PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Multi-omics data integration reveals metabolome as the top predictor of the cervicovaginal microenvironment

Nicholas A. Bokulich[1☯], **Paweł Łaniewski**[2☯], **Anja Adamov**[1], **Dana M. Chase**[3], **J. Gregory Caporaso**[4], **Melissa M. Herbst-Kralovetz**[2,5]*

1 Laboratory of Food Systems Biotechnology, Institute of Food, Nutrition, and Health, ETH Zürich, Switzerland, **2** Department of Basic Medical Sciences, College of Medicine-Phoenix, University of Arizona, Phoenix, Arizona, United States of America, **3** Arizona Oncology, Phoenix, Arizona, United States of America, **4** Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, Arizona, United States of America, **5** Department of Obstetrics and Gynecology, College of Medicine-Phoenix, University of Arizona, Phoenix, Arizona, United States of America

☯ These authors contributed equally to this work.
* mherbst1@arizona.edu

Abstract

Emerging evidence suggests that host-microbe interaction in the cervicovaginal microenvironment contributes to cervical carcinogenesis, yet dissecting these complex interactions is challenging. Herein, we performed an integrated analysis of multiple "omics" datasets to develop predictive models of the cervicovaginal microenvironment and identify characteristic features of vaginal microbiome, genital inflammation and disease status. Microbiomes, vaginal pH, immunoproteomes and metabolomes were measured in cervicovaginal specimens collected from a cohort (n = 72) of Arizonan women with or without cervical neoplasm. Multi-omics integration methods, including neural networks (mmvec) and Random Forest supervised learning, were utilized to explore potential interactions and develop predictive models. Our integrated analyses revealed that immune and cancer biomarker concentrations were reliably predicted by Random Forest regressors trained on microbial and metabolic features, suggesting close correspondence between the vaginal microbiome, metabolome, and genital inflammation involved in cervical carcinogenesis. Furthermore, we show that features of the microbiome and host microenvironment, including metabolites, microbial taxa, and immune biomarkers are predictive of genital inflammation status, but only weakly to moderately predictive of cervical neoplastic disease status. Different feature classes were important for prediction of different phenotypes. Lipids (e.g. sphingolipids and long-chain unsaturated fatty acids) were strong predictors of genital inflammation, whereas predictions of vaginal microbiota and vaginal pH relied mostly on alterations in amino acid metabolism. Finally, we identified key immune biomarkers associated with the vaginal microbiota composition and vaginal pH (MIF), as well as genital inflammation (IL-6, IL-10, MIP-1α).

OPEN ACCESS

Citation: Bokulich NA, Łaniewski P, Adamov A, Chase DM, Caporaso JG, Herbst-Kralovetz MM (2022) Multi-omics data integration reveals metabolome as the top predictor of the cervicovaginal microenvironment. PLoS Comput Biol 18(2): e1009876. https://doi.org/10.1371/journal.pcbi.1009876

Editor: Nicola Segata, University of Trento, ITALY

Received: September 27, 2021

Accepted: January 28, 2022

Published: February 23, 2022

Data Availability Statement: Bacterial 16s RNA gene sequence data analyzed in this study were deposited in SRA (PRJNA518153). In accordance with our IRB approval and the informed consents deposition of the individual immunoproteome, individual metabolome data and individual patient metadata into public databases is not allowed. A data use agreement with the University of Arizona will be required to share these datasets if proposed research is consistent with the scope of the informed consent and requests can be made to COMPHX-WomensHealth@arizona.edu. The code

for the data analysis and the modelling is available at https://github.com/bokulich-lab/HPV-multiOmics.

Funding: This study was supported by the Flinn Foundation Grant #1974 to D.M.C. and M.M.H.-K., Flinn Foundation Grant #2244 to M.M.H.-K., National Institutes of Health NCI Informatics Technology for Cancer Research Award 1U24CA248454-01 to J.G.C., and the National Institutes of Health NCI awards for the Partnership of Native American Cancer Prevention U54CA143924 (UACC) to M.M.H.-K and U54CA143925 (NAU) to J.G.C. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: ASV, amplicon sequencing variants; AUC, area under the curve; BV, bacterial vaginosis; CIN, cervical intraepithelial lesion; Ctrl, control; CVL, cervicovaginal lavage; HPV, human papillomavirus; HSIL, high grade squamous intraepithelial lesions; ICC, invasive cervical carcinoma; LD, *Lactobacillus* dominance; LSIL, low grade squamous intraepithelial lesions; NHW, non-Hispanic white; NLD, non-*Lactobacillus* dominance.

Author summary

This work was undertaken to improve our understanding of interactions between microbes, metabolites and the host in the cervicovaginal microenvironment. We employed a multi-omics approach to investigate relationships between microbiome, vaginal pH, metabolome, immunoproteome in women with and without cervical neoplasm identifying a tight link to abundance of *Lactobacillus* spp. We established predictive models and identified key signatures related to vaginal microbiota, vaginal pH and genital inflammation. Integration of multiple different "omics" data types resulted in only modest increases in prediction accuracy compared to models trained on a single data type. Since the most predictive data type was not known *a priori*, this multi-omics approach yielded insights that would not have been possible with any single data type. Metabolomics data was predictive of different features of the cervicovaginal microenvironment and host response but integrating multi-omics data is likely to be essential for realizing the advances promised by microbiome research.

Introduction

Despite the availability of preventive measures, such as human papillomavirus (HPV) vaccination and Pap smear screening, cervical cancer remains a major public health problem, particularly in low- and middle-income countries [1]. Infection with high-risk HPV types is a well-established risk factor for cervical cancer [2], but is not sufficient for development of the highest risk precancerous cervical dysplasia and progression to cancer [3]. This suggests that other factors in the local cervicovaginal microenvironment play a role during cervical carcinogenesis [4].

The human microbiome (collectively the microbiota, or microbial communities residing in and on the human body, and their theater of activity [5]) is a key regulator of mucosal homeostasis at various body sites, including the female reproductive tract [6]. The cervix and vagina in the majority of healthy, reproductive-age women are colonized by one or few *Lactobacillus* species [7]. These beneficial microorganisms produce lactic acid (lowering vaginal pH, typically below 4.5) and other antimicrobial products. Collectively, multifaceted interactions between *Lactobacillus* and the host create a protective microenvironment against invading pathogens, including HPV [8,9]. However, during dysbiosis *Lactobacillus* spp. are depleted and replaced by a diverse consortium of anaerobes, resulting in elevated vaginal pH [10,11].

Multiple cross-sectional studies in various racial/ethnic cohorts consistently demonstrated that HPV-positive women exhibit more diverse, non-*Lactobacillus* dominant (NLD) vaginal microbiota compared to HPV-negative women [12–15]. Women with cervical dysplasia or cancer also commonly lack *Lactobacillus* dominance (LD) [16–21]. Furthermore, bacterial vaginosis (BV), which is microbiologically characterized as an overgrowth of anaerobes, has been linked to an increased risk of HPV acquisition and persistence [22–24]. Limited longitudinal studies also demonstrated that LD correlates with HPV clearance and regression of dysplasia, whereas NLD microbiota is associated with HPV persistence [25–28]. Recent systematic reviews and meta-analyses of available studies support a causal link between dysbiotic vaginal microbiota and cervical cancer through the impact of bacteria on HPV acquisition, persistence, and progression to dysplasia [29–31].

Metabolomics studies have reported that HPV infection and cervical dysplasia relate to depletion of amino acid, peptide, and nucleotide signatures in the cervicovaginal microenvironment [32,33]. Intriguingly, these metabolic alterations are also associated with depletion of

Lactobacillus spp., connecting HPV infection to vaginal dysbiosis [32,34]. Alternatively, cervical carcinoma profoundly perturbs lipid signatures, such as sphingomyelins [32], which are also biomarkers of chronic inflammation [35] and associated with genital inflammation [32].

It is well documented that persistent HPV infection suppresses immune responses, which may contribute to progression of HPV-mediated neoplasm [36]. Yet, the impact of the microbiome on host defenses in the context of cervical neoplasia has not been comprehensively studied. Recently we showed that dysbiotic microbiota correlates with increased pro-inflammatory cytokines, growth factors, and immune checkpoint proteins in the cervicovaginal fluids [17,37,38]. Another cross-sectional study suggested a link between dysbiotic fusobacteria and immunosuppressive host responses [18]. Taken together, these reports strongly implicate interactions between HPV, microbiota, and host response mechanisms in the local microenvironment in the progression of (or protection from) neoplastic disease.

Here we employ multiple machine learning algorithms (neural networks and Random Forest classification and regression) to integrate omics datasets including vaginal microbiome [17], pH [17], metabolome [32] and immunoproteome [17,37,38] collected from women with and without cervical neoplasia. We present new predictive models of *Lactobacillus* dominance, vaginal pH, genital inflammation and cervical neoplastic disease, and discuss the relative contribution of different features and feature types to our top-performing models.

Results

Participant and clinical sample characteristics

In a previous multicenter study, we enrolled 100 pre-menopausal, non-pregnant participants, including HPV-negative (Ctrl HPV-) and HPV-positive women without cervical neoplasm (Ctrl HPV+), women with low-grade (LSIL) and high-grade squamous intraepithelial lesions (HSIL), and women newly diagnosed with invasive cervical carcinoma (ICC) [17]. Microbiome [17], metabolome [32] and immunoproteome analyses [17,37,38] were performed on collected cervicovaginal samples (**Fig 1**). The vaginal microbiota compositions were determined by 16S rRNA gene sequencing revealing 763 amplicon sequencing variants (ASVs). Cervicovaginal metabolic fingerprints were profiled by liquid chromatography-mass spectrometry and identified 467 unique metabolites. Levels of immune mediators and other cancer-related proteins in cervicovaginal lavage (CVL) samples were evaluated using multiplex cytometric bead arrays for 68 targets. These data, which were previously analyzed independently, were integrated resulting in 72 samples with complete microbiome, metabolome and immunoproteome data for the bioinformatics analyses presented here. Seventy-two patients were classified into five disease groups: Ctrl HPV- (n = 18), Ctrl HPV+ (n = 9), LSIL (n = 10), HSIL (n = 27) and ICC (n = 8). Sixty-one women (85%) were Caucasian and 11 women (15%) were of other races. Thirty-five women (49%) identified themselves as Hispanic/Latina. The average age of participants was 38 years old (ranging from 22 to 58). Forty-nine women (68%) were overweight [body mass index (BMI) >25]. Age, race, ethnicity and BMI were not significantly different among the disease groups. Thirty-eight women (53%) exhibited high vaginal pH (>5.0). Vaginal pH significantly varied among the disease groups ranging from 17% women with high pH in Ctrl HPV+ group to 88% women in ICC group ($P = 0.0002$).

Clustering of omics features according to patient covariates

We performed a principal coordinate analysis (PCoA) of the microbiome data using the Jaccard distance where the first two coordinates explained 20.8% of the observed sample variance (**Fig 2A–2D**). For the metabolome and immunoproteome features we performed a principal component analysis (PCA). The first two components accounted for 47.5% of the sample

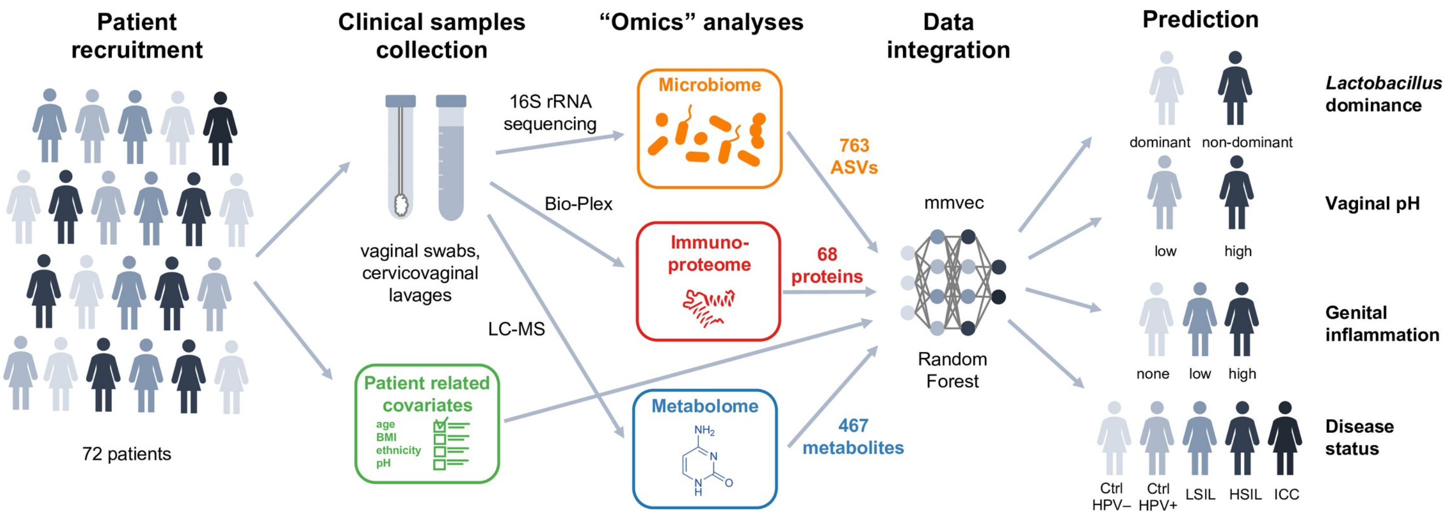

Fig 1. Schematic of a multi-omics approach to study the complex interplay between HPV, host and microbiota in women across cervical neoplasia. In this multicenter study n = 72 women were enrolled with invasive cervical carcinoma (ICC), high- and low-grade squamous intraepithelial lesions (HSIL, LSIL), as well as, HPV-positive and healthy HPV-negative controls (Ctrl). Two vaginal swabs and cervicovaginal lavage (CVL) were collected from each participant. Vaginal swabs were used for microbiome analysis and to evaluate vaginal pH. CVL samples were used for metabolome and immunoproteome analyses. The vaginal microbiota compositions were determined by 16S rRNA gene sequencing revealing 763 amplicon sequencing variants (ASVs). Cervicovaginal metabolic fingerprints in CVL samples were profiled by liquid chromatography-mass spectrometry and identified 467 unique metabolites. Levels of immune mediators and other cancer-related proteins in CVL samples (68 targets) were evaluated using multiplex cytometric bead arrays. Principal component, hierarchical clustering, neural network (mmvec) and Random Forest analyses were utilized to explore associations among multi-omics data sets to predict *Lactobacillus* dominance (dominant vs. non-dominant), vaginal pH (low ≤5 vs. high >5), evidence of genital inflammation (high, low, none) and disease status (Ctrl HPV–, Ctrl HPV+, LSIL, HSIL, ICC).

