Microbial function and genital inflammation in young South African women at high risk of HIV infection

Arghavan Alisoltani[1,2], Monalisa T. Manhanzva[1], Matthys Potgieter[3,4], Christina Balle[5], Liam Bell[6], Elizabeth Ross[6], Arash Iranzadeh[3], Michelle du Plessis[6], Nina Radzey[1], Zac McDonald[6], Bridget Calder[4], Imane Allali[3,7], Nicola Mulder[3,8,9], Smritee Dabee[1,10], Shaun Barnabas[1], Hoyam Gamieldien[1], Adam Godzik[2], Jonathan M. Blackburn[4,8], David L. Tabb[8,11,12], Linda-Gail Bekker[8,13], Heather Jaspan[5,8,10], Jo-Ann S. Passmore[1,8,14,15], Lindi Masson[1,8,14,16*]

[1]Division of Medical Virology, Department of Pathology, University of Cape Town, South Africa; [2]Division of Biomedical Sciences, University of California Riverside School of Medicine, Riverside, CA, USA; [3]Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, South Africa; [4]Division of Chemical and Systems Biology, Department of Integrative Biomedical Sciences, University of Cape Town, South Africa. [5]Division of Immunology, Department of Pathology, University of Cape Town, South Africa; [6]Centre for Proteomic and Genomic Research, South Africa; [7]Laboratory of Human Pathologies Biology, Department of Biology and Genomic Center of Human Pathologies, Mohammed V University in Rabat, Morocco; [8]Institute of Infectious Disease and Molecular Medicine (IDM), University of Cape Town, South Africa; [9]Centre for Infectious Diseases Research (CIDRI) in Africa Wellcome Trust Centre, University of Cape Town, South Africa; [10]Seattle Children's Research Institute, University of Washington, USA; [11]Bioinformatics Unit, South African Tuberculosis Bioinformatics Initiative, Stellenbosch University, South Africa; [12]DST–NRF Centre of Excellence for Biomedical Tuberculosis Research, Stellenbosch University, South Africa; [13]Desmond Tutu HIV Centre, University of Cape Town, South Africa; [14]Centre for the AIDS Programme of Research in South Africa, South Africa; [15]National Health Laboratory Service, South Africa; [16]Disease Elimination Program, Life Sciences Discipline, Burnet Institute, Melbourne, Australia

*Address correspondence to: Dr Lindi Masson; 85 Commercial Road, Melbourne, Victoria, Australia 3004; GPO Box 2284, Melbourne, Victoria, Australia 3001; Email: Lindi.Masson@burnet.edu.au

ABSTRACT

Female genital tract (FGT) inflammation is an important risk factor for HIV acquisition. Previous studies have identified the FGT microbiota as a major determinant of altered inflammatory cytokine profiles, however, the relative importance of microbial activities and functions is unknown. Liquid chromatography-tandem mass spectrometry was used to evaluate the FGT metaproteomic profiles of 113 young South African women. FGT inflammation was defined by inflammatory cytokine concentration profile. A total of 3,186 microbial and human proteins were identified and both were analyzed for the functional and the former also for taxonomic distributions. The functional annotations of bacterial proteins predicted genital inflammation more accurately than bacterial relative abundance determined by 16S rRNA analysis ($p<0.0001$). The majority of microbial biological pathways were underrepresented in women with inflammation, including a signature of reduced cell wall organization and peptidoglycan biosynthesis. This signature remained associated with FGT inflammatory profiles after adjusting for *Lactobacillus* relative abundance and potential confounders. These biological pathways were also associated with inflammatory cytokine responses to *Lactobacillus* isolates from the same women in a vaginal epithelial cell model. Comparison of metaproteomics taxonomic assignment with 16S rRNA gene survey analysis revealed differences in bacterial relative abundance between the methods, however both approaches showed increased non-optimal bacteria and decreased lactobacilli in women with FGT inflammation. These findings suggest that microbial function is a critical determinant of FGT inflammation, and that specific microbial properties may be harnessed for biotherapeutic development.

MAIN

Despite the large reduction in AIDS-related deaths as a result of the expansion of HIV antiretroviral programs, global HIV incidence has declined by only 16% since 2010[1]. Of particular concern are the extremely high rates of HIV infection in young South African women[1]. The behavioral and biological factors underlying this increased risk are not fully understood, however, a key predisposing factor that has been identified in this population is genital inflammation[2-4]. We have previously shown that women with elevated concentrations of pro-inflammatory cytokines and chemokines in their genital tracts were at higher risk of becoming infected with HIV[3]. Furthermore, the protection offered by a topical antiretroviral tenofovir microbicide was compromised in women with evidence of genital inflammation[4].

The mechanisms by which female genital tract (FGT) inflammation increases HIV risk are likely multifactorial, and genital inflammation may facilitate the establishment of a productive HIV infection by recruiting and activating HIV target cells, directly promoting HIV transcription and reducing epithelial barrier integrity[5-7]. Utilizing proteomics, Arnold, *et al.* showed that proteins associated with tissue remodeling processes were up-regulated in the FGT, while protein biomarkers of mucosal barrier integrity were down-regulated in individuals with an inflammatory profile[7]. This suggests that inflammation increases tissue remodeling at the expense of mucosal barrier integrity, which may in turn increase HIV acquisition risk[7]. Furthermore, neutrophil-associated proteins, especially certain proteases, were positively associated with inflammation and may be involved in disrupting the mucosal barrier[7]. This theory was further validated by the finding that antiproteases associated with HIV resistance were increased in a cohort of sex workers from Kenya[8].

