower tumor purity scores (p<0.01) and higher cell fractions of cancer-associated fibroblasts (CAFs) (p<0.05), macrophages (p<0.05), and CD4 T cells (p<0.05) as revealed by EPIC analysis (Fig. 7D). The PD-L1 expression status was assessed in the 14 ASCC cases using the Combined Positive Score (CPS). Notably, 57% of positive cases (8 out of 14) exhibited moderate to high PD-L1 expression levels (CPS > 5%), while the remaining samples showed low PD-L1 expression levels (CPS < 5%; 6 out of 14) (Fig. 7B). This analysis indicates a complex relationship between TILs and tumor microenvironment factors, shaping the immune profile of ASCC tumors and potentially influencing treatment approaches.

Comparative transcriptome analysis of HPV-related squamous cell carcinomas.

We analyzed relevant HPV-associated cancer studies to compare the gene expression signatures identified in ASCC with head and neck squamous cell carcinomas (HNSCC) and cervical squamous cell carcinomas (CSCC) cases. In a
previous study, Zhang et al. conducted RNA-seq on 36 HNSCC (18 HPV+ and 18 HPV-), identifying two HPV+ subtypes. One subtype was enriched in "immune response" related genes, while the other was enriched in "keratinocyte differentiation" related genes (48), which is consistent with our ASCC findings. We applied the gene signature distinguishing these subtypes in HNSCC across our sample cohort, sorted by immune score (Figure 8A; Supplementary Data 8). Additionally, we employed our gene signature, derived from the most significantly deregulated genes in the HGSIL vs. ASCC comparison, on HNSCC samples, grouped by the subtypes defined by the authors (Figure 8B; Supplementary Data 8). Results indicate similar gene expression patterns between locations, with variations in gene composition, yet lined with similar biological processes. For CSCC, we utilized den Boon et al.'s study, despite being microarray-based, due to its comprehensive analysis of premalignant (CIN1, CIN2, and CIN3) and CSCC specimens (49). Like our approach, we established a gene signature by comparing CIN2/CIN3 (comparable to HGSIL) versus CSCC and visualized the gene expression profile in our sample cohort (Figure 8C; Supplementary Data 8). This analysis and the application of our signature to cervical lesion samples, sorted by immune score (Figure 8D; Supplementary Data 8), showed an almost mutually exclusive relationship between immune and epidermal differentiation processes. This suggests a significant decrease in keratinocyte differentiation as the disease progresses, alongside a significant increase in immune response genes.

Mutational profiling of cancer driver genes among ASCC and other squamous cell carcinomas

We conducted mutational profiling on ASCC biopsies from 23 patients based on RNA-seq data, revealing 51 somatic missense mutations in cancer driver genes among 87% of ASCC cases (20 out of 23). We identified mutations in KMT2C (also known as MLL3, 30%), PIK3CA (20%), EP300 (20%), NOTCH1 (15%), IDH1 (15%), PRDM1 (15%), FGFR2 (15%), SETD2 (15%), FGFR3 (10%), MAP3K1 (10%), and MET (10%). Single cases of mutations were found affecting TP53, TET2, ATM, TSC1, EZH2, CASP8, ARID1B, APC, NCOR1, SF3B1, STK11, BRCA1, KDM6A, and STAG2 (Fig. 9A). Several of these mutated genes are commonly found in HPV-driven squamous cancers like cervix, head and neck, vulva, and anus, including KMT2C, EP300, PIK3CA, NOTCH1, FGFR2, ATM, TP53, and BRCA1 (50, 51, 17).

Consistent with our results, comparable frequencies of KMT2C, PIK3CA and the chromatin remodeler EP300, have been reported at the genomic level
through NGS or targeted sequencing among the most mutated genes in ASCC (51,52,53,17).

Our data revealed \textit{KMT2C} mutations at comparable rates in the early stages of anal lesions, reaching 30% in HGSIL and 42% in LGSIL (Fig. 9A), suggesting a potential pivotal role for \textit{KMT2C} as a driver gene in anal carcinogenesis progression. Additionally, increased mutation frequencies for \textit{EP300} (21% in ASCC, 4% in HGSIL, and 13% in LGSIL) and \textit{PI3KCA} (17% in ASCC, 8% in HGSIL, 4% in LGSIL) were observed compared to earlier stages of anal lesions (Fig. 9A), indicating potential shifts in the molecular landscape during disease progression.

A higher mutation rate of 3.5 (21 mutations in 6 samples) was observed in ASCC with a low immune signature compared to the high immune signature group (p < 0.01), which had a mutation rate of 1.76 (30 mutations in 17 samples). This implies distinct tumor subpopulations with mutations in cancer driver genes (Fig. 9A).

Furthermore, all mutations in \textit{KMT2C} (7 mutations in 6 cases), \textit{PRDM1} (3 mutations in 3 cases) and \textit{FGFR2} (3 mutations in 2 cases) occurred in HPV16-infected cases, comprising 25% of total mutations (Fig. 9A). \textit{PRDM1} is a master regulator of lymphoid cell differentiation and a tumor suppressor gene in lymphoma (54). It has been identified as a master regulator for HPV16 E6/E7 proteins (55) Aberrant \textit{FGFR} signaling and HPV16 E5 expression have been shown to be correlated with cervical cancer progression (56). Furthermore, the interaction between HPV16 E5 and \textit{FGFR2} alters keratinocyte differentiation and inhibits tumor-suppressive genes, suggesting a role in the early stages of HPV infection and transformation (56).