https://doi.org/10.1371/journal.pcbi.1009876.g001

variance for metabolome samples (Fig 2E–2H) and for 44.3% of the variance for immunoproteome samples (Fig 2I–2L).

Evaluating the clustering of omics datasets based on defined patient covariates, we found that microbiome samples cluster significantly according to vaginal pH ($pH \leq 5.0$ defined as "low" and $pH > 5.0$ as "high", Fig 2C) and by Lactobacillus dominance ("LD" representing samples with relative abundance ≥ 80% of *Lactobacillus* ASVs, Fig 2D). For metabolome features a significant clustering was observed for all four patient covariates (Fig 2E–2H). For immunoproteome data a significant clustering was only found for genital inflammation (defined through a binned scoring system [17] (Fig 2J).

Interconnection of vaginal microbiome, metabolome, and immune biomarkers

Microbe-metabolite interactions were predicted using mmvec [39]. Numerous lipids (including sphingolipids and long-chain unsaturated fatty acids) were associated with multiple ASVs belonging to *Prevotella* (including *Prevotella bivia*), *Peptoniphilus*, *Streptococcus anginosus*, *Atopobium vaginae*, *Sneathia sanguinegenes*, *Veillonellales*, *Finegoldia*, and other taxonomic groups (Fig 3). *Lactobacillus* ASVs (*Lactobacillus crispatus*, *Lactobacillus iners*, *Lactobacillus_H*), some *Prevotella* (including *Prevotella bivia*), and other ASVs, were correlated with a range of metabolites including phenylalanylglycine, the anti-inflammatory nucleotide cytosine, glycerophosphoglycerol, glycerol, N-acetyl methionine sulfoxide, and maltopentaose (Fig 3). These separations roughly mirror genital inflammation and disease status categories, corresponding with our present findings (described below) and previous work showing association between many of these lipids, ICC, and high inflammation; and between these non-lipid metabolites, LD, and low inflammation [17,32]. Three-hydroxybutyrate, previously associated

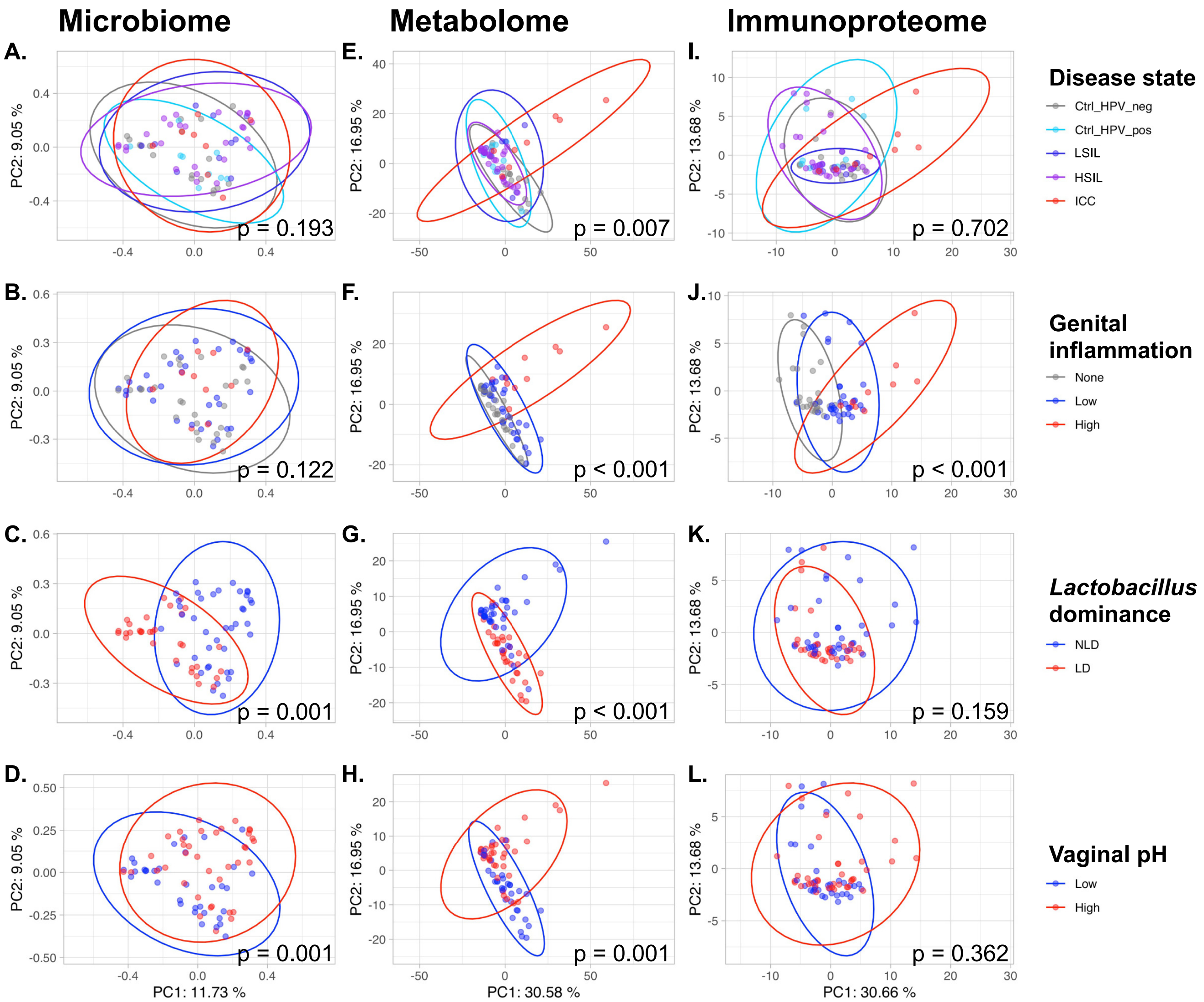

Fig 2. Metabolome features cluster most significantly according to patient covariate groups. A-D. Principal coordinate analysis (PCoA) of the Jaccard distance calculated from microbiome samples. The differences among the groups were tested for significance using a PERMANOVA on the distance matrices. **E-L.** For metabolome (E-H) and immunoproteome (I-L) features the principal component analysis (PCA) was performed on log-transformed and scaled features (zero mean and unit variance). The differences among groups were assessed using the multivariate analysis of variance (MANOVA) model for the first two principal components.

https://doi.org/10.1371/journal.pcbi.1009876.g002

with ICC [32], pipecolate, N-acetylcadaverine, and deoxycarnitine were highly correlated with a range of *Streptococcus*, *Prevotella* (including *P. bivia*), *Megasphaera*, *Finegoldia*, *A. vaginae*, *Sneathia amnii*, and *S. sanguinegens* ASVs. Interestingly, 3-hydroxybutyrate was also correlated to *L. iners*.

To further dissect relationships among the metabolite, microbiome, and immunoproteome, Random Forest regression with 10-fold cross-validation was used to determine the ability to predict the abundance of individual metabolites based on microbiome and immunoproteome profiles, revealing very strong predictive strength for a wide variety of targets (S1 Fig and S2 Table). This includes the inflammation- and ICC-associated lipids 1-palmitoyl-2-arachidonoyl-gpe (16:0/20:4), 1-palmitoyl-2-linoleoyl-gpc (16:0/18:2), 1,2-dilinoleoyl-gpc (18:2/18:2),

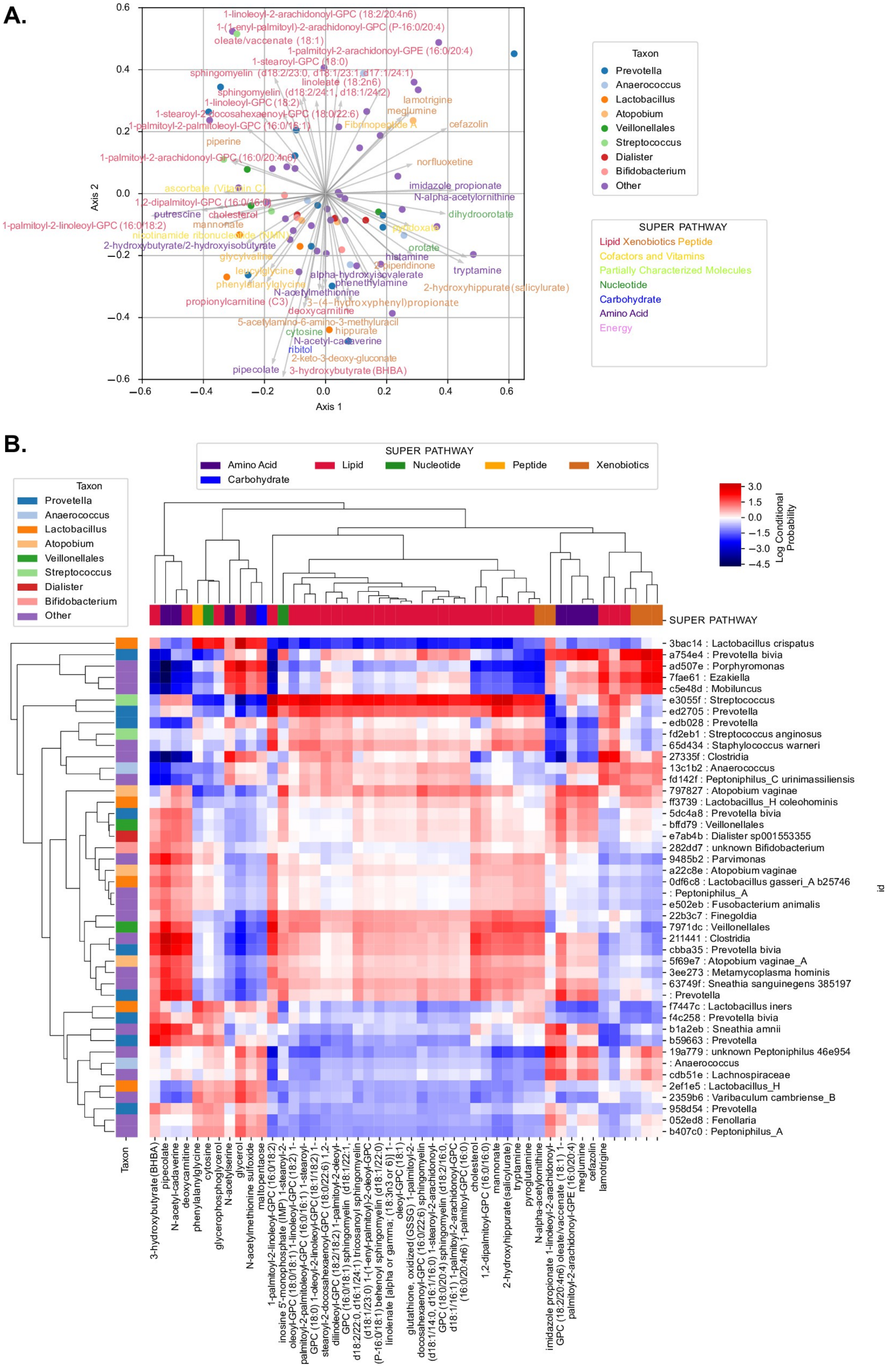

Fig 3. Microbiome-metabolome interaction probabilities via mmvec predict strong associations between lipid metabolites with *Prevotella*, *Streptococcus*, *Atopobium*, *Sneathia* and other clades. A. The principal component analysis

(PCA) biplot displays the top correlations, colored by genus (for microbial features) or by super pathway (for metabolite features). The correlations were tested using mmvec. This method uses neural networks for estimating microbe-metabolite interactions through their co-occurrence probabilities. Microbes (points) and metabolites (arrows) that appear closer to each other in the biplot have a higher likelihood of co-occurring. **B.** The heatmap depicts the correlation coefficients between ASVs and metabolites; hierarchical clustering was done via average weighted Bray-Curtis distance. ASVs were determined using the consensus taxonomy (see Materials and Methods section).

https://doi.org/10.1371/journal.pcbi.1009876.g003

1-palmitoyl-2-docosahexaenoyl-gpc (16:0/22:6), several sphingomyelins, 1-stearoyl-2-docosahexaenoyl-gpc (18:0/22:6), 1-linoleoyl-2-arachidonoyl-gpc (18:2/20:4n6), 1-palmitoyl-2-arachidonoyl-gpc (16:0/20:4n6), and the bile acid glycochenodeoxycholate (**S1 Fig** and **S2 Table**). Many of these associations are driven by high abundances of these lipids, sphingomyelins, and other metabolites in cancer cases: cancer biomarkers are the top predictive features for all of these metabolites (**S2 Fig**), and when ICC cases are removed from the dataset microbial features (including several *Sneathia*, *Atopobium*, *Prevotella*, *Finegoldia*, and *Mobiluncus* ASVs) are included among the top predictive features, though high predictive strength remains for many (but not all) of these targets (**S3** and **S4 Figs**). The ability to accurately predict the abundance of these metabolites through cross-validation highlights the close correspondence between the metabolome, microbiome, and immunoproteome across patients, both respective and irrespective of cancer diagnosis.