Bacterial vaginosis (BV), non-optimal cervicovaginal bacteria, and sexually transmitted infections (STIs), which are highly prevalent in South African women, are likely important drivers of genital inflammation in this population[9,10]. BV and non-optimal bacteria have been consistently associated with a marked increase in pro-inflammatory cytokine concentrations[9-12]. Conversely, women who have “optimal” vaginal microbiota, primarily consisting of *Lactobacillus* spp., have low levels of genital inflammation and a reduced risk of acquiring HIV[13]. Lactobacilli and the lactic acid that they produce may actively suppress inflammatory responses and thus may play a significant role in modulating

immune profiles in the FGT[14-16]. However, partly due to the complexity and diversity of the microbiome, the immunomodulatory mechanisms of specific vaginal bacterial species are not fully understood[17]. Further adding to this complexity is the fact that substantial differences in the cervicovaginal microbiota exist by geographical location and ethnicity and that the properties of different strains within particular microbial species are also highly variable[16,18].

In addition to providing insight into host mucosal barrier function, metaproteomic studies of the FGT can evaluate microbial activities and functions that may influence inflammatory profiles[19]. A recent study highlighted the importance of vaginal microbial function by demonstrating that FGT bacteria, such as *Gardnerella vaginalis,* modulate the efficacy of the tenofovir microbicide by active metabolism of the drug[20]. Additionally, Zevin et al. found that bacterial functional profiles were associated with epithelial barrier integrity and wound healing in the FGT[21], suggesting that microbial function may also be closely linked to genital inflammation. The aim of the present study was to utilize both metaproteomics and 16S rRNA amplicon sequencing to improve our understanding of the relationship between microbial function and inflammatory profiles in the FGTs of South African women.

RESULTS

This study included 113 young and adolescent HIV-uninfected women (aged 16-22 years) residing in Cape Town, South Africa[22]. STI and BV prevalence were high in this cohort, with 48/113 (42%) of the women having at least one STI and 56/111 (50%) of the women being BV positive by Nugent scoring (**Supplementary Table 1**). The use of injectable contraceptives, BV and chlamydia were associated with genital inflammation, as previously described[10,23].

FGT metaproteome associates with BV and inflammatory cytokine profiles

Liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis was conducted on the total protein acquired from lateral vaginal wall swabs for all participants. Using principal component analysis (PCA), it was found that women with BV clustered separately from women who were BV negative according to the relative abundance of all proteins identified (**Figure 1a**). Women who were considered to have high levels of genital inflammation based on pro-inflammatory cytokine concentrations (**Supplementary Figure 1**) tended to cluster separately from women with low inflammation, while no clear grouping was observed by STI status or chemokine profiles (**Figure 1b-d**). Similarly, BV was associated with the largest number of differentially abundant proteins, followed by inflammatory cytokine profiles, while few associations were observed between chemokine profiles and protein relative abundance and none between STI status and protein relative abundance (**Figure 1e**).

Comparing taxonomic assignment using metaproteomics and 16S rRNA amplicon sequencing

Of the proteins identified in lateral vaginal wall swab samples, 38.8% were human, 55.8% were bacterial, 2.7% fungal, 0.4% archaeal, 0.09% of viral origin and 2% grouped as “other” (not shown). However, as the majority of taxa identified had $\leq$2 proteins detected, a more stringent cut-off was applied to include only taxa with $\geq$3 detected proteins, or 2 proteins detected in multiple taxa. The final curated dataset included 44% human, 55% bacterial (n=81 taxa) and 1% fungal proteins (n=13 taxa) (**Supplementary Figure 2; Supplementary Table 2**). When the relative abundance of the most abundant bacterial taxa identified using metaproteomics was compared to the relative abundance of the most abundant bacteria identified using 16S rRNA amplicon sequencing, a large degree of similarity was found at genus level. A total of 6/9 of the genera identified using 16S rRNA analysis

were also identified using metaproteomics (**Figure 2; Supplementary Figure 3**). As expected, both approaches showed that multiple *Lactobacillus* species were less abundant in women with inflammation, while BV-associated bacteria (including *G. vaginalis*, *Prevotella* species, *Megasphaera* species, *Sneathia amnii* and *Atopobium vaginae*) were more abundant in women with inflammation compared to women without inflammation (**Supplementary Table 3)**. However, species level annotation differed between the 16S rRNA and metaproteomics analyses, with metaproteomics identifying more lactobacilli at the species level, while *Lachnovaginosum* genomospecies (previously known as BVAB1) was not identified using metaproteomics. Additionally, the relative abundance of the taxa differed between 16S rRNA and metaproteomics analyses.

Differences in protein profiles between women defined as having high, medium or low FGT inflammation

A total of 449, 165 and 39 proteins were significantly differentially abundant between women with high versus low inflammation, medium versus high inflammation and medium versus low inflammation, respectively (**Supplementary Table 4**; **Figure 3**). Microbial proteins that were more abundant in women with inflammation were mostly assigned to non-optimal bacterial taxa (including *G. vaginalis*, *Prevotella* species, *Megasphaera* species, *S. amnii* and *A. vaginae*). Less abundant microbial proteins in women with inflammation were primarily of *Lactobacillus* origin (**Figure 3**). Of the human proteins that were differentially abundant according to inflammation status, 93 were less abundant and 132 were more abundant in women with high compared to those with low inflammation. Human biological process gene ontologies (GOs) that were significantly underrepresented in women with inflammation included positive regulation of apoptotic signaling, establishment of endothelial intestinal barrier, ectoderm development and cornification [false discovery rate (FDR) adj. $p<0.0001$ for all; **Supplementary Figure 4**]. Significantly overrepresented pathways in women with inflammation included multiple inflammatory processes - such as chronic response to antigenic stimulus, positive regulation of IL-6 production and inflammatory response (FDR adj. $p<0.0001$ for all; **Supplementary Figure 4**).