Consistent with our findings, previous studies have recognized \textit{KMT2C} and \textit{EP300} as the most frequently mutated genes in metastatic ASCC (51). \textit{KMT2C} mutations are associated with abnormal H3K4 methylation, linked to oncogenic transformation in preclinical models (57). \textit{KMT2C} plays a crucial role in activating \textit{TP53} gene expression, demonstrated by targeted inactivation studies in mice (58).

Regarding \textit{EP300}, the oncoprotein HPV/E6 mediates \textit{TP53} degradation by binding to the histone acetyltransferase \textit{EP300}, inhibiting \textit{EP300}-mediated \textit{TP53} acetylation, and promoting \textit{TP53} degradation (59,60). Consequently, dysregulated histone/chromatin modulation within the context of impaired DNA repair mechanisms emerges as a driver of malignancy. We categorized mutated genes into cancer hallmarks and observed that Genome Instability predominated (Supplementary Data 9). Genes like \textit{KMT2C}, \textit{EP300}, \textit{IDH1},
SETD2, TET2, BRCA1, TP53, APC, ATM, KDM6A, NCOR1, SF3B1, and STAG2 defined a gene network critical for ASCC, regardless of HPV infection, aligning with TP53 association with HPV-HR negativity in our study, consistent with prior research (17,51,52).

To perform a comparative analysis of the mutational profile identified in ASCC with other squamous cell carcinomas, we analyzed two combined cervical cancer datasets (MSK-CESC and TCGA-CESC) and a head and neck cancer dataset (TCGA-HNSC) retrieved from cBioPortal online resource (http://www.cbioportal.org/). Only drivers and putative drivers’ somatic missense or truncating mutations were considered for frequency estimations among cohorts. The comparative analysis showed that one third of the most frequent cancer driver mutations identified in ASCC (8 out of 25 genes) were also frequently mutated (>5% of cases) in CSCC and HNSCC (KMT2C, EP300, PIK3CA, NOTCH1, TP53, CASP8, STK11 and KDM6A) (Fig. 9B).

Our mutational profiling of ASCC biopsies from 23 patients offered valuable insights into the somatic mutation landscape of cancer driver genes, particularly given their derivation from transcriptomic data. However, we recognize the significance of the limited sample size when drawing definitive conclusions.

Discussion

ASCC represents only 2% of all gastrointestinal tumors but is characterized by high morbidity and mortality. Unfortunately, treatment options for ASCC have not evolved in the past 20 years; concurrent chemoradiotherapy continues to be the standard care strategy for non-metastatic cases. For patients with metastasis at diagnosis or those who develop metastatic recurrences after chemoradiation therapy, the 5-year survival rate is below 20% (61). To date, platinum-based chemotherapy doublets are the most commonly used anticancer drugs for palliative chemotherapy, and no targeted agents have been approved. In clinical practice, prognostic factors of survival in ASCC are the T and N stage, sex, differentiation, tumor location, high-risk HPV infection, and occurrence of a complete response after CRT (17). These clinical parameters related to survival cannot be used to personalize therapy or predict treatment response in individual patients. Less is known regarding early-stage prognostic biomarkers of ASCC.

Comprehensive characterization of anal squamous precancerous and cancerous lesions at metatranscriptome and transcriptome levels allowed us to identify the most relevant changes that occur at the cell host and their associated microenvironment – the immune infiltrate and the microbiome –
during the progression from preinvasive to the invasive stages. Unsupervised analyses allowed us to identify two patient clusters (Cluster I and Cluster II) based on their histological diagnosis, microbial composition, cell cycle, immune infiltrate, immune response, viral infection (HIV and HPV), epidermal differentiation and activity of specific metabolic and signaling pathways. Cluster I was mainly composed by LGSIL and HGSIL differentiated and low proliferative cases with low immune infiltrate and almost infected by low-risk HPV types. Meanwhile Cluster II was significantly enriched in ASCC and HGSIL cases with higher immune signature score, low epidermal differentiation, a greater number of samples with a high cell cycle signature, and a higher prevalence of high-risk HPV. In this sense, Cluster II was associated with higher expression of Ki67 and p16, older patients, TGW, and females. These findings align with previous studies that have implicated specific viral infections, immune responses, and molecular pathways in the progression of anal lesions (4,10,17). The observed distinctions between Cluster I and Cluster II provide valuable insights into potential prognostic and therapeutic considerations in the management of anal squamous lesions (62).