Random Forest regression was also performed to predict concentration of immunoproteomic biomarkers based on microbiome and metabolome profiles, demonstrating strong predictive strength for several targets, including proinflammatory cytokines and chemokines (IL-1β, IL-6, IL-8, MIF, MIP-1β), the anti-inflammatory cytokine IL-10, growth factors (HGF, SCF, TGF-α,) apoptosis-related proteins (sFAS, TRAIL), the hormone prolactin, the cytokeratin CYFRA21-1, and other cancer biomarkers (AFP, sCD40L, CEA) (**S5 Fig**). Metabolites (primarily inflammation-associated lipids) are the most predictive features for each of these targets, but microbial features occur among the top 25 predictive features for many of these, most notably *Coriobacteriales bacterium* DNF00809, *S. amnii*, *Veillonellales*, *S. sanguingegens*, *P. bivia*, *Parvimonas*, *A. vaginae* dominating the top important features for predicting cervicovaginal CEA concentration, regardless of cancer diagnosis (**S6 Fig**). Several of these biomarkers are clearly related to ICC, as indicated by reduced predictive strength after ICC cases are removed from the dataset; however, most of these markers exhibit similar performance and important feature associations after removing ICC cases (**S7** and **S8 Figs**).

These findings indicate that both the metabolome and microbiome are highly correlated with and predictive of immunoproteomic biomarker concentrations in the cervicovaginal mucosa. Hence, metabolome and microbiome composition can be considered proxy measurements for genital inflammation and suggest immunological responses linked to cervicovaginal carcinogenesis, a relationship that is more explicitly tested below.

Metabolome and immunoproteome markers predict *Lactobacillus* dominance and vaginal pH

To evaluate the ability of metabolome and immunoproteome features to predict LD (as a proxy for their association with vaginal health), we used Random Forest classification with 10-fold cross-validation. We define LD as any sample in which *Lactobacillus* ASVs collectively comprise $\geq$ 80% of the vaginal microbiome, and grouped subjects into LD (n = 32) and NLD groups (n = 40). Microbiome data were excluded from the predictive model, as these measurements are non-independent due to compositionality constraints, i.e., changing the relative abundance of one feature (such as a *Lactobacillus* ASV) will alter the relative abundance of other features.

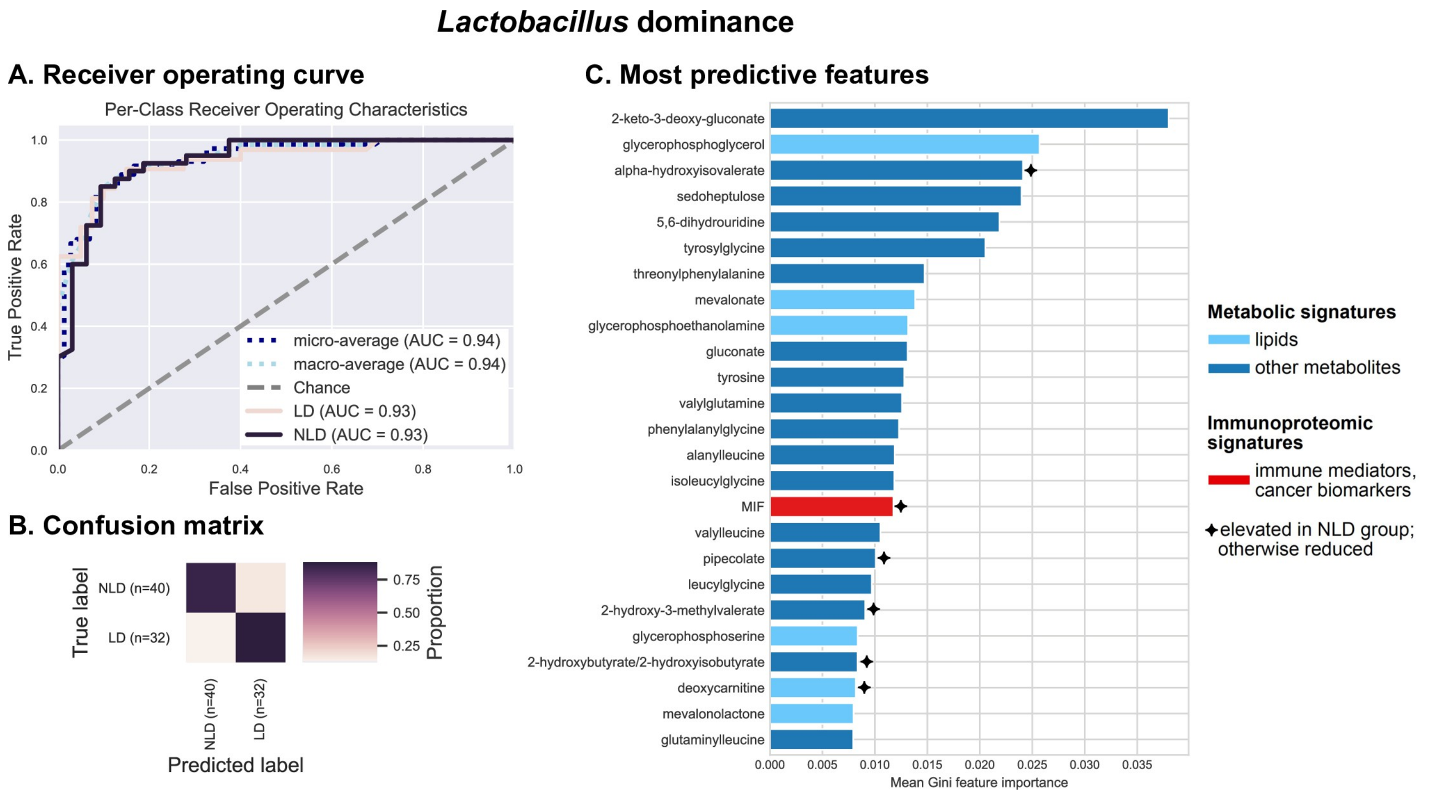

Fig 4. Metabolites (particularly xenobiotics, carbohydrates, amino acids and peptides) and the inflammatory cytokine MIF can accurately predict *Lactobacillus* dominance. Integrated vaginal metabolome and immunoproteome profiles were used as predictive features for training cross-validated Random Forest classifiers to predict whether a subject's vaginal microbiota is *Lactobacillus* dominant (LD $\geq$ 80% relative abundance consists of *Lactobacillus* ASVs) or non-LD (NLD $<$ 80% relative abundance consists of lactobacilli). Combined measurements predict the *Lactobacillus* dominance at an overall accuracy rate of 86.1%. A 1.6-fold improvement over baseline accuracy was observed. Receiver operating characteristics (ROC) analysis showing true and false positive rates for each group, indicating excellent predictive accuracy for both LD (AUC = 0.93) and NLD groups (AUC = 0.93) (**A**). The confusion matrix illustrates the proportion of times each sample receives the correct classification when evaluating the classifier at a threshold of 0.5 (**B**). The graphs depict the 25 most strongly predictive features ranked by their mean Gini importance score across all 10 trained classifiers, a measure of their overall contribution to classifier accuracy (**C**).

https://doi.org/10.1371/journal.pcbi.1009876.g004

Results demonstrate a very high predictive accuracy (average AUC = 0.94), indicating a near-perfect ability to predict LD or NLD across subjects via cross-validation (Fig 4A and 4B). In other words, cervicovaginal metabolome and immunoproteome profiles are tightly linked to the abundance of *Lactobacillus* spp., suggesting that host immunological response is associated with vaginal microbiome composition. The top predictive features consist primarily of non-lipid metabolites, consistent with the mmvec results (Fig 3), though the immunoproteomic biomarkers macrophage migration inhibitory factor (MIF) also rank among the top 25 most important predictive features (Fig 4C). MIF is more abundant in NLD women (S9 Fig), consistent with higher inflammation and ICC.

Vaginal pH is an important feature of the cervicovaginal microenvironment which relates to *Lactobacillus* dominance (Fig 2). We assessed the predictive relationship between pH and cervicovaginal metabolites, microbiota, and immunoproteome using cross-validated Random Forest classification models. Typically, women with LD microbiota have a vaginal pH of 4.5 or lower. However, for the purposes of this analysis, samples were grouped into "low" (pH $\leq$ 5.0, n = 34) and "high" pH groups (pH $>$ 5.0, n = 38). Vaginal pH level is closely related to demographic characteristics, and Hispanic women tend to have slightly higher average vaginal pH compared to NHW [7,17]. We also observed that, in our cohort, the majority of women (75%) with pH 5.0 had LD microbiota (defined as $>$80% *Lactobacillus* abundance). Thus, we defined