Microbial function predicts genital inflammation with greater accuracy than taxa relative abundance

The accuracies of (i) microbial relative abundance (determined using 16S rRNA sequencing), (ii) functional prediction based on 16S rRNA data, (iii) microbial protein relative abundance (determined using metaproteomics) and (iv) microbial molecular function (determined by aggregation of GOs of proteins identified using metaproteomics) for prediction of genital inflammation status were evaluated using random forest analysis. The three most important species predicting genital inflammation by 16S rRNA gene data were *Sneathia sanguinegens*, *A. vaginae,* and *Lactobacillus iners* (**Figure 4a**). The top functional pathways based on 16S rRNA data for the identification of women with inflammation were tyrosine metabolism, biofilm formation, and prolactin signaling (**Figure 4b**). Similarly, the most important proteins predicting inflammation by metaproteomics included two *L. iners* proteins (chaperone protein DnaK and pyrophosphate phospho-hydrolase), two *A. vaginae* proteins (glyceraldehyde-3-phosphate dehydrogenase and phosphoglycerate kinase), *Megasphaera* genomosp. type 1 cold-shock DNA-binding domain protein, and *L. crispatus* ATP synthase subunit alpha (**Figure 4c**). The top molecular functions associated with genital inflammation included oxidoreductase activity (of multiple proteins expressed primarily by lactobacilli, but also *Prevotella* and *Atopobium* species), beta-phosphoglucomutase activity (expressed by lactobacilli) and GTPase activity (of multiple proteins produced by mainly *Prevotella, G. vaginalis*, and some lactobacilli) as illustrated in **Figure 4d**. These findings suggest that the functions of key BV-associated bacteria and lactobacilli are critical for determining the level of genital inflammation. It was also found that the out-of-bag (OOB) error rate distribution was significantly lower for molecular function, followed by protein relative abundance, functional prediction based on16S rRNA data and then taxa relative abundance (**Figure 4e**). This suggests that the activities of the bacteria present in the FGT play a critical role in driving genital inflammation, in addition to the presence and abundance of particular taxa.

Differences in microbial stress response, lactate dehydrogenase and metabolic pathways between women with high versus low FGT inflammation

Microbial functional profiles were further investigated by comparing molecular functions, biological processes and cellular component GO enrichment for detected proteins between women with low versus high inflammation using differential expression analysis, adjusting for potentially confounding

variables including age, contraceptives, prostate specific antigen (PSA) and STIs (**Figure 5; Supplementary Table 5**). The majority of microbial molecular functions (90/107), biological processes (57/83) and cellular components (14/23) were underrepresented in women with genital inflammation, with most of the differentially abundant metabolic processes, including peptide, nucleoside, glutamine, and glycerophospholipid metabolic processes, decreased in women with inflammation. The phosphoenolpyruvate-dependent sugar phosphotransferase system (a major microbial carbohydrate uptake system including proteins assigned to *Lactobacillus* species, *S. amniii*, *A. vaginae,* and *G. vaginalis*) was also underrepresented in women with evidence of genital inflammation compared to women without inflammation (adj. p<0.0001; **Figure 5a**). Additionally, L-lactate dehydrogenase activity was inversely associated with inflammation (**Supplementary Table 5**) and both large and small ribosomal subunit cellular components were underrepresented in those with FGT inflammation (**Figure 5b**). The only metabolic processes that were overabundant in women with inflammation were malate and ATP metabolic processes. Overrepresented microbial biological processes in women with genital inflammation also included response to oxidative stress and cellular components associated with cell division (e.g. mitotic spindle midzone) and ubiquitination (e.g. Doa10p ubiquitin ligase complex, Hrd1p ubiquitin ligase ERAD-L complex, RQC complex; **Figure 5)**.

Underrepresented *Lactobacillus* peptidoglycan, cell wall and membrane pathways in women with FGT inflammation

Among the top biological processes inversely associated with inflammation were cell wall organization, regulation of cell shape and peptidoglycan biosynthetic process (all adj. p=<0.0001; **Figure 5a**). The GO terms, regulation of cell shape, and peptidoglycan biosynthetic process included identical proteins (e.g. D-alanylalanine synthetase, D-alanyl-D-alanine-adding enzyme and Bifunctional protein GlmU), while cell wall organization included the same proteins plus D-alanyl carrier protein. Although proteins involved in these processes were only detected for *Lactobacillus* species and may thus be biomarkers of the increased relative abundance of lactobacilli, each remained significantly associated with inflammation after adjusting for the relative abundance of *L. iners* and non-*iners Lactobacillus* species (determined using 16S rRNA gene sequencing), as well as STI status, PSA detection and hormonal contraceptive use [beta-coefficient: -1.68; 95% confidence interval (CI): -2.84 to -0.52; p=0.004 for cell wall organization; beta-coefficient: -1.71; 95% CI: -2.91 to

-0.53; p=0.005 for both regulation of cell shape and peptidoglycan biosynthetic process]. This suggests that the relationship between these processes and inflammation is independent of the relative abundance of *Lactobacillus* species and that the cell wall and membrane properties of lactobacilli may play a role in modulating FGT inflammation. Furthermore, we found that multiple cell wall and membrane cellular components, including the S-layer, peptidoglycan-based cell wall, cell wall and membrane, were similarly underrepresented in women with FGT inflammation (**Figure 5b**).