Microbiome changes in preinvasive and invasive stages of anal cancer

A comparison of the microbiota composition at phylum and species levels reveals expected differences between SILs and ASCC regarding the prevalence of HR HPV subtypes but also identifies several viruses and bacteria species significantly associated with anal cancer not previously reported. In this sense, *Fusobacterium nucleatum, Fusobacterium gonidiaformans* and *Bacteroides fragilis*, previously associated with CRC progression at early stages (17), were significantly enriched in ASCC compared with premalignant lesions. More importantly, these taxa together with HPV16 contributed with gene encoding enzymes (e.g.: Acca, glyQ, eno, pgk and por) and oncoproteins (FadA and dnaK) and a distinctive ASCC metabolic profile characterized by the enrichment of pathways related to oxidative, energetics or biosynthetic processes, including glycolysis, lipid, amino acid, and nucleotide biosynthesis that could facilitate and promote the survival and proliferation of cancerous cells (10, 29, 30). Among these enzymes and proteins, Acca, glyA, glyQ, eno, pgk, and por were identified in our previous study as associated with precancerous anal lesions, highlighting their roles as metabolic markers in cancer progression (10). In line with our results, Serrano and Villar also found pgk and eno overexpressed in the microbiome of HGSIL subjects, while they propose succinyl-CoA and cobalamin as markers associated with HGSIL (8). This reinforces the idea that HPV-infected cells can modify metabolism by regulating genes involved in cellular growth and metabolism, which is crucial to oncogenesis (63). Considering and validating these microbial proteins as markers could offer alternative tools in cancer prevention.
Our integrative analysis of the host transcriptome provided valuable insights into the molecular landscape underlying anal cancer development. The transition from HGSIL to ASCC was characterized by a statistically significant number of DEG, with notable alterations in keratin expression and overexpression of members of the MAGE gene family in ASCC. Functional analysis revealed key biological processes and pathways associated with each stage. In HGSIL compared to LGSIL, activated processes included nuclear division, chromatin modification, and cell proliferation, aligning with histopathological features indicative of high-grade lesions (35). Conversely, the transition from HGSIL to ASCC revealed immune response activation, marked by upregulation of IFN pathways, highlighting the role of the immune system in the progression to anal squamous cell carcinoma (38). Indeed, patients of Cluster II were characterized by a higher immune signature and immune cell infiltration, as assessed by gene expression profiling of immune cell fractions, IHC of CD3 and CD8 TILs as well as PDL-1 expression.

Therefore, in comparing HGSIL and ASCC, the data underscored the predominance of immune response activation in ASCC, contrasting with the cell proliferation and DNA modification processes observed in the transition from LGSIL to HGSIL. Noteworthy findings included the suppression of the p53 pathway potentially linked to the overexpression of HPV16 E6 protein, highlighting the intricate interplay between viral oncoproteins and host cellular processes in the progression to ASCC (64).

Shared and unique immune and molecular changes across squamous cell carcinomas

Through the integration of transcriptomic studies on HNSCC and CSCC with our ASCC transcriptome, we found shared gene expression patterns across tumor sites indicating a shift towards immune response genes and a decrease in keratinocyte differentiation genes during disease progression from preinvasive to invasive stages. These patterns align with the known biology of HPV carcinogenesis, where HPV E6 oncoprotein downregulates keratinocyte differentiation genes and upregulates mesenchymal lineage genes (65). Regarding the immune response, the elevated immune score and the high frequency of TILs cell fraction observed in cluster II samples from our analysis are also in line with a higher prevalence of HR-HPV. HPV-positive tumors may have increased numbers of TILs, myeloid dendritic cells, and proinflammatory chemokines, which are thought to improve treatment response in patients with head and neck and cervical cancers (66,67). Our results showed that a strong immune response is associated with better treatment outcomes, as
indicated by TIDE score analysis. Studies have shown that TILs may improve treatment responses or outcomes in CSCC patients undergoing chemotherapy or radiotherapy (68,69). TIL-based immunotherapy has shown promise as an alternative treatment for advanced cervical cancer, with positive results (70).

In advanced ASCC, immunotherapy trials primarily focus on targeting PD1/PDL1 and E6/E7 proteins (71). Therefore, combining TIL therapy with checkpoint blockade and HPV E6/E7 vaccination offers a potent anti-tumor therapy with the potential to eradicate malignancy in ASCC completely.

Furthermore, ASCC exhibited somatic missense mutations in cancer driver genes, with KMT2C, PIK3CA and EP300, being the most mutated genes in agreement with previous reports. The prevalence of these mutations varied at different stages of anal precancerous lesions, suggesting their involvement in early stages of anal cancer development. Additionally, there were distinct tumor subpopulations with different mutation rates and immune signatures. Mutations in KMT2C, PRDM1, and FGFR2 were predominantly found in HPV16-infected cases, indicating their association with HPV-related carcinogenesis (51,52). The comparative analysis of mutational profiles across different squamous cell carcinomas, including ASCC, HNSCC, and CSCC, revealed significant overlaps in the mutation landscape. By examining datasets from cBioPortal, encompassing a substantial number of cases, we identified common mutations in several cancer driver genes among these carcinomas. Approximately one third of the frequently mutated genes in ASCC were also prevalent in HNSCC and CSCC, suggesting potential shared molecular mechanisms underlying these cancers. Key drivers such as KMT2C, EP300, PIK3CA, NOTCH1, TP53, CASP8, STK11, and KDM6A emerged as recurrently altered across these cohorts. However, the remaining two thirds of the mutated genes appear to be specific to ASCC, indicating distinct genetic alterations driving the development and progression of anal cancer.