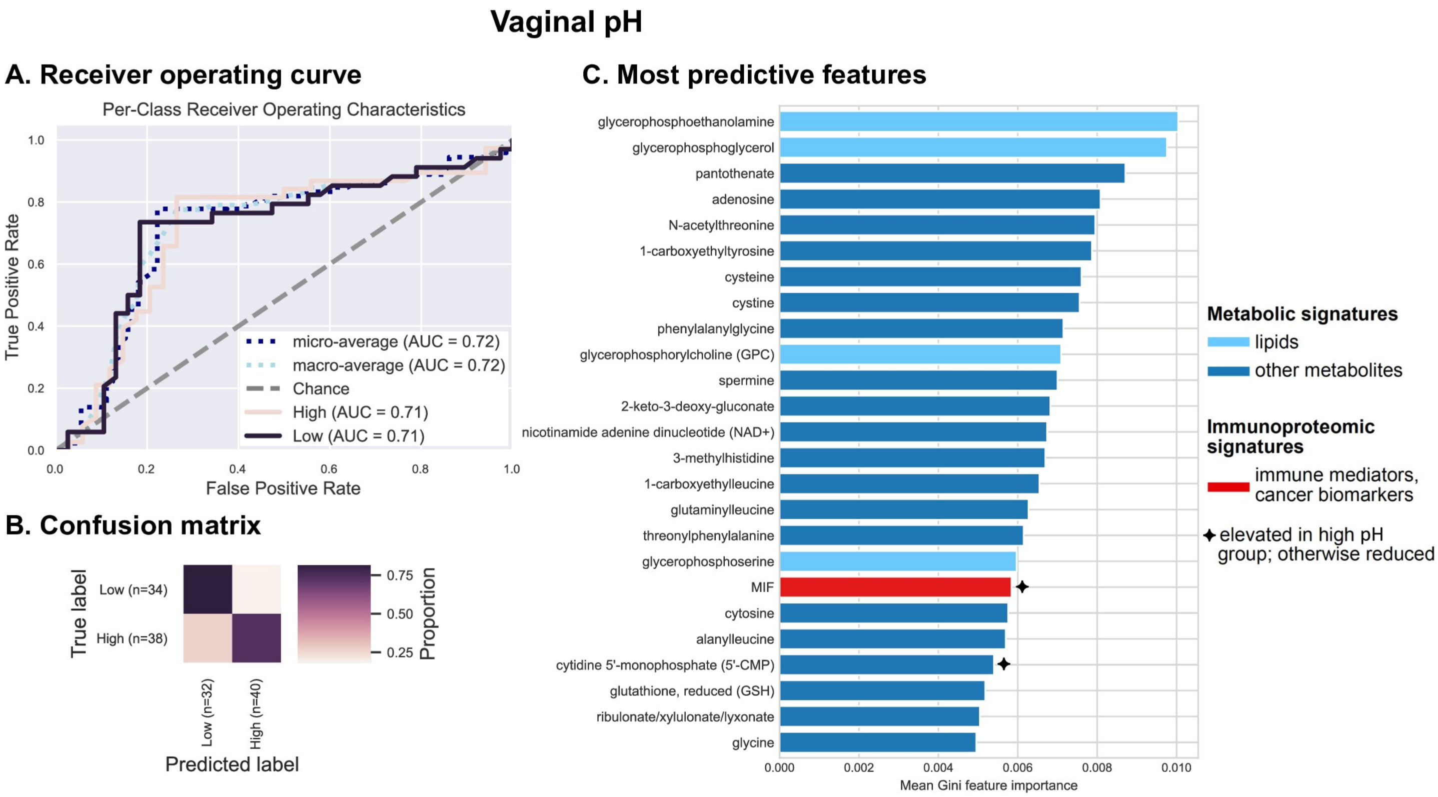

Fig 5. Metabolites (particularly amino acids, peptides and nucleotides) and inflammatory cytokine MIF are the best predictors of vaginal pH. Integrated vaginal microbiome, metabolome, and immunoproteome profiles were used as predictive features for training cross-validated Random Forest classifiers to predict whether a subject's vaginal pH was low (≤ 5.0) or high (> 5.0). Combined measurements predict vaginal pH at an overall accuracy rate of 77.8%. A 1.5-fold improvement over baseline accuracy was observed. Receiver operating characteristics (ROC) analysis showing true and false positive rates for each group, indicating weak predictive accuracy (micro-average AUC = 0.72) for both low (AUC = 0.71) and high pH groups (AUC = 0.71) (**A**). The confusion matrix illustrates the proportion of times each sample receives the correct classification when evaluating the classifier at a threshold of 0.5 (**B**). The graphs depict the 25 most strongly predictive features ranked by their mean Gini importance score across all 10 trained classifiers, a measure of their overall contribution to classifier accuracy (**C**).

https://doi.org/10.1371/journal.pcbi.1009876.g005

pH $\leq$ 5.0 as "low" for the purposes of this study. Results indicate a weak to moderate predictive relationship (AUC = 0.72) (Fig 5A). Predictive power was lost because a large proportion (26.4%) of women with low vaginal pH were predicted to belong to the high pH group (**Figs 5B** and **S10**). Results also indicate that this binary pH model, as expected, exhibits many of the same characteristics as the LD/NLD prediction model: many of the same top predictive features were identified (Fig 5C). Notably, the top predictive features consist primarily of non-lipid metabolites, and MIF is again in the top 25 most important predictors, both associated with high pH as well as NLD (S9 and S11 **Figs**). Hence, together these findings recapitulate the associations between LD, low vaginal pH, and low inflammation, and between NLD, high pH, higher inflammation, and carcinogenesis, as well as the microbial and metabolic context of these states, explored in more detail below.

Metabolome, immunoproteome, and microbiome accurately predict genital inflammation but only moderately predict cancer status

Next, we tested the relationship between the cervicovaginal environment and genital inflammation, as a crucial characteristic of ICC progression. We have previously utilized a scoring system to quantify genital inflammation in our cohort [17]. To assign genital inflammatory scores (0–7), levels of seven cytokines and chemokines, including IL-1α, IL-1β, IL-8, MIP-1β,

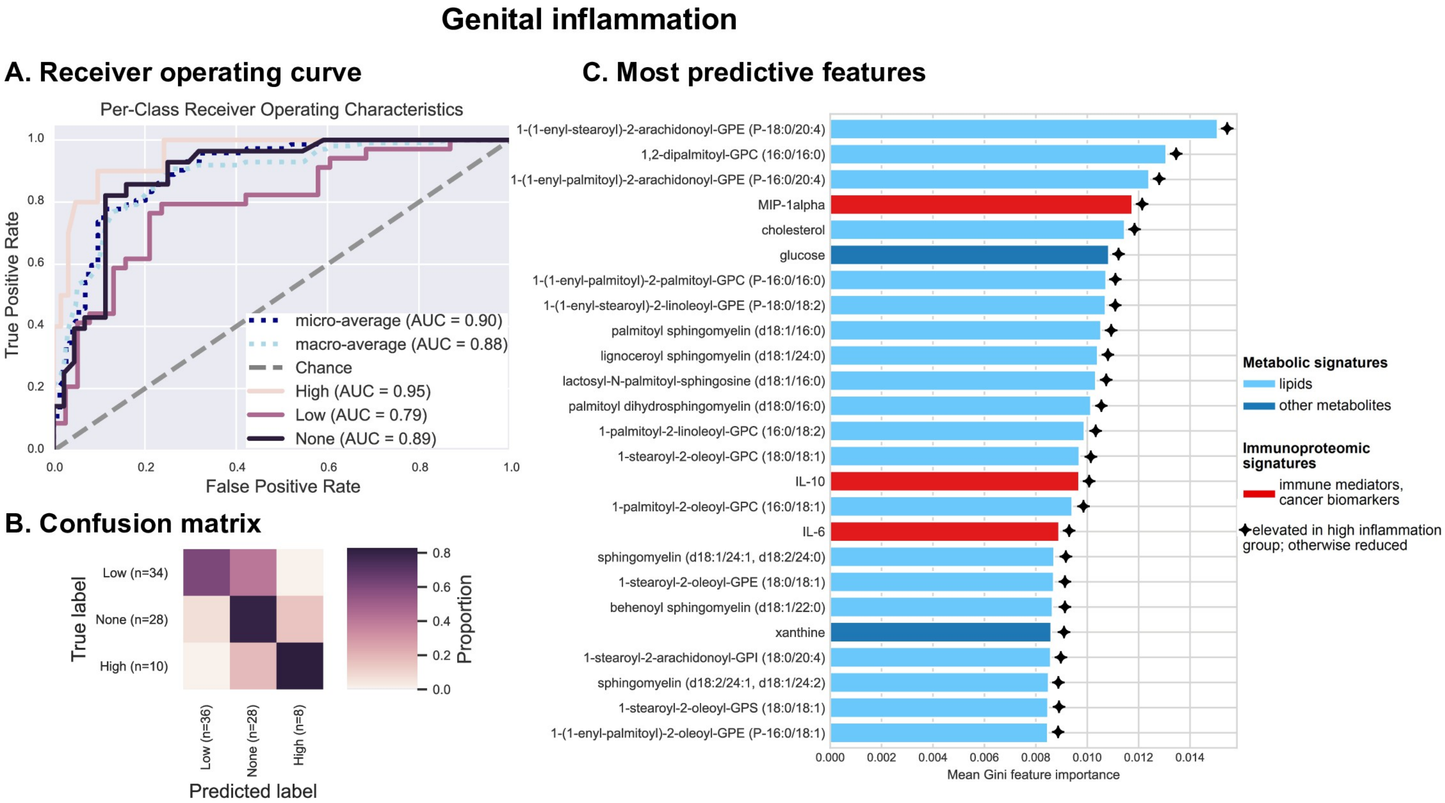

Fig 6. Various metabolites (particularly long-chain fatty acids, sphingolipids and glucose), protein biomarkers (IL-6, IL-10, MIP-1α) are the best predictors of the genital inflammation. Integrated vaginal microbiome, metabolome, and immunoproteome profiles (excluding the 7 cytokines used to score genital inflammation) were used as predictive features for training cross-validated Random Forest classifiers to predict whether a subject's genital inflammation score was "no inflammation" (0), low (1–4), or high (≥ 5.0). Combined measurements predict inflammation score at an overall accuracy rate of 77.8%. A 1.7-fold improvement over baseline accuracy was observed. Receiver operating characteristics (ROC) analysis showing true and false positive rates for each group, indicating moderate average accuracy (micro-average AUC = 0.90) and weak to good predictive accuracy for each group (**A**). The confusion matrix illustrates the proportion of times each sample receives the correct classification when evaluating the classifier at a threshold of 0.5 (**B**). The graphs depict the 25 most strongly predictive features ranked by their mean Gini importance score across all 10 trained classifiers, a measure of their overall contribution to classifier accuracy (**C**).

https://doi.org/10.1371/journal.pcbi.1009876.g006

MIP-3α, RANTES, and TNFα, were measured in cervicovaginal lavages (CVL) and patients were assigned a score based on whether the level of each immune mediator was in the upper quartile. For the purposes of classification, subjects were grouped into no (score = 0, n = 28), low (0 < score < 5, n = 34), or high inflammation (score ≥ 5, n = 10) groups, and Random Forest classifiers were trained and tested via 10-fold cross-validation to assess the ability to predict genital inflammation across subjects based on cervicovaginal microbiome, metabolome, and immunoproteome (excluding the seven inflammatory markers that are used to measure inflammatory score). Results indicate moderately high predictive accuracy (macro-average AUC = 0.88) (Fig 6A). Predictive accuracy is very good for high (AUC = 0.95) and no inflammation (AUC = 0.89), but lowest for low inflammation (AUC = 0.79), due to misclassification of some samples as either high or no inflammation (Fig 6B). Similar to pH classification but to a lesser extent, this reflects the shortcoming of binning samples for classification into categorical groups, a necessary limitation due to the small sample size of the current study. Regression models predicting actual inflammation score demonstrate high accuracy at lower inflammation scores, but lower accuracy at the upper range due to sparsity of high-inflammation samples for cross-validation (S12 Fig). Larger sample sizes in future studies will enable more accurate prediction of low-inflammation samples through prediction of actual inflammation scores, refining our current estimates of associations between genital inflammation and cervicovaginal microenvironment. As it stands, categorical classification performs moderately well, and can identify a range of features predictive of inflammation, primarily lipids, but also several immunoproteomic biomarkers including MIP-1α, IL-10 and IL-6 (**Figs** 6C and S13).

Given the ability to predict genital inflammation, a crucial feature of ICC progression, based on features of the cervicovaginal microenvironment, we sought to determine if cervical neoplasm status could also be predicted based on these features using cross-validated Random Forest classification. Samples (n = 72) were grouped into control HPV- (n = 18), control HPV + (n = 9), LSIL (n = 10), HSIL (n = 27), and ICC (n = 8). This yielded low predictive accuracy (micro-average AUC = 0.74, macro-average AUC = 0.65) (**S14 Fig**). Although many of the same carcinogenesis-related metabolites and immune markers were top predictors in these models, accurate differentiation could not be achieved, primarily because of the low sample size and large class imbalances, but also due to the large number of classes with borderline differences (e.g., high similarity led to misclassification between control HPV–and control HPV + groups, and between LSIL and HSIL groups). Given the low per-group sample sizes, approaches to mitigate class imbalances were not feasible in the current study, but larger sample sizes and pooled analyses will facilitate better estimates in future studies. However, it should be noted that ICC predictive accuracy was moderately high (AUC = 0.76), in spite of the low sample size and class imbalance (**S14 Fig**). This indicates that ICC could be predicted with fairly high accuracy across subjects, but non-ICC groups could not be reliably distinguished due to the similarities between these groups. Combining LSIL and HSIL prior to classification increases accuracy, indicating ambiguity between these groups, as reflected in the imprecise distinction between these histological classifications. Hence, ICC elicits signature characteristics in the cervicovaginal microenvironment across subjects that can be used to identify these subjects, but intermediate stages of progression (HPV infection, LSIL, HSIL) cannot be fully distinguished (**Fig 2E and 2I**). Larger sample sizes and longitudinal measurement in future studies may improve our ability to diagnose ICC or even predict cancer risk based on cervicovaginal microenvironment characteristics (metabolome, immunoproteome, microbiome).

Integrative omics modestly increases predictive accuracy

To test whether integration of multiple omics dataset leads to increased predictive accuracy of our models, we evaluated the performance of each Random Forest classifier with different combinations of data types with the expectation that more data types could only yield better predictive accuracy. Results indicate that integrating data led to modest increases in accuracy for most classification tasks, but with mixed results (**Fig 7**). For LD, combining multiple datasets did not increase accuracy (**Fig 7A**). Metabolites alone could predict LD status with high accuracy; immunoproteome data exhibited lower accuracy. For pH prediction, metabolites, immunoproteome and microbiome datasets on their own could predict pH with moderate accuracy; integrating all three omics datasets led to an overall increase in mean accuracy (**Fig 7B**).

Genital inflammation was the one measurement that showed little change in accuracy with integration of multiple omics datasets (**Fig 7C**). Both metabolome and immunoproteome datasets yielded nearly identical high predictive accuracy, whereas microbiome data exhibited poor predictive accuracy. Combining all three datasets led to a slight increase in predictive accuracy.