In vitro **confirmation of the role of** ***Lactobacillus*** **cell wall and membrane pathways in regulating genital inflammation**

In order to further understand the relationships between *Lactobacillus* cell wall and membrane properties and inflammation, 64 *Lactobacillus* species were isolated from the same women and their *in vitro* inflammatory profiles characterized as previously described[24]. From these lactobacilli, the 22 isolates that induced the lowest levels of pro-inflammatory cytokine production by vaginal epithelial cells and the 22 isolates that induced the greatest cytokine responses were selected and their protein profiles evaluated using LC-MS/MS (**Supplementary Figure 5**). Cell wall organization, regulation of cell shape and peptidoglycan biosynthetic process that were associated with genital inflammation in the FGT metaproteomic analysis of vaginal swab samples were also significantly overrepresented in isolates that induced low levels of inflammation *in vitro* (**Figure 6**). Furthermore, plasma membrane, cell wall and membrane cellular components were similarly overrepresented in isolates that induced low levels of inflammation (**Figure 6**).

DISCUSSION

Understanding biological risk factors for HIV acquisition is critical for the development of effective HIV prevention strategies. FGT inflammation, defined by elevated inflammatory cytokine levels, has been identified as a key factor associated with HIV infection risk. However, the causes and underlying mechanisms are not fully understood. In this study, we demonstrate that, while the presence of particular microorganisms, including non-optimal bacteria and *Lactobacillus* species, is important for modulating FGT inflammation, their activities and functions are likely of greater importance. We showed that the molecular functions of bacterial proteins predicted FGT inflammation status more accurately than taxa relative abundance determined using 16S rRNA amplicon sequencing, as well as functional predictions based on 16S rRNA analysis. The majority of microbial biological processes were underrepresented in women with inflammation (57/83), including metabolic pathways and a strong signature of reduced cell wall organization and peptidoglycan biosynthesis. This signature remained associated with FGT inflammation after adjusting for the relative abundance of lactobacilli, as well as other potential confounders (such as hormone contraceptive use, semen exposure, STIs), and was also associated with inflammatory cytokine induction in vaginal epithelial cells in response to *Lactobacillus* isolates *in vitro*.

While we found a large degree of similarity between metaproteomics and 16S rRNA amplicon sequence taxonomic assignment determined for the same women[25], important differences in the predicted relative abundance of most bacteria were observed. Differences in the relative abundance of certain taxa observed using metaproteomics and 16S rRNA gene data is not surprising as taxonomic assignment from metaproteomic approaches also indicates functional activity (measured by the amount of expressed protein), rather than simply the abundance of a particular bacterial taxa. Additionally, others have noted that 16S rRNA amplicon sequencing does not have the ability to distinguish between live bacteria and transient DNA[26]. We also observed differences that are likely due to database limitations. For example, *Lachnovaginosum* genomospecies (previously known as BVAB1) was not identified in the metaproteomics analysis. This is due to the fact that this species was not present in the UniProt or NCBI databases[27]. However, the *Clostridium* and *Ruminococcus* species which were identified are likely to be *Lachnovaginosum* genomospecies since these taxa fall into the same Clostridiales order. Additionally, *Clostridium* and *Ruminococcus* species were more

abundant in women with high inflammation, as expected for *Lachnovaginosum* genomospecies[10]. Metaproteomic analysis was able to identify lactobacilli to species level, while sequencing of the V4 region of the 16S rRNA gene had more limited resolution for this genus, as expected.

The finding that fungal protein relative abundance was substantially lower than bacterial protein relative abundance is not surprising as it has been estimated that the total number of fungal cells is orders of magnitude lower than the number of bacterial cells in the human body[28]. Although *Candida* proteins were detected, the curated dataset (following removal of taxa that had only 1 protein detected or 2 proteins detected in only a single sample) did not include any *Candida* species. This may be due to the relative rarity of fungal cells in these samples and it would be interesting to evaluate alternative methodology for sample processing to better resolve this population in the future[29]. A limitation of microbial functional analysis using metaproteomics is that, at present, there is generally sparse population of functional information in the databases for the majority of the microbiome. Thus, the findings of this study will be biased toward the microbes that are better annotated. Nonetheless, microbial function was closely associated with genital inflammatory profiles and it is possible that this relationship may be even stronger should better annotation exist. To enable researchers to further study associations between microbial composition and function with BV, STIs, and FGT inflammation by conducting detailed analysis of specific proteins, taxa and GO terms of interest, we have developed the "FGT-METAP" online application. The application is available at (http://immunodb.org/FGTMetap/) and allows users to repeat the analyses presented here, as well as obtain additional information by analyzing individual proteins, species, and pathways.

A significant functional signature associated with FGT inflammation included cell wall, peptidoglycan and cell membrane biological processes and cellular components. Although these functions were linked exclusively to *Lactobacillus* species, the associations with FGT inflammation were upheld after adjusting for *Lactobacillus* relative abundance determined by 16S rRNA amplicon sequencing. This profile was confirmed when these GOs were compared between *Lactobacillus* isolates that induced high versus low levels of cytokine production *in vitro*. Previous studies have suggested that the peptidoglycan structure, the proteins present in the cell wall, as well as the cell membrane, may influence the immunomodulatory properties of *Lactobacillus* species[14,30]. In a murine model,

administration of peptidoglycan extracted from gut lactobacilli was able to rescue the mice from colitis[31]. This activity was dependent on the NOD2 pathway and correlated with an upregulation of the indoleamine 2,3-dioxygenase immunosuppressive pathway. The immunomodulatory properties of peptidoglycan were furthermore dependent on the strain of lactobacilli from which the peptidoglycan was isolated[31]. In a previous study, we showed that adhesion of lactobacilli to vaginal epithelial cells *in vitro* was inversely associated with cytokine responses, suggesting that direct interaction between the isolates and vaginal epithelial cells is critical for immunoregulation[24]. In addition to cell wall and membrane properties, L-lactate dehydrogenase activity was identified as a molecular function inversely associated with inflammation. Similarly, D-lactate dehydrogenase relative abundance and D-lactate production by the *Lactobacillus* isolates included in this study were also associated with inflammatory profiles in an *in vitro* analysis[24]. These findings support the results of previous studies showing that lactic acid has immunoregulatory properties[15]. Taken together, these findings suggest that lactobacilli may modulate the inflammatory environment in the FGT through multiple mechanisms.