Our study has a number of limitations given its cross-sectional nature and the small sample size utilized for data collection due the rarity of anal cancer. The small sample size might not fully represent the biological diversity and variability within the population under investigation, potentially limiting the generalization of the findings. Furthermore, the high risk of false discovery poses a considerable concern, especially in exploratory analyses or when multiple comparisons are conducted. Due to the cross-sectional design adopted in this study, establishing causal associations becomes challenging. However, this is the first cross-sectional study that identifies metatranscriptomics and transcriptomics changes among premalignant and the malignant stages of anal cancer. Furthermore, these findings provide valuable insights into novel prognostic biomarkers that may help to stratify patients with precancerous lesions in low- vs. high-risk groups of progression.
to the malignant stage. Future research employing larger sample sizes and longitudinal designs would be needed to address these limitations and corroborate our findings.

**Methods**

**Sex as a biological variable**

Sex at birth (male or female) and gender identity (cisgender men (CGM), transgender women (TGW), and cisgender women (CGW)), were incorporated into our study design as biological variables.

**Sample Collection and RNA Sequencing**

We collected 70 anal biopsies from patients with different stages of anal lesions: 31 LGSIL, 16 HGSIL, and 23 ASCC, stored in RNAlater (Thermo Fisher Scientific, USA). Clinical data including age, HPV status, ART treatment, and HIV status were recorded at enrollment. RNA was extracted using miRNeasy Tissue/Cells Advanced Kits (Qiagen), and its quality was assessed on an Agilent 2100 Bioanalyzer. Samples with RNA integrity number (RIN) >7 were chosen for RNA-seq. Directional RNA-seq libraries were prepared using Illumina Total RNA Prep with Ribo-Zero Plus kit. Sequencing was performed on an Illumina Novaseq 6000 platform, yielding approximately 80 million clusters per sample with >92% >Q30 quality scores.

**DNA purification and HPV detection and genotyping**

Samples were collected using Qiagen specimen collection devices (Qiagen, USA) by qualified staff at Fundación Huésped and Hospital Udaondo. DNA purification utilized QIAMP DNA kits (Qiagen, USA). DNA integrity and concentration were assessed by Nanodrop spectrophotometry. HPV detection was performed at Institute Malbrán via PCR using biotinylated Broad-Spectrum General Primers BSGP5+/GP6, designed to amplify a 140 bp fragment of the HPV-L1 gene. Genotyping was conducted using reverse line blot hybridization (RLB) for 36 HPV genotypes (validated by Global HPV LabNet) (72). Biotinylated amplicons were denatured and hybridized with genotype-specific oligonucleotide probes immobilized as parallel lines on membrane strips.

**Metatranscriptomic Data Analysis.**

For metatranscriptomics analysis, the obtained RNA-seq data were processed using the Biobakery suite of tools: KneadData was used to separate the human and the non-human reads; taxonomic profiling was performed using MetaPhlAn to identify and quantify microbial taxa at species level present in the anal samples (73).
Species richness and diversity were calculated using the R function `estimate_richness` from R package `phyloseq` (74). We considered the observed species and Chao1 indices for richness, and the Shannon and Simpson indices for diversity. Beta diversity was measured by Bray–Curtis, weighted UniFrac, and unweighted UniFrac. For Principal Coordinate Analysis, the Aitchison distance was used as the distance metric to analyze the compositional data. To test whether the samples cluster beyond that expected by sampling variability we applied permutational multivariate analysis of variance (PERMANOVA) Differences in richness and diversity indices between groups were determined using the Wilcoxon rank sum test with a significance level of 0.05. For relative abundance analysis and visualization, we used R `phyloseq` packages.

Differential abundance analysis

For determining the relative differential abundance and the multivariable association between subjects’ metadata and microbial features, we used the MaAsLin2 package from the bioBakery suite in R/Bioconductor (75). We used default parameters for normalization (TSS method), transformation (Log), analysis method (LM), correction method (BH), and significance threshold (q-value < 0.25). The minimum abundance for each feature was set to 0.001 (0.1%) while the minimum percent of samples for which a feature was detected (prevalence, Pr) at minimum abundance was used as follows: 0.05 (5%) for viruses, 0.1 (10%) for bacteria and pathways and 0.2 (20%) for gene families.

Pathways and gene family analysis

Metatranscriptomics pathway analysis was conducted using the HMP Unified Metabolic Analysis Network 3 (HUMAnN3) pipeline to investigate potential variations in metabolic pathways. HUMAnN3 employs a multifaceted approach, extracting species profiles from KneadData output, aligning reads to pan-genomes, executing translated searches on unclassified reads, and quantifying gene families and pathways. By default, gene families are UniRef90 annotated and metabolic pathways are annotated using MetaCyc database (76,77). The UniRef90 gene family abundance from HUMAnN3 was then regrouped to Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology (KO) (78) We used the KEGGREST package in R/Bioconductor for KO identifiers and KEGG Mapper reconstruct tool for KEGG pathway maps (79,80).

Data visualization

We used the R package “phyloseq” to create a heatmap representation of taxa abundances. For the unsupervised ordination of samples, we applied the NMDS method and Bray distance in the plot_heatmap function. Heatmap
visualization of differentially represented gene proteins was done with R/Bioconductor and the MultiExperiment Viewer software (MeV v4.9).