Discussion

The vaginal microbiota, HPV infection and cervical neoplasm are related in ways that are still not fully understood. Emerging evidence suggests that *Lactobacillus* dominance (LD) in the

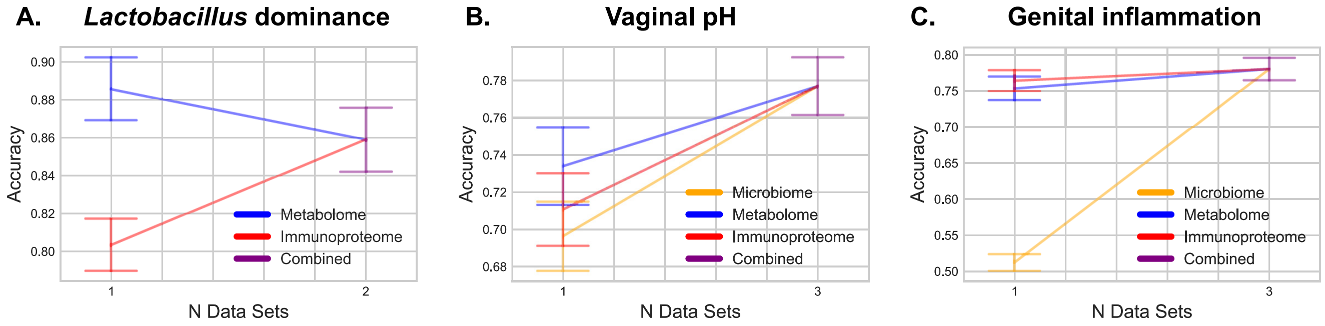

Fig 7. Integrating multiple–omics datasets does not dramatically improve overall prediction accuracy; however, different integration of various measurements are needed for the best prediction of distinct features. Graphs show stepwise accuracy levels for *Lactobacillus* dominance (**A**), vaginal pH (**B**) and genital inflammation (**C**) when Random Forest models are trained on a single omics dataset or combined data containing 2–3 omics datasets. *Lactobacillus* dominance can be explained mostly by metabolome data, vaginal pH by metabolome and microbiome datasets, and genital inflammation by metabolome and immunoproteome datasets. Combining omics datasets leads to higher average accuracy scores for *Lactobacillus* dominance and vaginal pH and genital inflammation classifications, but not for *Lactobacillus* dominance classification.

https://doi.org/10.1371/journal.pcbi.1009876.g007

vagina and cervix relates to HPV clearance and disease regression, whereas dysbiotic anaerobes contribute to HPV persistence and progression of cervical neoplasm [29–31]. Host response to HPV and microbiota, which may result in genital inflammation, immune evasion, and altered metabolism, likely contribute to establishment of persistent infection and disease progression [32,33,40–43]. Thus, improving our understanding of microbiota-virus-host interactions in the local cervicovaginal microenvironment is imperative for the development of novel diagnostic, preventative and therapeutic approaches, which might help reduce cervical cancer burden among unvaccinated women in the future [44].

We investigated relationships between multiple clinical "omics" datasets (microbiome, vaginal pH, metabolome, immunoproteome) collected from women who had not been vaccinated against HPV at different stages of cervical neoplasia (**Fig 1**). Using integrated multi-omics, we aimed to establish predictive models and identify key signatures related to vaginal microbiota structure, vaginal pH, genital inflammation and cervical neoplasm status. We identified specific metabolites that were predictive of *Lactobacillus* dominance, vaginal pH, and genital inflammation (**Figs 4–6**). These findings demonstrate that vaginal microbiota and host defense responses strongly influence cervicovaginal metabolic fingerprints [32,33,45] and indicate that cervicovaginal metabolic signatures might be promising biomarkers for gynecological conditions, including cervical cancer. In addition, select immune mediators and cancer biomarkers also exhibited high importance scores in our analyses for predictions of LD and vaginal pH (MIF), as well as genital inflammation (IL-6, IL-10, MIP-1α), further confirming the link between vaginal microbiota and host immune responses [17,37,40,46,47]. Intriguingly, microbial features did not rank among the top predictors of vaginal pH or genital inflammation, suggesting they had less predictive power than metabolites. On the other hand, our neural network and Random Forest models showed that the abundance of bacterial taxa highly corresponded to levels of key metabolites, immune mediators, and cancer biomarkers related to cervicovaginal health or dysbiosis (**Fig 3**), suggesting tight coupling of the microbiome, metabolome, and immunoproteome [39,48–51].

Using our approach, we predicted the cervical cancer group with good accuracy, however we were unable to accurately predict other cervical neoplasm status. Relatively low sample size and imbalance in disease classification, which are limitations of our study, might have impacted these predictions. Larger numbers of subjects and temporal data on subjects will likely improve predictive models in the future, and better support causal links between microbial dysbiosis and HPV-mediated carcinogenesis. In addition, pathophysiological responses across the continuum of cervical neoplasm might not be uniform among patients with different disease classifications (for example LSIL and HSIL). Indeed, clinical studies have shown contrasting results related to genital inflammation and cervical dysplasia. Some studies report that infection with high-risk HPV types or precancerous dysplasia has not been associated with increased levels of genital inflammation [17,40,43]. Another report showed increased inflammatory cytokines in patients with cervical dysplasia, but it did not control for microbiota composition [42].

Our integrated analyses revealed that different classes of metabolites are important for prediction of different phenotypes: lipids were strong predictors of genital inflammation, while amino acids, peptides and nucleotides were predictive of the vaginal microbiota composition. Sphingolipids and long-chain unsaturated fatty acids in particular ranked as top predictors of genital inflammation. Emerging studies have demonstrated that sphingolipids are implicated in multiple pathological processes, such as inflammatory diseases, diabetes, and cancer [52]. In a previous report we showed that women with cervical cancer had elevated sphingolipids in the cervicovaginal fluids, suggesting that cancer drives associations of phospholipids with inflammation. In addition, we observed the correlation with inflammation even after excluding cancer patients [32]. In fact, sphingolipids are bioactive metabolites, which may mediate inflammatory signaling through TNFα activation [35]. Using neural network analysis, we also showed the co-occurrence of many lipid metabolites and dysbiotic vaginal bacterial taxa (including multiple BV-associated bacteria and *Streptococcus*), linking microbiota to inflammatory markers.

Predictions of vaginal microbiota and vaginal pH relied mostly on alterations in amino acid metabolism, which was in accordance with previous reports on cervicovaginal metabolomes [33,34,45]. Specifically we found that 3-hydroxybutyrate, a ketone body, was strongly correlated with abundance of pathobionts or dysbiotic bacterial taxa, such as *Streptococcus*, *Prevotella*, *Megasphaera*, *Atopobium* and *Sneathia*, and unexpectedly with one ASV classified to the predominant vaginal *Lactobacillus* spp., *L. iners*. *L. iners*-dominant vaginal microbiota has been shown to more often transition to dysbiotic NLD microbiota compared to other *Lactobacillus* spp. [53]. Furthermore, *L. iners* produces a different ratio of lactic acid isoforms [54], which vary in bactericidal capacities [55]; therefore, the protective role of *L. iners* in the cervicovaginal microenvironment is still questionable [56]. We have previously demonstrated that 3-hydroxybutyrate (measured in the cervicovaginal fluids) is an excellent discriminator of cervical cancer patients compared to healthy controls [32]. Several clinical studies also identified 3-hydroxybutyrate (but measured in serum or tissue effusions) as a potential biomarker of other gynecologic malignancies, such as endometrial cancer [57] and ovarian cancer [58,59].

Other key metabolites that we identified to highly correlate with dysbiotic microbiota were pipecolate and deoxycarnitine. In a previous study on metabolomes of women with BV, these two metabolites positively associated with BV status and the presence of "clue cells" [34], a key clinical characteristic of BV. We also revealed that deoxycarnitine in cervicovaginal fluids can discriminate HPV-positive and HPV-negative women without neoplasia [32], linking vaginal dysbiosis with HPV infection. In addition, *Lactobacillus* spp. (particularly *L. crispatus*) positively correlated with N-acetyl methionine sulfoxide, a reactive oxygen species. Production of hydrogen peroxide, another reactive oxygen species, by vaginal *Lactobacillus* spp. has been postulated to have a protective effect against invading pathogens [60,61]. Similarly, an increase of N-acetyl methionine sulfoxide in the cervicovaginal microenvironment might contribute to host protection via oxidative stress.

Through our integrated multi-omics approach, we also identified key immune biomarkers associated with the vaginal microbiota composition and vaginal pH, for instance MIF, a pleiotropic cytokine regulating inflammatory reactions and stress responses [62]. MIF was identified as a top predictive factor of vaginal pH and LD in our Random Forest analysis, which took into account multiple different "omics" data types (**Figs 4 and 5**), suggesting that *Lactobacillus* colonization may be closely involved in regulating markers of genital inflammation, including MIF. In accordance, several reports have demonstrated significantly increased levels of MIF in cervicovaginal fluids of women with vaginal dysbiosis or BV compared to women with healthy LD microbiota [47,63,64]. Previously, we identified cervicovaginal MIF as a potential biomarker for cervical cancer [37]. Immunohistochemical studies demonstrated overexpression of MIF in cervical cancer tissues compared to healthy cervix and dysplasia [65–67]. MIF has been shown to promote cell proliferation, inhibit apoptosis [66] and directly induce secretion of VEGF, an angiogenesis factor [65]. Thus, elevated MIF production induced by dysbiotic vaginal microbiota might contribute to cervical carcinogenesis. Our integrated analysis further highlighted the importance of this key immune mediator, and links its expression to vaginal microbiome and metabolome characteristics. Other immunoproteome biomarkers (IL-6, IL-10, MIP-1α) identified to be associated with genital inflammatory scores likely relate to cancer-induced inflammation rather than a host defense response to dysbiotic vaginal microbiota [37]. Overall, our data indicate that mucosal inflammation is likely associated with cervical neoplasm via the effect of vaginal microbiota on induction of specific inflammatory mediators and metabolites.

Many of the predictive models used in this study integrate metabolome, immunoproteome, and microbiome data. We hypothesized that integrating multiple data types would lead to a cumulative increase in predictive accuracy, as accumulating more features should more completely model the host environment. We instead observed that our metabolomics data nearly always drove classifier accuracy, and inclusion of other data types resulted in modest, if any, increases in classifier performance accuracy. There are a few explanations for this that are not mutually exclusive. First, the metabolites profile might contain features that are proxy information for other feature types (e.g., microbial metabolites as a proxy for microbiome), and hence only gain minimal benefit for integration with those other data types and serve as a good predictor of those other features. This is supported by our finding that metabolome and immunoproteome can almost perfectly predict LD where most of the important features are metabolites (**Fig 3**). However, those classifiers do not achieve perfect accuracy even for this simple microbiome summary statistic of LD, therefore we expect that the microbiome provides context about the cervicovaginal microenvironment that is not present in the other feature types used here. Second, our supervised classification approaches may need improvement for integrating data types. This is likely, given that integrating microbiome multi-omics data is currently a very active area of bioinformatics research. In this case, higher accuracy will be possible as feature extraction and normalization methods designed for microbiome multi-omics improve. Third, there may be more variance in microbiomes than metabolomes across individuals (or across samples from the same individual), requiring a larger training data set for microbiome-based classification than for metabolite-based classification. In this case it is possible that a larger training set would allow for accurate microbiome- or immunoproteome-based classification.

Given that we observed only a modest increase in classifier performance accuracy with the use of multiple "omics" data types, it may seem that the benefit of including these additional data does not justify their cost. We provide a few counterpoints to this idea. First, we cannot know, *a priori*, which data type will provide the best predictive accuracy in any given study of a new system (as in our study). The information gained in this multi-omics survey can now be used to prioritize data to collect in future studies, with the caveat that larger sample sizes and additional populations are needed to fully resolve the predictive power of various omics types for cervicovaginal microenvironment across human populations. While the metabolome data in our study appears most predictive, and this finding has been presented in other recent studies [48–50], we suspect that this is system-specific rather than a general principle. Second, integrating multiple feature types may lead to more consistent performance, as shown here, and even modest increases in accuracy are valuable. Furthermore, different feature types were differentially useful for predicting different characteristics of the cervicovaginal environment. Profiling different feature types therefore enabled discoveries that would not have been possible had we focused only on a single feature type. As a result, we still see considerable value in collecting multi-omics data despite achieving consistently high performance from a single feature type in the samples and system under investigation here. We believe that collecting multi-omics data in human microbiome studies will enable a broader understanding of the complex mechanistic interplay between microbes, metabolites, the host immune system, and host phenotype. As we continue to amass data relating microbes and metabolites to the host immune system and phenotype, we suspect that our ability to model features (such as genital inflammation) based on combinations of microbes and metabolites will improve. This will enable design of treatments based on an understanding of, for example, how the presence of a metabolite will impact the abundance of a group of microbes, which in turn will drive or suppress an immune response.