In this study, we observed a similar host proteome profile associated with FGT inflammation compared to previous studies[7]. Multiple inflammatory pathways were overrepresented in women with FGT inflammation, while signatures of reduced barrier function were observed, with underabundant endothelial, ectoderm and tight junction biological processes. These findings similarly suggest that genital inflammation may result in reduced epithelial barrier integrity.

The link between FGT microbial function and local inflammatory responses described in this study suggests that both the presence of specific microbial taxa in the FGT and their properties and activities likely play a critical role in modulating inflammation. Currently the annotation of microbial functions is sparse, but with ever increasing amounts of high quality high-throughput data, the available information will improve steadily. The findings of the present study contribute to our understanding of the mechanisms by which the microbiota may influence local immunity, and in turn alter the risk of HIV infection. Additionally, the analyses described herein identify specific microbial properties that may be harnessed for biotherapeutic development.

METHODS

Participants

This study included sexually experienced HIV-negative adolescent girls and young women (16-22 years) from the Women's Initiative in Sexual Health (WISH) study in Cape Town, South Africa[22]. While the parent study cohort comprised 149 women, the present sub-study included 113 women who had both vaginal swabs available for metaproteomics analysis and menstrual cup (MC) cervicovaginal secretions available for cytokine profiling. The University of Cape Town Faculty of Health Sciences Human Research Ethics Committee (HREC REF: 267/2013) approved this study. Women ≥18 years provided written informed consent, those <18 years provided informed assent and informed consent was obtained from parents/guardians.

Definition of genital inflammation based on cytokine concentrations

For cytokine measurement, MCs (Softcup®, Evofem Inc, USA) were inserted by the clinician and kept in place for an hour, after which they were transported to the lab at 4ºC and processed within 4 hours of removal from the participants. MCs were centrifuged and the cervicovaginal secretions were re-suspended in phosphate buffered saline (PBS) at a ratio of 1ml of mucus: 4 ml of PBS and stored at -80ºC until cytokine measurement. Prior to cytokine measurement, MC secretions were pre-filtered using 0.2 µm cellulose acetate filters (Sigma-Aldrich, MO, USA). The concentrations of 48 cytokines were measured in MC samples using Luminex (Bio-Rad Laboratories Inc®, CA, USA)[23]. K-means clustering was used to identify women with low, medium and high pro-inflammatory cytokine [interleukin (IL)-1α, IL-1β, IL-6, IL-12p40, IL-12p70, tumor necrosis factor (TNF)-α, TNF-β, TNF-related apoptosis-inducing ligand (TRAIL), interferon (IFN)-γ] and chemokine profiles [cutaneous T-cell-attracting chemokine (CTACK), eotaxin, growth regulated oncogene (GRO)-α, IL-8, IL-16, IFN-γ-induced protein (IP)-10, monocyte chemoattractant protein (MCP)-1, MCP-3, monokine induced by IFN-γ (MIG), macrophage inflammatory protein (MIP)-1α, MIP-1β, regulated on activation, normal T cell expressed and secreted (RANTES)] (**Supplementary Figure 1**).

STI and BV diagnosis and evaluation of semen contamination

Vulvovaginal swab samples were screened for common STIs, including *Chlamydia trachomatis*, *Neisseria gonorrhoeae*, *Trichomonas vaginalis*, *Mycoplasma genitalium,* herpes simplex virus (HSV)

types 1 and 2, *Treponema pallidum* and *Haemophilus ducreyi*, using real-time multiplex PCR assays[22]. Lateral vaginal wall swabs were collected for BV assessment by Nugent scoring. Blood was collected for HIV rapid testing (Alere Determine™ HIV-1/2 Ag/Ab Combo, Alere, USA). PSA was measured in lateral vaginal wall swabs using Human Kallikrein 3/PSA Quantikine ELISA kits (R&D Systems, MN, USA).

Discovery metaproteomics and analysis

For shotgun LC-MS/MS, lateral vaginal wall swabs were collected, placed in 1 ml PBS, transported to the laboratory at 4°C and stored at -80°C. Samples were computationally randomized for processing and analysis. The stored swabs were thawed overnight at 4°C before each swab sample was vortexed for 30 seconds, the mucus scraped off the swab on the side of the tube and vortexed for an additional 10 seconds. The samples were then clarified by centrifugation and the protein content of each supernatant was determined using the Quanti-Pro bicinchoninic acid (BCA) assay kit (Sigma-Aldrich, MO, USA). Equal protein amounts (100 µg) were denatured with urea exchange buffer, filtered, reduced with dithiothreitol, alkylated with iodoacetamide and washed with hydroxyethyl piperazineethanesulfonic acid (HEPES). Acetone precipitation/formic acid (FA; Sigma-Aldrich, MO, USA) solubilisation was added as a further sample clean-up procedure. The samples were then incubated overnight with trypsin (Promega, WI, USA) and the peptides were eluted with HEPES and dried via vacuum centrifugation. Reversed-phase liquid chromatography was used for desalting using a step-function gradient. The eluted fractions were dried via vacuum centrifugation and kept at -80°C until analysis. LC-MS/MS analysis was conducted on a Q-Exactive quadrupole-Orbitrap MS (Thermo Fisher Scientific, MA, USA) coupled with a Dionex UltiMate 3000 nano-UPLC system (120min per sample). The solvent system included Solvent A: 0.1% Formic Acid (FA) in LC grade water (Burdick and Jackson, NJ, USA); and Solvent B: 0.1% FA in Acetonitrile (ACN; Burdick & Jackson, NJ, USA). The MS was operated in positive ion mode with a capillary temperature of 320°C, and 1.95 kV electrospray voltage was applied. All 113 samples were evaluated in 11 batches over a period of 3 months. A quality control consisting of pooled lateral vaginal swab eluants from the study participants was run at least twice with every batch. Preliminary quality control analysis was conducted, and flagged samples/batches were rerun.