Transcriptomic, functional enrichment and immune infiltrate analysis.

The raw short-read sequences were preprocessed using Rfastp from the R/Bioconductor package Rsubread. Quality checks, adapter removal, and trimming of low-quality bases were conducted with Rfastp. Reads were aligned to the human genome reference GRCh38 using the Subread aligner algorithm from Rsubread. Gene expression abundance at the whole-genome level was calculated using featureCounts from Rsubread. Differential gene expression analysis between anal lesion stages (LGSIL vs. HGSIL and HGSIL vs. ASCC) utilized edgeR, with fold changes and adjusted p-values computed based on normalized log2 count per million values. Genes with a log-fold change >1 and adjusted p-value <0.05 were considered differentially expressed.

Functional enrichment analysis of differentially expressed genes employed the clusterProfiler package for Gene Set Enrichment Analysis (GSEA) (81). Functional enrichment results were visualized using enrichplot for Gene Ontology, Hallmark of Cancers, and Disease Ontology terms. Heatmaps were generated using MultiExperiment Viewer (MeV) 4.9.0.

Tumor purity, immune cell infiltration, and T cell dysfunction/exclusion scores were estimated using ESTIMATE, EPIC, and TIDE algorithms, respectively, on normalized count matrices.

For comparative transcriptomics analysis, the GSE74927 dataset for HNSCC and GSE63514 for CSCC were utilized. Raw data were imported into R using GEOquery to obtain normalized matrices for each study. Differential gene expression analysis employed DESeq2 for GSE74927 and limma for GSE63514.

To visualize gene expression patterns, we defined the following gene signatures: for ASCC, we used the gene signature obtained from our comparison of HGSIL vs. ASCC; for HNSCC, we used the gene signature provided in the study by Zhang et al., derived from the differential expression analysis between two HPV+ subgroups of HNSCC (48); for CSCC, we obtained a gene signature from the comparison between CIN2/CIN3 samples (comparable to HGSIL) and CSCC. Gene expression profiles across ASCC, HNSCC, and CSCC matrices were visualized after filtering out genes with <50% variance within each signature. Functional enrichment analysis of resulting genes used the ClusterProfiler package. Heatmaps were generated in MeV 4.9.0 based on immune scores.

Mutational analysis based on RNA-seq data.
The preprocessed reads previously used for the transcriptomic analysis were aligned and mapped to the human genome reference GRCh38 using the Subjunc aligner algorithm provided by Rsubread R/Bioconductor package. Subjunc aligner was developed for aligning RNA-seq reads and for the detection of exon-exon junctions at the same time. The Subjunc mapping results (BAM files) were used for genomic variants detection using the exactSNP variant caller algorithm provided by Rsubread package. The VariantAnnotation R/Bioconductor package was subsequently used for SNPs and InDels filtering of the obtained VCF files based on quality (QUAL > 20) and coverage (DP>10) metrics. Identified variants were annotated, filtered and interpreted using OpenCRAVAT and their aggregated variant databases and resources (GnomAD, Cancer Genome Interpreter, Cancer Hotspots, CIVIC, Cosmic, SIFT, PolyPhen2) for the prediction of somatic mutations in cancer driver genes.

In addition, to perform a comparative analysis of the mutational profile identified in ASCC with other squamous cell carcinomas, we analyzed HNSCC and CSCC datasets obtained from cBioPortal online resource (http://www.cbioportal.org/). Briefly, the mutational profiles of the 25 cancer driver genes mutated in ASCC were retrieved from two combined cervical cancer datasets (MSK-CESC and TCGA-CESC, n=468) and a head and neck cancer dataset (TCGA-HNSC, n=510). Only drivers and putative drivers’ somatic missense or truncating mutations were considered for frequency estimations among cohorts.

Immunohistochemistry analysis of ASCC.

Immunostaining utilized a Roche Benchmark XT system with anti-CD3 (Clone 2GV6, Ventana - Roche), anti-CD8 (Clone SP57, Ventana - Roche), anti-PD-L1 (Clone SP263, Ventana - Roche), and anti-p16 (Clone 6H12, Leica Biosystems) antibodies. Evaluation involved two independent pathologists, with discrepancies resolved by a senior pathologist in four cases. CD3 and CD8 expression levels were averaged across intra- and peritumoral areas and categorized as low (0–34%), moderate (35–64%), or high (65–100%) based on total tumor-related lymphocyte staining. PD-L1 expression was assessed using the Combined Positive Score (CPS) for gastric/gastroesophageal junction adenocarcinoma.

Statistics

We used R/Bioconductor for different statistical comparisons outside of MaAsLin’s analysis. To analyze continuous variables, we utilized either two tailed t-tests or Wilcoxon tests as appropriate. For categorical data, we employed Chi-squared and Fisher tests. Box-plots were created in R using the ggplot package.
Study approval.

This study was approved by the institutional review boards of Fundación Huésped and Hospital de Gastroenterología "Dr. Carlos Bonorino Udaondo," both in Buenos Aires, Argentina. All participants included in this study signed informed consent before being involved in the project.