In our previous work, we investigated pairwise associations between pH [17] and microbiome, microbiome and immunoproteome [17,37,38], microbiome and metabolome [32], as well as microbiome and metabolome [32] in the cervicovaginal microenvironment to better understand the complex host-microbe interactions contributing to cervical carcinogenesis. In this study, we employed a multi-omics approach and machine-learning algorithms (neural networks and Random Forest) to move beyond pairwise associations by integrating all available omics datasets and establish predictive models of cervical neoplasm, genital inflammation, pH and microbiome. We also aimed to identify key signatures related to these different features of cervicovaginal microenvironment. Intriguingly, our integrated analyses revealed metabolome as the top predictor of genital inflammation, microbiome, and vaginal pH when integrated with other feature types. In addition, we identified new links between microbial, immune, and metabolic signatures linked to cervical carcinogenesis, which have not been reported previously (e.g., interconnection of 3-hydroxybutyrate and MIF with pathobionts and dysbiotic microbiota).

Although our study provided new insights into the multifaceted host-microbe interplay during cervical carcinogensis, there is much work to be done to improve our approaches for integrated multi-omics analyses. For example, developing machine learning classification tools for microbiome multi-omics data that can handle multiple observations per subject to make better use of longitudinal data, and interactive visualization tools that can assist with exploration and interpretation of multi-omics network data will facilitate work. Combining these approaches with novel methods [68] and databases [69,70] for accurate taxonomic classification of vaginal microbiota will further advance our ability to identify microbial species linked to carcinogenesis and prevention. We posit that integrated multi-omics approaches are essential to enabling many of the advances in human medicine that are promised by microbiome research.

Materials and methods

Ethics statement

The research and related activities involving human subjects were approved by the Institutional Review Boards at University of Arizona (no. 1510171298), University of Arizona Cancer

Center/Dignity Health St. Joseph's Hospital and Medical Center (no. PHXB-15-0027-70-15) and Maricopa Integrated Health Systems (no. 2015–040). All participants provided informed written consent and all research was performed in accordance with the federal guidelines and regulations and the Declaration of Helsinki.

Study population and clinical sample collection

Seventy-two premenopausal, non-pregnant women were recruited at three clinical sites located in Phoenix, Arizona: St. Joseph's Hospital and Medical Center, University of Arizona Cancer Center and Maricopa Integrated Health Systems (now Valleywise Health Medical Center). The participants were grouped as follows: Ctrl HPV- (n = 18), Ctrl HPV+ (n = 9), LSIL (n = 10), HSIL (n = 27) and ICC (n = 8). Classification of patients into the five groups and detailed inclusion/exclusion criteria were described previously [17]. Cervicovaginal lavage (CVL) and two vaginal swabs were collected by a physician using the standardized clinical protocol and processed as described previously [17]. Briefly, the first vaginal swab was collected using ESwab Collection System (cat. no. 480C, COPAN Diagnostics Inc., Murrieta, CA) and stored at -80˚C prior to microbiome analysis. Vaginal pH was measured using the second vaginal swab, nitrazine paper and a pH scale ranging from 4.5 to 7.5 [17]. CVL sample was collected using 10 ml of sterile 0.9% saline solution, cleared by centrifugation and aliquoted to avoid freeze-thaw cycles. CVL samples were also stored at -80˚C prior to immunoproteome and metabolome analyses. Demographic data were collected from surveys and/or medical records.

Omics analyses

Immunoproteome, metabolome and microbiome datasets used in this study were described previously [17,32,37,38].

For immunoproteome analysis, levels of 68 proteins were determined in CVL samples using multiplex cytometric bead arrays: customized MILLIPLEX MAP Human Cytokine/Chemokine I (cat. no. HCYTOMAG-60K), Th17 (cat. no. HTH17MAG-14K), High Sensitivity T Cell (cat. no. HSTCMAG-28SK), Circulating Cancer Biomarker 1 (cat. no. HCCBP1MAG-58K) and Immuno-Oncology Checkpoint Protein 1 (cat. HCKP1-11K) Magnetic Bead Panels (Millipore, Billerica, MA) or enzyme-linked immunosorbent assays: Human IL-1F9 (IL-36γ) ELISA kit (cat. no. ELH-IL1F9, RayBiotech, Norcross, GA) in accordance with the manufacturer's protocols [17,37,38]. Data were collected with a Bio-Plex 200 instrument and analyzed using Manager 5.0 software (Bio-Rad, Hercules, CA). Levels of seven cytokines (IL-1α, IL-1β, IL-8, MIP-1β, MIP-3α, RANTES, and TNFα) were used to determine the genital inflammatory scores; patients were assigned one point for each mediator when the level was in the upper quartile. Patients with inflammatory scores 0, 1–4, 5–7 were considered to have no, low or high genital inflammation, respectively.

Global untargeted metabolome analysis of CVL samples was performed by Metabolon, Inc (Durham, NC) using a Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer operated at 35,000 mass resolution [32]. Metabolites were identified and quantified using Metabolon's Laboratory Information Management Systems (LIMS).

For microbiome analysis, DNA was extracted from vaginal swabs using PowerSoil DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA) following the manufacturer's instructions [17]. Amplicon library preparation and 16S rRNA sequencing were performed by Second Genome Inc. (San Francisco, CA). The V4 region of bacterial 16S rRNA gene was amplified from the genomic DNA using fusion primers and sequenced on the MiSeq platform (Illumina, San Diego, CA).

Bioinformatics analysis

Microbial DNA sequence data were processed and analyzed using the QIIME 2 version 2019.7 [71]. DADA2 [72] was used (via the q2-dada2 QIIME 2 plugin) to quality filter the sequence data, removing PhiX, chimeric, and erroneous reads, and merge paired-end reads. Forward and reverse reads were trimmed to 250 nt prior to denoising with dada2, otherwise default parameter settings were used. Taxonomy was assigned to sequence variants using q2-feature-classifier [73] with the classify-sklearn naive Bayes classification method against (a) the GreenGenes 16S rRNA reference database 13_8 release [74] assuming a uniform taxonomic distribution [68]; (b) the Genome Taxonomy Database (GTDB) [75], assuming a uniform taxonomic distribution; and (c) GTDB, with taxonomic class weights (expected species distributions) assembled from a collection of 1,017 human cervicovaginal microbiota samples derived from the Vaginal Human Microbiome Project (the same reference set used to construct the STIRRUPS database [69]) using q2-clawback [68]. RESCRIPt [70] was used to merge these taxonomies via determination of the last common ancestor (LCA) consensus taxonomy assignment for each feature (giving priority to majority classifications, and using superstring matching to facilitate compatibility between the Greengenes and GTDB taxonomies). Any sequence that failed to classify at phylum level was discarded prior to downstream analysis. Microbial feature tables were evenly sampled at 50,000 sequences per sample prior to supervised classification. We did not apply CLR prior to supervised classification or diversity analyses, as this and many other normalization methods for compositional data were designed for differential abundance tests, and their appropriate application to supervised classification problems is still an open question [76]. Following the recommendations of Knights et al. [77] we did apply rarefaction to avoid introducing library size biases and used the rarefied counts as input, not relative abundances.

Prior to the application of supervised learning, samples were selected based on their availability of all omics features and defined targets, resulting in 72 samples. Additionally, features with the same value for all samples were discarded (69 features affected). Supervised learning was performed in q2-sample-classifier [78] via 10-fold nested cross-validation (classify-samples-ncv method), using Random Forest classification or regression models [79] grown with 500 trees. We did not apply transformation to different omics data before merge as the scaling of measurements of different features is not necessary for decision tree based supervised classification approaches [80], and prediction results do not change as a result of monotone transformation of the training data. Using scikit-learn implementations, the trained classifiers were evaluated on their performance on the test sets of each fold. Evaluation metrics that were employed include the area under curve (AUC) of the receiver operating characteristic (ROC) curve and the confusion matrix calculated at a probability threshold of 0.5. Feature importances were calculated as the mean of the Gini importance scores across all 10 trained classifiers. Trained regressors were evaluated based on the R-squared measure and the scatter plot of true versus predicted values of the test sets. An overview of which combination of omics features was used to train classifiers for selected targets is provided in **S15 Fig**.

Microbe-metabolite interactions were estimated using mmvec [39]. This method uses neural networks for estimating microbe-metabolite interactions through their co-occurrence probabilities. Features with fewer than 10 observations were filtered prior to mmvec analysis. Conditional rank probabilities were used to construct principal coordinate analysis biplots (visualized using matplotlib [81]) that illustrate the co-occurrence probabilities of each metabolite and microbe.

Supporting information

S1 Table. Random Forest regression predictive accuracy for predicting log concentration of 95 selected metabolites, ordered by accuracy (most to least). Low mean squared error and high r-squared values indicate close correspondence between predicted and true values. These values correspond to plots displayed in **S1 Fig**, for the top 20 most accurately predicted features.
(XLS)

S2 Table. Random Forest regression predictive accuracy for predicting log concentration of cancer biomarkers, ordered by accuracy (most to least). Low mean squared error and high r-squared values indicate close correspondence between predicted and true values. These values correspond to plots displayed in **S5 Fig**, for the top 20 most accurately predicted features.
(XLS)

S1 Fig. Microbiome and immunoproteome data accurately predict metabolite abundances. Random Forest regressors with 10-fold cross-validation were used to predict the abundance of each selected metabolite in S1 Table based on combined microbiome and immunoproteome datasets. Scatterplots display the linear regression of predicted vs. true log concentrations for the top 20 most accurately predicted metabolites. Dotted lines indicate an ideal 1:1 slope. Grey lines and shading indicate the regression trend line and 95% CI.
(PDF)

S2 Fig. Microbiome and immunoproteome data accurately predict metabolite abundances. Feature importance of top 15 features used in the final Random Forest regression model for each metabolite prediction displayed in **S1 Fig**. *Microbial features are displayed in red, with the first 6 characters of the ASV ID followed by the genus/species-level Greengenes taxonomy.
(PDF)

S3 Fig. Microbiome and immunoproteome data accurately predict metabolite abundances with cancer cases removed. Plots display the predictive accuracy of the top 20 metabolites displayed in **S1 Fig**, but with cancer cases removed. Predictive accuracy remains high for most metabolites, indicating that cancer cases do not drive the associations observed for that metabolite.
(PDF)

S4 Fig. Microbiome and immunoproteome data accurately predict metabolite abundances with cancer cases removed. Feature importance of top 15 features used in the final Random Forest regression model for each metabolite prediction displayed in **S3 Fig**. *Microbial features are displayed in red, with the first 6 characters of the ASV ID followed by the genus/species-level Greengenes taxonomy.
(PDF)

S5 Fig. Microbiome and metabolome data accurately predict cancer biomarker abundances. Random Forest regressors with 10-fold cross-validation were used to predict the abundance of each selected biomarker in S2 Table based on combined microbiome and metabolome datasets. Scatterplots display the linear regression of predicted vs. true log concentrations for the top 20 most accurately predicted biomarkers. Dotted lines indicate an ideal 1:1 slope. Grey lines and shading indicate the regression trend line and 95% CI.
(PDF)

S6 Fig. Microbiome and metabolome data accurately predict cancer biomarker abundances. Feature importance of top 15 features used in the final Random Forest regression model for each cancer biomarker prediction displayed in S5 Fig. *Microbial features are displayed in red, with the first 6 characters of the ASV ID followed by the genus/species-level Greengenes taxonomy.
(PDF)

S7 Fig. Microbiome and metabolome data accurately predict cancer biomarker abundances with cancer cases removed. Plots display the predictive accuracy of the top 20 biomarkers displayed in S5 Fig, but with cancer cases removed. Predictive accuracy remains high for most metabolites, indicating that cancer cases do not drive the associations observed for that metabolite.
(PDF)

S8 Fig. Microbiome and metabolome data accurately predict cancer biomarker abundances with cancer cases removed. Feature importance of top 20 features used in the final Random Forest regression model for each cancer biomarker prediction displayed in S7 Fig. *Microbial features are displayed in red, with the first 6 characters of the ASV ID followed by the genus/species-level Greengenes taxonomy.
(PDF)

S9 Fig. Abundances of top 25 most predictive features for *Lactobacillus* dominance (LD) vs. non-LD (NLD) Random Forest classification. Boxplots display quartile distributions, swarmplots display individual values of top important feature abundances in LD and NLD groups.
(PDF)

S10 Fig. Vaginal pH distribution is skewed toward low (typical) end of pH range, preventing accurate random forest prediction of pH values. Left, histogram displays number of samples per pH value, binned into increments of 0.5 (min = 4.5, max = 7.5). Right, scatterplot displays true vs. predicted log10 vaginal pH for each subject (using 10-fold cross-validation random forest regressors to predict vaginal pH across subjects), indicating very poor regression results due to pH skew.
(PDF)

S11 Fig. Abundances of top 25 most predictive features for vaginal pH Random Forest classification. Boxplots display quartile distributions, swarmplots display individual values of top important feature abundances in “typical” ($pH \leq 5.0$) and “high” ($pH > 5.0$) groups.
(PDF)

S12 Fig. Genital inflammation score distribution is skewed toward no and low inflammation, reducing predictive accuracy of high-inflammation samples. Left, histogram displays number of samples per genital inflammation score. Right, scatterplot displays true vs. predicted inflammation scores for each subject (using 10-fold cross-validation random forest regressors to predict inflammation score across subjects).
(PDF)

S13 Fig. Abundances of top 25 most predictive features for genital inflammation score Random Forest classification. Boxplots display quartile distributions, swarmplots display individual values of top important feature abundances in no (score = 0), low ($0 < \text{score} < 5$), and high inflammation ($\text{score} \geq 5$) groups.
(PDF)

S14 Fig. Microbiome, metabolome, and immunoproteome data weakly predict disease state. Receiver operating characteristics (ROC) analysis showing true and false positive rates for each group, using random forest classifiers with 10-fold cross-validation to test predictive accuracy across subjects. Higher area under the curve (AUC) indicates better accuracy. Micro-average is calculated across each sample, and hence impacted by class imbalances. Macro-average gives equal weight to the classification of each sample, eliminating the impact of class imbalances on average AUC. Notably, invasive cervical carcinoma (ICC) cases are predicted moderately well, indicating a characteristic signal associated with ICC but not with intermediate stages of progression. HSIL and LSIL = high- and low-grade squamous intraepithelial lesions, respectively.
(PDF)

S15 Fig. Overview of omics features used to predict selected targets with supervised classification models. Column names depict selected targets and row names selected omics features.
(PDF)

Acknowledgments

We would like to thank the patients who enrolled in the study and acknowledge Kelli Williamson, Ann De Jong, Eileen Molzen, Liane Fales, Maureen Sutton for the kind assistance in patient recruitment and sample collection and Drs. Dominique Barnes and Alison Goulder for the assistance with clinical sample and data collection.