The MetaNovo pipeline was used to filter the UniProt (August 2017 release) database for protein identification with default settings[32]. Raw files were then processed with MaxQuant version 1.5.7.4. The detailed parameters are provided in **Supplementary Table 6**. In brief, methionine oxidation and acetylation of the protein N-terminal amino acid were considered as variable modifications and carbamidomethyl (C) as a fixed modification. The digestion enzyme was trypsin with a maximum of two missed cleavages. For quality control analysis, raw protein intensities (**Supplementary Figure 6a**) and $\log_{10}$-transformed iBAQ intensities (**Supplementary Figure 6b**) of the quality control pool, as well as the raw protein intensities of the clinical samples (**Supplementary Figure 6c**), were compared between batches. As variation in raw intensity was observed, all data was adjusted for batch number prior to downstream analysis. The cleaned and transformed iBAQ values were used to identify significantly differentially abundant proteins according to inflammatory cytokine profiles, BV status or STIs using the limma R package[33]. The p-values were obtained from the moderated t-test after adjusting for confounding variables including age, contraceptives, PSA and STIs. Proteins with FDR adjusted p-values ≤ 0.05 and $\log_2$-transformed fold change ≥ 1.2 or ≤ -1.2 were considered as significantly differentially abundant. To define significantly differentially abundant taxa and GOs, aggregated iBAQ values of proteins of the same species or GO IDs were compared between women with low, medium and high inflammation using the limma R package. The aggregation was preformed separately for host and microbial proteins. The p-values along with GO IDs were uploaded to REVIGO (http://revigo.irb.hr/) to visualize top 50 pathways with the lowest p-values.

The effects of BV, STIs and chemokine and pro-inflammatory cytokine profiles on the metaproteome were investigated by PCA using the mixOmics R package[34]. The ComplexHeatmap package[35] was used to generate heatmaps and cluster the samples and metaproteomes based on the hierarchical and k-means clustering methods. Additionally, R packages EnhancedVolcano, and FlipPlots were applied for plotting volcano and sankey diagrams, respectively. Basic R functions and the ggplot2 R package were used for data manipulation, transformation, normalization and generation of graphics. To identify the key factors distinguishing women defined as having low, medium and high inflammation in their FGTs, we used random forest analysis using the R package randomForest[36] to

validate and compare the importance of variables obtained in decision trees using the following settings: (i) the type of random forest was classification, (ii) the number of trees was 2000, and (iii) the number of variables evaluated at each split was one-third of the total number of variables [proteins/molecular function GO IDs/16S based operational taxonomic units (OTUs)]. To compare the accuracy of the protein-, function- and species-based models, we iterated each random forest model 100 times for each of the input datasets separately. The distribution of the OOBs for 100 models for each data type were then compared using t-tests.

16S rRNA gene sequencing and analysis

Bacterial 16S rRNA gene sequencing and analysis was conducted as previously described[25]. Briefly, the V4 hypervariable region of the 16S rRNA gene was amplified using modified universal primers[37]. Duplicate samples were pooled, purified using Agencourt AMPure XP beads (Beckman Coulter, CA, USA) and quantified using the Qubit dsDNA HS Assay (Life Technologies, CA, USA). Illumina sequencing adapters and dual-index barcodes were added to the purified amplicon products using limited cycle PCR and the Nextera XT Index Kit (Illumina, CA, USA). Amplicons from 96 samples and controls were pooled in equimolar amounts and the resultant libraries were purified by gel extraction and quantified (Qiagen, Germany). The libraries were sequenced on an Illumina MiSeq platform (300 bp paired-end with v3 chemistry). Following de-multiplexing, raw reads were preprocessed, merged and filtered using DADA2[38]. Primer sequences were removed using a custom python script and the reads truncated at 250 bp. Taxonomic annotation was based on a customized version of SILVA reference database. Only samples with $\geq$ 2000 reads were included in further analyses. The representative sequences for each amplicon sequence variant (ASV) were BLAST searched against NCBI non-redundant protein and 16S databases for further validation, as well as enrichment of the taxonomic classifications. Microbial functional predictions were performed using Piphillin server[39] against the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (99% identity cutoff). The relative abundance of functional pathways (KEGG level 3) was used for RF analysis.

***Lactobacillus* isolation**

Lactobacilli were isolated from cervicovaginal secretions collected using MCs by culturing in de Man Rogosa and Sharpe (MRS) broth for 48 hours at 37°C under anaerobic conditions. The cultures were

streaked onto MRS agar plates under the same culture conditions, single colonies were picked and then pre-screened microscopically by Gram staining. Matrix Assisted Laser Desorption Ionization Time of Flight (MALDI-TOF) was conducted to identify the bacteria to species level. A total of 115 lactobacilli were isolated and stored at -80°C in a final concentration of 60% glycerol. From these, 64 isolates from 25 of the study participants were selected for evaluation of inflammatory profiles.