Data availability

The raw data have been submitted to NCBI GEO database with accession number GSE253560. Supporting data values of figures and table 1 are available as supplementary files. The rest of the data are available from the corresponding author upon reasonable request. All codes and scripts used for data preprocessing and analysis are available at the following Github repository: https://github.com/mabba777/ASCC-transcriptomics.

Author contributions

EL: Investigation, formal analysis, writing the article. VF, MIF, PC: resources; logistics of obtaining samples and clinical data of participants. MES, AMG, JAB, MAP, MG, JR, MK, SI, SW: methodology, research assistance, clinical data of participants. OC, JCR: resources. MCA: Conceived the study, supervision, formal analysis, writing the article.

Acknowledgements

The authors want to acknowledge the invaluable contributions from all study participants and from all the research team at Fundación Huésped, where participants were recruited. This work was supported by the NIH grant CA221208.

Conflict of interest

The authors have declared that no conflict of interest exists.

References


2. Darragh TM, et al. The Lower Anogenital Squamous Terminology Standardization project for HPV-associated lesions: background and consensus recommendations from the College of American Pathologists and


Table 1: Clinical and demographics data of patients.

<table>
<thead>
<tr>
<th>Baseline characteristics</th>
<th>LGSIL n=31</th>
<th>HGSIL n=16</th>
<th>ASCC n=23</th>
<th>P-value *</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (years)</strong></td>
<td>33.22 ± 12.88</td>
<td>41.31 ± 11.14</td>
<td>52.23 ± 13.30</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>Sex at birth</strong></td>
<td></td>
<td></td>
<td></td>
<td>0.011</td>
</tr>
<tr>
<td>Male: 61 (CGM+TGW)</td>
<td>31 (100%)</td>
<td>16 (100%)</td>
<td>14 (61%)</td>
<td></td>
</tr>
<tr>
<td>Female: 9</td>
<td>0</td>
<td>0</td>
<td>9 (39%)</td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td>0.027</td>
</tr>
<tr>
<td>CGM: 53 (all MSM)</td>
<td>28 (90%)</td>
<td>11 (69%)</td>
<td>14 (61%)</td>
<td></td>
</tr>
<tr>
<td>TGW: 8</td>
<td>3 (10%)</td>
<td>5 (31%)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>CGW: 9</td>
<td>0</td>
<td>0</td>
<td>9 (39%)</td>
<td></td>
</tr>
<tr>
<td><strong>HPV DNA status</strong></td>
<td></td>
<td></td>
<td></td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Low Risk</td>
<td>21 (67%)</td>
<td>5 (32%)</td>
<td>2 (9%)</td>
<td></td>
</tr>
<tr>
<td>High Risk</td>
<td>3 (10%)</td>
<td>8 (50%)</td>
<td>16 (70%)</td>
<td></td>
</tr>
<tr>
<td>Undetected</td>
<td>7 (23%)</td>
<td>3 (18%)</td>
<td>5 (21%)</td>
<td></td>
</tr>
<tr>
<td><strong>HIV status</strong></td>
<td></td>
<td></td>
<td></td>
<td>0.004</td>
</tr>
<tr>
<td>Positive</td>
<td>28 (90%)</td>
<td>15 (94%)</td>
<td>9 (52%)</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>3 (10%)</td>
<td>1 (6%)</td>
<td>8 (48%)</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td></td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><strong>ART (HIV positive cases)</strong></td>
<td></td>
<td></td>
<td></td>
<td>0.321</td>
</tr>
<tr>
<td>Treated</td>
<td>27 (96%)</td>
<td>15 (100%)</td>
<td>5 (83%)</td>
<td></td>
</tr>
<tr>
<td>Untreated</td>
<td>1 (4%)</td>
<td>0</td>
<td>1 (17%)</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

* Fisher Exact Test
Table 2: Metabolic pathways enriched in ASCC compared with SILs.