Author Contributions

Conceptualization: Dana M. Chase, Melissa M. Herbst-Kralovetz.

Data curation: Nicholas A. Bokulich, Paweł Łaniewski.

Formal analysis: Nicholas A. Bokulich, Anja Adamov.

Funding acquisition: Dana M. Chase, J. Gregory Caporaso, Melissa M. Herbst-Kralovetz.

Investigation: Nicholas A. Bokulich, Paweł Łaniewski, Anja Adamov, J. Gregory Caporaso, Melissa M. Herbst-Kralovetz.

Methodology: Nicholas A. Bokulich, Anja Adamov.

Project administration: Melissa M. Herbst-Kralovetz.

Resources: Dana M. Chase, Melissa M. Herbst-Kralovetz.

Software: Nicholas A. Bokulich, Anja Adamov, J. Gregory Caporaso.

Supervision: J. Gregory Caporaso, Melissa M. Herbst-Kralovetz.

Validation: Nicholas A. Bokulich.

Visualization: Paweł Łaniewski, Anja Adamov.

Writing – original draft: Nicholas A. Bokulich, Paweł Łaniewski.

Writing – review & editing: Anja Adamov, Dana M. Chase, J. Gregory Caporaso, Melissa M. Herbst-Kralovetz.

References

1. Arbyn M, Weiderpass E, Bruni L, de Sanjose S, Saraiya M, Ferlay J, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Glob Health. 2020; 8(2):e191–e203. https://doi.org/10.1016/S2214-109X(19)30482-6 PMID: 31812369

2. Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papillomavirus and cervical cancer. Lancet. 2007; 370(9590):890–907. https://doi.org/10.1016/S0140-6736(07)61416-0 PMID: 17826171

3. Gravitt PE, Winer RL. Natural history of HPV infection across the lifespan: role of viral latency. Viruses. 2017; 9(10). https://doi.org/10.3390/v9100267 PMID: 28934151

4. Łaniewski P, Ilhan ZE, Herbst-Kralovetz MM. The microbiome and gynaecological cancer development, prevention and therapy. Nat Rev Urol. 2020; 17(4):232–50. https://doi.org/10.1038/s41585-020-0286-z PMID: 32071434

5. Berg G, Rybakova D, Fischer D, Cernava T, Verges MC, Charles T, et al. Microbiome definition re-visited: old concepts and new challenges. Microbiome. 2020; 8(1):103. https://doi.org/10.1186/s40168-020-00875-0 PMID: 32605663

6. Integrative HMPRNC. The Integrative Human Microbiome Project. Nature. 2019; 569(7758):641–8. https://doi.org/10.1038/s41586-019-1238-8 PMID: 31142853

7. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, et al. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A. 2011; 108 Suppl 1:4680–7. https://doi.org/10.1073/pnas.1002611107 PMID: 20534435

8. Anahtar MN, Gootenberg DB, Mitchell CM, Kwon DS. Cervicovaginal microbiota and reproductive health: The virtue of simplicity. Cell host & microbe. 2018; 23(2):159–68. https://doi.org/10.1016/j.chom.2018.01.013 PMID: 29447695

9. Martin DH, Marrazzo JM. The vaginal microbiome: Current understanding and future directions. J Infect Dis. 2016; 214 Suppl 1:S36–41. https://doi.org/10.1093/infdis/jiw184 PMID: 27449871

10. Onderdonk AB, Delaney ML, Fichorova RN. The human microbiome during bacterial vaginosis. Clin Microbiol Rev. 2016; 29(2):223–38. https://doi.org/10.1128/CMR.00075-15 PMID: 26864580

11. Hillier SL, Marrazzo J, Holmes KK. Bacterial Vaginosis. In: Holmes KK, Sparling PF, Stamm WE, Piot P, Wasserheit JN, Corey L, et al., editors. Sexually Transmitted Diseases, Fourth Edition: McGraw-Hill Education; 2007. p. 737–68. https://doi.org/10.1097/OLQ.0b013e3181559c5c PMID: 17891031

12. Lee JE, Lee S, Lee H, Song YM, Lee K, Han MJ, et al. Association of the vaginal microbiota with human papillomavirus infection in a Korean twin cohort. PLoS One. 2013; 8(5):e63514. https://doi.org/10.1371/journal.pone.0063514 PMID: 23717441

13. Chen Y, Hong Z, Wang W, Gu L, Gao H, Qiu L, et al. Association between the vaginal microbiome and high-risk human papillomavirus infection in pregnant Chinese women. BMC Infect Dis. 2019; 19(1):677. https://doi.org/10.1186/s12879-019-4279-6 PMID: 31370796

14. Gao W, Weng J, Gao Y, Chen X. Comparison of the vaginal microbiota diversity of women with and without human papillomavirus infection: a cross-sectional study. BMC Infect Dis. 2013; 13:271. https://doi.org/10.1186/1471-2334-13-271 PMID: 23758857

15. Tuominen H, Rautava S, Syrjanen S, Collado MC, Rautava J. HPV infection and bacterial microbiota in the placenta, uterine cervix and oral mucosa. Sci Rep. 2018; 8(1):9787. https://doi.org/10.1038/s41598-018-27980-3 PMID: 29955075

16. Mitra A, MacIntyre DA, Lee YS, Smith A, Marchesi JR, Lehne B, et al. Cervical intraepithelial neoplasia disease progression is associated with increased vaginal microbiome diversity. Sci Rep. 2015; 5:16865. https://doi.org/10.1038/srep16865 PMID: 26574055

17. Łaniewski P, Barnes D, Goulder A, Cui H, Roe DJ, Chase DM, et al. Linking cervicovaginal immune signatures, HPV and microbiota composition in cervical carcinogenesis in non-Hispanic and Hispanic women. Sci Rep. 2018; 8(1):7593. https://doi.org/10.1038/s41598-018-25879-7 PMID: 29765068

18. Audirac-Chalifour A, Torres-Poveda K, Bahena-Roman M, Tellez-Sosa J, Martinez-Barnetche J, Cortina-Ceballos B, et al. Cervical microbiome and cytokine profile at various stages of cervical cancer: a pilot study. PLoS One. 2016; 11(4):e0153274. https://doi.org/10.1371/journal.pone.0153274 PMID: 27115350

19. Oh HY, Kim BS, Seo SS, Kong JS, Lee JK, Park SY, et al. The association of uterine cervical microbiota with an increased risk for cervical intraepithelial neoplasia in Korea. Clin Microbiol Infect. 2015; 21 (7):674 e1–9. https://doi.org/10.1016/j.cmi.2015.02.026 PMID: 25752224

20. Kwasniewski W, Wolun-Cholewa M, Kotarski J, Warchol W, Kuzma D, Kwasniewska A, et al. Microbiota dysbiosis is associated with HPV-induced cervical carcinogenesis. Oncol Lett. 2018; 16(6):7035–47. https://doi.org/10.3892/ol.2018.9509 PMID: 30546437

21. Godoy-Vitorino F, Romaguera J, Zhao C, Vargas-Robles D, Ortiz-Morales G, Vazquez-Sanchez F, et al. Cervicovaginal fungi and bacteria associated with cervical intraepithelial neoplasia and high-risk human papillomavirus infections in a Hispanic population. Front Microbiol. 2018; 9:2533. https://doi.org/10.3389/fmicb.2018.02533 PMID: 30405584

22. Watts DH, Fazzari M, Minkoff H, Hillier SL, Sha B, Glesby M, et al. Effects of bacterial vaginosis and other genital infections on the natural history of human papillomavirus infection in HIV-1-infected and high-risk HIV-1-uninfected women. J Infect Dis. 2005; 191(7):1129–39. https://doi.org/10.1086/427777 PMID: 15747249

23. Gillet E, Meys JF, Verstraelen H, Bosire C, De Sutter P, Temmerman M, et al. Bacterial vaginosis is associated with uterine cervical human papillomavirus infection: a meta-analysis. BMC Infect Dis. 2011; 11:10. https://doi.org/10.1186/1471-2334-11-10 PMID: 21223574

24. Guo YL, You K, Qiao J, Zhao YM, Geng L. Bacterial vaginosis is conducive to the persistence of HPV infection. Int J STD AIDS. 2012; 23(8):581–4. https://doi.org/10.1258/ijsa.2012.011342 PMID: 22930296

25. Brotman RM, Shardell MD, Gajer P, Tracy JK, Zenilman JM, Ravel J, et al. Interplay between the temporal dynamics of the vaginal microbiota and human papillomavirus detection. J Infect Dis. 2014; 210 (11):1723–33. https://doi.org/10.1093/infdis/jiu330 PMID: 24943724

26. Di Paola M, Sani C, Clemente AM, Iossa A, Perissi E, Castronovo G, et al. Characterization of cervicovaginal microbiota in women developing persistent high-risk Human Papillomavirus infection. Sci Rep. 2017; 7(1):10200. https://doi.org/10.1038/s41598-017-09842-6 PMID: 28860468

27. Mitra A, MacIntyre DA, Ntritsos G, Smith A, Tsilidis KK, Marchesi JR, et al. The vaginal microbiota associates with the regression of untreated cervical intraepithelial neoplasia 2 lesions. Nat Commun. 2020; 11(1):1999. https://doi.org/10.1038/s41467-020-15856-y PMID: 32332850

28. Usyk M, Zolnik CP, Castle PE, Porras C, Herrero R, Gradissimo A, et al. Cervicovaginal microbiome and natural history of HPV in a longitudinal study. PLoS Pathog. 2020; 16(3):e1008376. https://doi.org/10.1371/journal.ppat.1008376 PMID: 32214382

29. Norenhag J, Du J, Olovsson M, Verstraelen H, Engstrand L, Brusselaers N. The vaginal microbiota, human papillomavirus and cervical dysplasia: a systematic review and network meta-analysis. BJOG. 2020; 127(2):171–80. https://doi.org/10.1111/1471-0528.15854 PMID: 31237400

30. Wang H, Ma Y, Li R, Chen X, Wan L, Zhao W. Associations of cervicovaginal lactobacilli with high-risk HPV infection, cervical intraepithelial neoplasia, and cancer: a systematic review and meta-analysis. J Infect Dis. 2019; 220(8):1243–1254. https://doi.org/10.1093/infdis/jiz325 PMID: 31242505

31. Brusselaers N, Shrestha S, Van De Wijgert J, Verstraelen H. Vaginal dysbiosis, and the risk of human papillomavirus and cervical cancer: systematic review and meta-analysis. Am J Obstet Gynecol. 2018; 221(1):9–18.e8. https://doi.org/10.1016/j.ajog.2018.12.011 PMID: 30550767

32. Ilhan ZE, Łaniewski P, Thomas N, Roe DJ, Chase DM, Herbst-Kralovetz MM. Deciphering the complex interplay between microbiota, HPV, inflammation and cancer through cervicovaginal metabolic profiling. EBioMedicine. 2019; 44:675–90. https://doi.org/10.1016/j.ebiom.2019.04.028 PMID: 31027917

33. Borgogna JC, Shardell MD, Santori EK, Nelson TM, Rath JM, Glover ED, et al. The vaginal metabolome and microbiota of cervical HPV-positive and HPV-negative women: a cross-sectional analysis. BJOG. 2020; 127(2):182–92. https://doi.org/10.1111/1471-0528.15981 PMID: 31749298