Vaginal epithelial cell stimulation and measurement of cytokine concentrations

As described previously[24], vaginal epithelial cells (VK2/E6E7 ATCC® CRL-2616™) were maintained in complete keratinocyte serum free media (KSFM) supplemented with 0.4 mM calcium chloride, 0.05 mg/ml of bovine pituitary extract, 0.1 ng/ml human recombinant epithelial growth factor and 50 U/ml penicillin and 50 U/ml streptomycin (Sigma-Aldrich, MO, USA). The VK2 cells were seeded into 24-well tissue culture plates, incubated at 37°C in the presence of 5% carbon dioxide and grown to confluency. Lactobacilli, adjusted to 4.18 x 10^6 colony forming units (CFU)/ml in antibiotic-free KSFM, were used to stimulate VK2 cells for 24 hours at 37°C in the presence of 5% carbon dioxide. Production of IL-6, IL-8, IL-1α, IL-1β, IP-10, MIP-3α, MIP-1α, MIP-1β and IL-1RA were measured using a Magnetic Luminex Screening Assay kit (R&D Systems, MN, USA) and a Bio-Plex™ Suspension Array Reader (Bio-Rad Laboratories Inc®, CA, USA). VK2 cell viability following bacterial stimulation was confirmed using the Trypan blue exclusion assay. PCA was used to compare the overall inflammatory profiles of the isolates and to select the 22 most inflammatory and 22 least inflammatory isolates for characterization using proteomics (**Supplementary Figure 5**).

Characterization of *Lactobacillus* protein expression *in vitro* using mass spectrometry

Forty-four *Lactobacillus* isolates were adjusted to 4.18 x 10^6 CFU/ml in MRS and incubated for 24 hours under anaerobic conditions. Following incubation, the cultures were centrifuged and the pellets washed 3x with PBS. Protein was extracted by resuspending the pellets in 100 mM triethylammonium bicarbonate (TEAB; Sigma-Aldrich, MO, USA) 4% sodium dodecyl sulfate (SDS; Sigma-Aldrich, MO, USA). Samples were sonicated in a sonicating water bath and subsequently incubated at 95°C for 10min. Nucleic acids were degraded using benzonase nuclease (Sigma-Aldrich, MO, USA) and samples were clarified by centrifugation at 10 000xg for 10 min. Quantification was performed using the Quanti-Pro BCA assay kit (Sigma-Aldrich, MO, USA). HILIC beads (ReSyn Biosciences, South

Africa) were washed with 250 µl wash buffer [15% ACN, 100 mM Ammonium acetate (Sigma-Aldrich, MO, USA) pH 4.5]. The beads were then resuspended in loading buffer (30% ACN, 200mM Ammonium acetate pH 4.5). A total of 50 µg of protein from each sample was transferred to a protein LoBind plate (Merck, NJ, USA). Protein was reduced with tris (2-carboxyethyl) phosphine (Sigma-Aldrich, MO, USA) and alkylated with methylmethanethiosulphonate (MMTS; Sigma-Aldrich, MO, USA). HILIC magnetic beads were added at an equal volume to that of the sample and a ratio of 5:1 total protein and incubated on the shaker at 900 rpm for 30 min. After binding, the beads were washed four times with 95% ACN. Protein was digested by incubation with trypsin for four hours and the supernatant containing peptides was removed and dried down in a vacuum centrifuge. LC-MS/MS analysis was conducted with a Q-Exactive quadrupole-Orbitrap MS as described above. Raw files were processed with MaxQuant version 1.5.7.4 against a database including the *Lactobacillus* genus and common contaminants. Log_{10}-transformed iBAQ intensities were compared using Mann-Whitney U test and p-values were adjusted for multiple comparisons using an FDR step-down procedure.

FGT-METAP application

We have developed FGT-METAP (Female Genital Tract Metaproteomics), an online application for exploring and mining the FGT microbial composition and function of South African women at high risk of HIV infection. The app is available at (http://immunodb.org/FGTMetap/) and allows users to repeat the analyses presented here, as well as perform additional analyses and investigation using the metaproteomic data generated in this study.

ACKNOWLEDGEMENTS

We would like to acknowledge the participants of the Women's Initiative in Health Study. This work was supported by the Carnegie Corporation of New York (PI: L. Masson), South African National Research Foundation (PI: L. Masson), the Poliomyelitis Research Foundation (PI: Masson), and the South African Medical Research Council (PI: L. Masson). The WISH cohort was supported by the European and Developing Countries Clinical Trials Partnership Strategic Primer grant (SP.2011.41304.038; PI J.S. Passmore) and the South African Department of Science and Technology (DST/CON 0260/2012; PI J.S. Passmore). J.M.B. thanks the NRF for a SARChI Chair. D.L.T. was supported by the South African Tuberculosis Bioinformatics Initiative (SATBBI), a Strategic Health Innovation Partnership grant from the South African Medical Research Council and South African Department of Science and Technology.

AUTHOR CONTRIBUTIONS

A.A. analyzed the data, developed the FGT-METAP application, and wrote the manuscript; M.T.M, C.B., S.D., and S.B. conducted some of the laboratory experiments, analyzed the data and contributed to manuscript preparation; M.P., A.I., N.R., I.A., N.M. B.C. and D.L.T. analyzed the data and contributed to manuscript preparation; L.B., E.W., M.P, Z.M., H.G. conducted some of the laboratory experiments and contributed to manuscript preparation; N.M. and J.M.B. supervised the data analysis and contributed to manuscript preparation; A.G supervised the development of FGT-METAP and hosting the application and contributed to manuscript preparation; L.G.B. managed the clinical site for the WISH study, collected some of the clinical data and contributed to manuscript preparation; H.J. supervised the data collection, analyzed the data and contributed to manuscript preparation; J.S.P. was Principal Investigator of the WISH cohort, supervised the collection of some of the data and contributed to manuscript preparation; L.M. conceptualized the study, supervised the data collection, analyzed some of the data and wrote the manuscript.