<table>
<thead>
<tr>
<th>PWY_ID</th>
<th>Name</th>
<th>Description</th>
<th>coef</th>
<th>pval</th>
<th>qval</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWY-7219</td>
<td>adenosine_ribonucleotides_de_novo_biosynthesis</td>
<td>Nucleotides synthesis</td>
<td>2.44</td>
<td>0.0015</td>
<td>0.0103</td>
</tr>
<tr>
<td>PWY-7221</td>
<td>guanosine_ribonucleotides_de_novo_biosynthesis</td>
<td>Nucleotides synthesis</td>
<td>1.98</td>
<td>0.0035</td>
<td>0.0113</td>
</tr>
<tr>
<td>SER_GLYSYN</td>
<td>PWY_superpathway_of_L_serine_and_glycine_biosynthesis</td>
<td>Amino acids synthesis</td>
<td>1.93</td>
<td>0.0141</td>
<td>0.0206</td>
</tr>
<tr>
<td>PWY-6122</td>
<td>5_aminoimidazole_ribonucleotide_biosynthesis_II</td>
<td>Nucleotides synthesis</td>
<td>1.76</td>
<td>0.0010</td>
<td>0.0103</td>
</tr>
<tr>
<td>PWY-6277</td>
<td>superpathway_of_5_aminoimidazole_ribonucleotide_biosynthesis</td>
<td>Nucleotides synthesis</td>
<td>1.76</td>
<td>0.0010</td>
<td>0.0103</td>
</tr>
<tr>
<td>PWY-6121</td>
<td>5_aminoimidazole_ribonucleotide_biosynthesis_I</td>
<td>Nucleotides synthesis</td>
<td>1.72</td>
<td>0.0012</td>
<td>0.0103</td>
</tr>
<tr>
<td>VALSYN-PWY</td>
<td>L_valine_biosynthesis</td>
<td>Amino acids synthesis</td>
<td>1.65</td>
<td>0.0022</td>
<td>0.0113</td>
</tr>
<tr>
<td>PWY-7208</td>
<td>superpathway_of_pyrimidine_nucleobases_salvage</td>
<td>Nucleotides synthesis</td>
<td>1.55</td>
<td>0.0034</td>
<td>0.0113</td>
</tr>
<tr>
<td>PWY-7228</td>
<td>superpathway_of_guanosine_nucleotides_de_novo_biosynthesis_I</td>
<td>Nucleotides synthesis</td>
<td>1.54</td>
<td>0.0018</td>
<td>0.0103</td>
</tr>
<tr>
<td>COA-PWY</td>
<td>coenzyme_A_biosynthesis_I</td>
<td>Coenzymes synthesis</td>
<td>1.48</td>
<td>0.0046</td>
<td>0.0113</td>
</tr>
<tr>
<td>PWY-7220</td>
<td>adenosine_deoxyribonucleotides_de_novo_biosynthesis_II</td>
<td>Nucleotides synthesis</td>
<td>1.47</td>
<td>0.0015</td>
<td>0.0103</td>
</tr>
<tr>
<td>PWY-7222</td>
<td>guanosine_deoxyribonucleotides_de_novo_biosynthesis_II</td>
<td>Nucleotides synthesis</td>
<td>1.47</td>
<td>0.0015</td>
<td>0.0103</td>
</tr>
<tr>
<td>PWY-7663</td>
<td>gondoate_biosynthesis</td>
<td>Fatty acid synthesis</td>
<td>1.46</td>
<td>0.0133</td>
<td>0.0206</td>
</tr>
<tr>
<td>PWY-5973</td>
<td>cis_vaccenate_biosynthesis</td>
<td>Fatty acid synthesis</td>
<td>1.41</td>
<td>0.0126</td>
<td>0.0206</td>
</tr>
<tr>
<td>PWY-6151</td>
<td>S_adenosyl_L_methionine_salvage_I</td>
<td>Amino acids synthesis</td>
<td>1.38</td>
<td>0.0051</td>
<td>0.0116</td>
</tr>
<tr>
<td>FASYN-INITIAL-PWY</td>
<td>superpathway_of_fatty_acid_biosynthesis_initiation</td>
<td>Fatty acid synthesis</td>
<td>1.31</td>
<td>0.0038</td>
<td>0.0113</td>
</tr>
<tr>
<td>UDPNAGSYN-PWY</td>
<td>UDP_N acetyl_D glucosamine_biosynthesis_I</td>
<td>Nucleotides sugar synthesis</td>
<td>1.25</td>
<td>0.0047</td>
<td>0.0113</td>
</tr>
<tr>
<td>PWY-6125</td>
<td>superpathway_of_guanosine_nucleotides_de_novo_biosynthesis_II</td>
<td>Nucleotides synthesis</td>
<td>1.22</td>
<td>0.0040</td>
<td>0.0113</td>
</tr>
<tr>
<td>PWY-2942</td>
<td>L_lysine_biosynthesis_III</td>
<td>Amino acids synthesis</td>
<td>1.16</td>
<td>0.0148</td>
<td>0.0206</td>
</tr>
<tr>
<td>PWY-6124</td>
<td>inosine_5._phosphate_biosynthesis_II</td>
<td>Amino acids synthesis</td>
<td>1.15</td>
<td>0.0040</td>
<td>0.0113</td>
</tr>
</tbody>
</table>
Figure 1: Richness, diversity, and microbial profile of LGSIL, HGSIL and ASCC. A. Principal Coordinate Analysis depicting the unsupervised distribution of samples, assessed at the species level based on microbiota composition, and evaluated through Euclidean distance. B. Beta diversity comparison between diagnosis groups and covariates. C. Observed and Chao1 richness indices obtained at species level by metatranscriptome analysis. D. Significantly altered phyla Fusobacteriota, Bacteroidota, Bacillota and Pseudomonadota, were related to ASCC. Statistical significance was calculated with the Wilcoxon signed-rank test. E. Heatmap representation of the relative
abundances of the most abundant bacterial species identified across all samples. Highlighted in red are the taxa significantly enriched in ASCC compared with SIL obtained by MaAsLin2 analysis. * p <0.05; ** p<0.01; *** p <0.001.