34. Srinivasan S, Morgan MT, Fiedler TL, Djukovic D, Hoffman NG, Raftery D, et al. Metabolic signatures of bacterial vaginosis. MBio. 2015; 6(2). https://doi.org/10.1128/mBio.00204-15 PMID: 25873373

35. Maceyka M, Spiegel S. Sphingolipid metabolites in inflammatory disease. Nature. 2014; 510(7503):58–67. https://doi.org/10.1038/nature13475 PMID: 24899305

36. Westrich JA, Warren CJ, Pyeon D. Evasion of host immune defenses by human papillomavirus. Virus Res. 2017; 231:21–33. https://doi.org/10.1016/j.virusres.2016.11.023 PMID: 27890631

37. Łaniewski P, Cui H, Roe DJ, Barnes D, Goulder A, Monk BJ, et al. Features of the cervicovaginal microenvironment drive cancer biomarker signatures in patients across cervical carcinogenesis. Sci Rep. 2019; 9(1):7333. https://doi.org/10.1038/s41598-019-43849-5 PMID: 31089160

38. Łaniewski P, Cui H, Roe DJ, Chase DM, Herbst-Kralovetz MM. Vaginal microbiota, genital inflammation and neoplasia impact immune checkpoint protein profiles in the cervicovaginal microenvironment. NPJ Precis Oncol. 2020; 4(22).https://doi.org/10.1038/s41698-020-0126-x PMID: 32802959

39. Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, Badri MH, et al. Learning representations of microbe-metabolite interactions. Nat Methods. 2019; 16(12):1306–14. https://doi.org/10.1038/s41592-019-0616-3 PMID: 31686038

40. Shannon B, Yi TJ, Perusini S, Gajer P, Ma B, Humphrys MS, et al. Association of HPV infection and clearance with cervicovaginal immunology and the vaginal microbiota. Mucosal Immunol. 2017; 10 (5):1310–9. https://doi.org/10.1038/mi.2016.129 PMID: 28120845

41. Castle PE, Hillier SL, Rabe LK, Hildesheim A, Herrero R, Bratti MC, et al. An association of cervical inflammation with high-grade cervical neoplasia in women infected with oncogenic human papillomavirus (HPV). Cancer Epidemiol Biomarkers Prev. 2001; 10(10):1021–7. PMID: 11588127

42. Mhatre M, McAndrew T, Carpenter C, Burk RD, Einstein MH, Herold BC. Cervical intraepithelial neoplasia is associated with genital tract mucosal inflammation. Sex Transm Dis. 2012; 39(8):591–7. https://doi.org/10.1097/OLQ.0b013e318255aeef PMID: 22801340

43. Kriek JM, Jaumdally SZ, Masson L, Little F, Mbulawa Z, Gumbi PP, et al. Female genital tract inflammation, HIV co-infection and persistent mucosal Human Papillomavirus (HPV) infections. Virology. 2016; 493:247–54. https://doi.org/10.1016/j.virol.2016.03.022 PMID: 27065342

44. Drolet M, Benard E, Perez N, Brisson M, Group HPVVIS. Population-level impact and herd effects following the introduction of human papillomavirus vaccination programmes: updated systematic review and meta-analysis. Lancet. 2019; 394(10197):497–509. https://doi.org/10.1016/S0140-6736(19)30298-3 PMID: 31255301

45. Nelson TM, Borgogna JC, Michalek RD, Roberts DW, Rath JM, Glover ED, et al. Cigarette smoking is associated with an altered vaginal tract metabolomic profile. Sci Rep. 2018; 8(1):852. https://doi.org/10.1038/s41598-017-14943-3 PMID: 29339821

46. Masson L, Arnold KB, Little F, Mlisana K, Lewis DA, Mkhize N, et al. Inflammatory cytokine biomarkers to identify women with asymptomatic sexually transmitted infections and bacterial vaginosis who are at high risk of HIV infection. Sex Transm Infect. 2016; 92(3):186–93. https://doi.org/10.1136/sextrans-2015-052072 PMID: 26511781

47. Lennard K, Dabee S, Barnabas SL, Havyarimana E, Blakney A, Jaumdally SZ, et al. Microbial composition predicts genital tract inflammation and persistent bacterial vaginosis in South African adolescent females. Infect Immun. 2018; 86(1). https://doi.org/10.1128/IAI.00410-17 PMID: 29038128

48. Le V, Quinn TP, Tran T, Venkatesh S. Deep in the Bowel: Highly Interpretable Neural Encoder-Decoder Networks Predict Gut Metabolites from Gut Microbiome. BMC Genomics. 2020; 21(Suppl 4):256. https://doi.org/10.1186/s12864-020-6652-7 PMID: 32689932

49. Mallick H, Franzosa EA, McLver LJ, Banerjee S, Sirota-Madi A, Kostic AD, et al. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nat Commun. 2019; 10(1):3136. https://doi.org/10.1038/s41467-019-10927-1 PMID: 31316056

50. Larsen PE, Dai Y. Metabolome of human gut microbiome is predictive of host dysbiosis. Gigascience. 2015; 4:42. https://doi.org/10.1186/s13742-015-0084-3 PMID: 26380076

51. Visconti A, Le Roy CI, Rosa F, Rossi N, Martin TC, Mohney RP, et al. Interplay between the human gut microbiome and host metabolism. Nat Commun. 2019; 10(1):4505. https://doi.org/10.1038/s41467-019-12476-z PMID: 31582752

52. Gomez-Larrauri A, Presa N, Dominguez-Herrera A, Ouro A, Trueba M, Gomez-Munoz A. Role of bioactive sphingolipids in physiology and pathology. Essays Biochem. 2020. https://doi.org/10.1042/EBC20190091 PMID: 32579188

53. Gajer P, Brotman RM, Bai G, Sakamoto J, Schutte UM, Zhong X, et al. Temporal dynamics of the human vaginal microbiota. Sci Transl Med. 2012; 4(132):132ra52. https://doi.org/10.1126/scitranslmed.3003605 PMID: 22553250

54. Witkin SS, Mendes-Soares H, Linhares IM, Jayaram A, Ledger WJ, Forney LJ. Influence of vaginal bacteria and D- and L-lactic acid isomers on vaginal extracellular matrix metalloproteinase inducer: implications for protection against upper genital tract infections. MBio. 2013; 4(4). https://doi.org/10.1128/mBio.00460-13 PMID: 23919998

55. Edwards VL, Smith SB, McComb EJ, Tamarelle J, Ma B, Humphrys MS, et al. The cervicovaginal microbiota-host interaction modulates Chlamydia trachomatis infection. mBio. 2019; 10(4). https://doi.org/10.1128/mBio.01548-19 PMID: 31409678

56. Petrova MI, Reid G, Vaneechoutte M, Lebeer S. Lactobacillus iners: Friend or Foe? Trends Microbiol. 2017; 25(3):182–91. https://doi.org/10.1016/j.tim.2016.11.007 PMID: 27914761

57. Troisi J, Sarno L, Landolfi A, Scala G, Martinelli P, Venturella R, et al. Metabolomic Signature of Endometrial Cancer. J Proteome Res. 2018; 17(2):804–12. https://doi.org/10.1021/acs.jproteome.7b00503 PMID: 29235868

58. Vettukattil R, Hetland TE, Florenes VA, Kaern J, Davidson B, Bathen TF. Proton magnetic resonance metabolomic characterization of ovarian serous carcinoma effusions: chemotherapy-related effects and comparison with malignant mesothelioma and breast carcinoma. Hum Pathol. 2013; 44(9):1859–66. https://doi.org/10.1016/j.humpath.2013.02.009 PMID: 23656974

59. Hilvo M, de Santiago I, Gopalacharyulu P, Schmitt WD, Budczies J, Kuhberg M, et al. Accumulated Metabolites of Hydroxybutyric Acid Serve as Diagnostic and Prognostic Biomarkers of Ovarian High-Grade Serous Carcinomas. Cancer Res. 2016; 76(4):796–804. https://doi.org/10.1158/0008-5472.CAN-15-2298 PMID: 26685161

60. Kovachev S. Defence factors of vaginal lactobacilli. Crit Rev Microbiol. 2018; 44(1):31–9. https://doi.org/10.1080/1040841X.2017.1306688 PMID: 28418713

61. McGroarty JA, Tomeczek L, Pond DG, Reid G, Bruce AW. Hydrogen peroxide production by Lactobacillus species: correlation with susceptibility to the spermicidal compound nonoxynol-9. J Infect Dis. 1992; 165(6):1142–4. https://doi.org/10.1093/infdis/165.6.1142 PMID: 1316413

62. Hertelendy J, Reumuth G, Simons D, Stoppe C, Kim BS, Stromps JP, et al. Macrophage migration inhibitory factor—a favorable marker in inflammatory diseases? Curr Med Chem. 2018; 25(5):601–5. https://doi.org/10.2174/0929867324666170714114200 PMID: 28714387

63. Campisciano G, Zanotta N, Licastro D, De Seta F, Comar M. In vivo microbiome and associated immune markers: New insights into the pathogenesis of vaginal dysbiosis. Sci Rep. 2018; 8(1):2307. https://doi.org/10.1038/s41598-018-20649-x PMID: 29396486

64. Dabee S, Barnabas SL, Lennard KS, Jaumdally SZ, Gamieldien H, Balle C, et al. Defining characteristics of genital health in South African adolescent girls and young women at high risk for HIV infection. PLoS One. 2019; 14(4):e0213975. https://doi.org/10.1371/journal.pone.0213975 PMID: 30947260

65. Cheng RJ, Deng WG, Niu CB, Li YY, Fu Y. Expression of macrophage migration inhibitory factor and CD74 in cervical squamous cell carcinoma. Int J Gynecol Cancer. 2011; 21(6):1004–12. https://doi.org/10.1097/IGC.0b013e31821c45b7 PMID: 21792010

66. Guo P, Wang J, Liu J, Xia M, Li W, He M. Macrophage immigration inhibitory factor promotes cell proliferation and inhibits apoptosis of cervical adenocarcinoma. Tumour Biol. 2015; 36(7):5095–102. https://doi.org/10.1007/s13277-015-3161-4 PMID: 25716200

67. Krockenberger M, Engel JB, Kolb J, Dombrowsky Y, Hausler SF, Kohrenhagen N, et al. Macrophage migration inhibitory factor expression in cervical cancer. J Cancer Res Clin Oncol. 2010; 136(5):651–7. https://doi.org/10.1007/s00432-009-0702-5 PMID: 19915866

68. Kaehler BD, Bokulich NA, McDonald D, Knight R, Caporaso JG, Huttley GA. Species abundance information improves sequence taxonomy classification accuracy. Nat Commun. 2019; 10(1):4643. https://doi.org/10.1038/s41467-019-12669-6 PMID: 31604942

69. Fettweis JM, Serrano MG, Sheth NU, Mayer CM, Glascock AL, Brooks JP, et al. Species-level classification of the vaginal microbiome. BMC Genomics. 2012; 13 Suppl 8:S17. https://doi.org/10.1186/1471-2164-13-S8-S17 PMID: 23282177

70. Robeson MS 2nd, O'Rourke DR, Kaehler BD, Ziemski M, Dillon MR, Foster JT, et al. RESCRIPt: Reproducible sequence taxonomy reference database management. PLoS Comput Biol. 2021; 17(11): e1009581. https://doi.org/10.1371/journal.pcbi.1009581 PMID: 34748542

71. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019; 37(8):852–7. https://doi.org/10.1038/s41587-019-0209-9 PMID: 31341288

72. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016; 13(7):581–3. https://doi.org/10.1038/nmeth.3869 PMID: 27214047

73. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018; 6(1):90. https://doi.org/10.1186/s40168-018-0470-z PMID: 29773078

74. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012; 6(3):610–8. https://doi.org/10.1038/ismej.2011.139 PMID: 22134646

75. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018; 36(10):996–1004. https://doi.org/10.1038/nbt.4229 PMID: 30148503

76. Quinn TP, Erb I, Gloor G, Notredame C, Richardson MF, Crowley TM. A field guide for the compositional analysis of any-omics data. Gigascience. 2019; 8(9). https://doi.org/10.1093/gigascience/giz107 PMID: 31544212

77. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. 2011; 35(2):343–59. https://doi.org/10.1111/j.1574-6976.2010.00251.x PMID: 21039646

78. Bokulich NA, Dillon MR, Bolyen E, Kaehler BD, Huttley GA, Caporaso JG. q2-sample-classifier: machine-learning tools for microbiome classification and regression. J Open Res Softw. 2018; 3(30). https://doi.org/10.21105/joss.00934 PMID: 31552137

79. Breiman L. Random Forests. Machine Learning. 2001; 45:5–32.

80. Hastie T, Tibshirani R, Friedman J. Boosting and Additive Trees. The Elements of Statistical Learning Springer Series in Statistics. New York, NY: Springer; 2009. p. 337–87.

81. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. 2007; 9 (3):90–5.