Figure 2: Viral composition of LGSIL, HGSIL and ASCC. A. Relative abundance heatmap showing the most prevalent viral species identified in the diagnosis groups using meta-transcriptome analysis (RNA Level). B. Alpha PV-10 was found to be linked to the SIL group, whereas Alpha PV-9 and Alpha PV-7 were associated with ASCC. Statistical significance was derived from MaAsLin2 analysis C. Percentage of patients with detectable viruses of the species Alpha PV 10, 9, and 7 assessed by meta-transcriptome. Statistical significance was determined through the application of the Fisher exact test. D. Tile plot visualizing the HPV types identified through DNA genotyping across the different diagnosis groups. E. Percentage distribution of HPV types, assessed by DNA genotyping and classified into low-risk and high-risk categories. F. Percentage of patients in each diagnostic group with detectable low-risk and high-risk HPV types identified through DNA genotyping. Statistical significance was determined through the application of the Fisher exact test.
Figure 3: Functional and taxonomic enrichment of microbial gene proteins associated with anal lesions. 

A. Heatmap representation of metabolic pathways enriched in ASCC compared with SILs represented by 60 gene proteins contributed by relevant gut microbiota taxa of which *F. nucleatum*, *B. fragilis*, and *C. ureolyticus* are predominant. 

B. Viral proteins identified as differentially abundant in ASCC relative to SILs contributed by high risk and low risk HPV.
Figure 4: Differential gene expression analysis and functional enrichment of transcriptomic data. A. Unsupervised hierarchical clustering of samples classified according to diagnosis groups B-C. Volcano plots representing significant differentially expressed genes (LogFC > 1, adj p-value < 0.05) from the comparisons between LGSIL and HGSIL (B) and between HGSIL and ASCC (C). Upregulated genes are indicated by red arrowheads, while downregulated genes are indicated by blue arrowheads. The top 20 significant genes are shown. D-G. Dot plots of Gene Set Enrichment Analysis obtained from the comparisons between LGSIL and HGSIL (D-E) and between HGSIL and ASCC (F-G). D. Dot plot of significantly activated and suppressed Gene Ontology pathways in HGSIL compared with LGSIL. E. Dot plot of significantly activated and suppressed Hallmarks of Cancer in HGSIL compared with LGSIL. F. Dot plot of significantly activated and suppressed Gene Ontology pathways in ASCC compared with HGSIL. G. Dot plot of significantly activated and suppressed Hallmarks of Cancer in ASCC compared with HGSIL.
Figure 5: Heatmaps illustrating the expression profiles of gene signatures across diagnostic groups—LGSIL, HGSIL, and ASCC. A. Epidermal differentiation signature. B. Immune signature. C. Cell cycle signature. The color coding bar at the bottom of each heatmap indicates the score (high or low) assigned to each sample based on the average expression of the gene signature divided by the median value.
Figure 6: Integrative analysis of host transcriptome of LGSIL, HGSIL, and ASSC. A. Tile plot illustrating signatures scores, HPV status, and HIV status of samples distributed according to the unsupervised clustering analysis. Statistical significance was determined through the application of the Fisher exact test. B. Immune profiling and cell fraction composition for each sample using Estimate and Epic, respectively. C. T cell dysfunction and exclusion score for each sample. Statistical significance was determined through the application of the Fisher exact test. D. Relative mRNA abundance of CDKN2A (p16) and MKI67 (Ki67) across samples in Cluster I versus Cluster II. E. Comparative analysis of clusters for age and gender. Statistical significance was determined through the application of a t-test for age and Chi-square test for gender. * p < 0.05; ** p<0.01; *** p <0.001.
Figure 7: Comprehensive analysis of p16, CD3, CD8 TILs density, and PD-L1 expression in the tumor microenvironment of ASCC. A. Immunohistochemistry (IHC) results of p16 staining on ASCC (10X). B. Representative IHC results depicting high and low expression levels of CD3, CD8, and PD-L1. C. Tile plot illustrating ASCC samples analyzed by IHC, showcasing scores for immune signature, CD3, CD8, and PD-L1 IHC results, along with EPIC cell fractions. D. Box-Plots comparing tumor purity, CAFs, and macrophage levels, as obtained by EPIC, between tumors with low (n=8) and high (n=6) CD3/CD8 TILs. Statistical significance was calculated with the Wilcoxon signed-rank test.
Figure 8: Comparative analysis of gene signature expression patterns and enriched pathways in HNSSC, cervical lesions and anal lesions. A. Heatmap visualization of HNSSC gene signature across our sample cohort, grouped by immune score within each diagnosis category. Additionally, the epidermal differentiation score is displayed. B. Heatmap visualization of the ASCC gene signature expression profile in HNSSC samples organized by subtype classification according to Zhang et al 2016. C. Heatmap visualization of CSCC gene signature across our sample cohort grouped by immune score within each diagnosis category. D. Heatmap visualization of the ASCC gene signature across cervical lesions, arranged in ascending order based on the immune gene profile within each diagnosis category.
Figure 9: Mutational profiles among squamous cell carcinomas. A. Tile plot of the most prevalent somatic cancer driver mutations identified in 23 ASCC cases through transcriptome-based sequencing. The upper color-coded bars provide an indication of the immune signature score and HR-HPV status for each respective sample. On the left barplot, the proportions of somatic mutations within each group are presented, relative to the total number of cases in that specific group. TSG: Tumor Suppressor Gene. B. Comparative frequency of the mutations identified in the ASCC cohort with respect to CSCC and HNSCC retrieved from the TCGA cohorts.