COMPETING INTERESTS STATEMENT

The authors have no competing interests to declare.

REFERENCES:

1. UNAIDS Data 2019. Available at: https://www.unaids.org/sites/default/files/media_asset/2019-UNAIDS-data_en.pdf. Accessed 20/02/20.
2. Morrison, C., et al. Cervical inflammation and immunity associated with hormonal contraception, pregnancy, and HIV-1 seroconversion. *Journal of Acquired Immune Deficiency Syndromes* **66**, 109-117 (2014).
3. Masson, L., et al. Genital inflammation and the risk of HIV acquisition in women. *Clinical Infectious Diseases* **61**, 260-269 (2015).
4. McKinnon, L.R., et al. Genital inflammation undermines the effectiveness of tenofovir gel in preventing HIV acquisition in women. *Nature Medicine* **24**, 491 (2018).
5. Osborn, L., Kunkel, S. & Nabel, G.J. Tumor necrosis factor-alpha and interleukin-1 stimulate the human immunodeficiency virus enhancer by activation of the nuclear factor kappa B. *Proceedings of the National Academy of Sciences* 86, 2336–240 (1989).
6. Li, Q., et al. Glycerol monolaurate prevents mucosal SIV transmission. *Nature* **458**, 1034-1038 (2009)
7. Arnold, K.B., et al. Increased levels of inflammatory cytokines in the female reproductive tract are associated with altered expression of proteases, mucosal barrier proteins, and an influx of HIV-susceptible target cells. *Mucosal Immunology* **9**, 194 (2016)
8. Burgener, A., et al. Comprehensive proteomic study identifies serpin and cystatin antiproteases as novel correlates of HIV-1 resistance in the cervicovaginal mucosa of female sex workers. *Journal of Proteome Research* **10**, 5139-49 (2011)
9. Masson, L., et al. Defining genital tract cytokine signatures of sexually transmitted infections and bacterial vaginosis in women at high risk of HIV infection: a cross-sectional study. *Sexually Transmitted Infections* **90**, 580-587 (2014)
10. Lennard, K., et al. Microbial composition predicts genital tract inflammation and persistent bacterial vaginosis in South African adolescent females. *Infection and Immunity* **86**, e00410-17 (2018).
11. Anahtar, M.N., et al. Cervicovaginal bacteria are a major modulator of host inflammatory responses in the female genital tract. *Immunity* **42**, 965-976 (2015).

12. Deese, J., et al. Injectable progestin-only contraception is associated with increased levels of pro-inflammatory cytokines in the female genital tract. *American Journal of Reproductive Immunology* **74**, 357-367 (2015).
13. Gosmann, C., et al. *Lactobacillus*-deficient cervicovaginal bacterial communities are associated with increased HIV acquisition in young South African women. *Immunity* **17**, 29-37 (2017).
14. Chapot-Chartier, M.P. & Kulakauskas, S. Cell wall structure and function in lactic acid bacteria. *Microbial Cell Factories* **13**, S9 (2014).
15. Hearps, A.C., et al. Vaginal lactic acid elicits an anti-inflammatory response from human cervicovaginal epithelial cells and inhibits production of pro-inflammatory mediators associated with HIV acquisition. *Mucosal Immunology* **10**, 1480 (2017).
16. Chetwin, E., et al. Antimicrobial and inflammatory properties of South African clinical *Lactobacillus* isolates and vaginal probiotics. *Scientific Reports* **13**, 1917 (2019).
17. McKinnon, L.R., et al. The evolving facets of bacterial vaginosis: implications for HIV transmission. *AIDS Research and Human Retroviruses* **35**, 219-228 (2019).
18. Ravel, J., et al. Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences* **108**, 4680-7 (2011).
19. Xiong, W., Abraham, P.E., Li, Z., Pan, C. & Hettich, R.L. Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. *Proteomics* **15**, 3424-38 (2015).
20. Klatt, N.R., et al. Vaginal bacteria modify HIV tenofovir microbicide efficacy in African women. *Science* **356**, 938-45 (2017).
21. Zevin, A.S., et al. Microbiome composition and function drives wound-healing impairment in the female genital tract. *PLoS Pathogens* **12**, e1005889 (2016).
22. Barnabas, S. L., *et al.* Converging epidemics of sexually transmitted infections and bacterial vaginosis in southern African female adolescents at risk of HIV. *International Journal of STD & AIDS* **29**, 531–539 (2017).
23. Dabee, S., et al. Defining characteristics of genital health in South African adolescent girls and young women at high risk for HIV infection. *PloS One* **14**, e0213975 (2019).
24. Manhanzva, M.T., et al. Inflammatory and antimicrobial properties differ between vaginal

Lactobacillus isolates from South African women with non-optimal versus optimal microbiota. *Scientific Reports*. Accepted

25. Balle, C., et al. Endocervical and vaginal microbiota in South African adolescents with asymptomatic *Chlamydia trachomatis* infection. *Scientific Reports* **8**, 11109 (2018).
26. Lagier, J.C., et al. Culturing the human microbiota and culturomics. *Nature Reviews Microbiology* **16**, 540-550 (2018).
27. Holm, J.B., et al. Comparative metagenome-assembled genome analysis of *Lachnovaginosum genomospecies*, formerly known as BVAB1. *BioRxiv* 1, 657197 (2019).
28. Huffnagle, G.B. & Noverr, M.C. The emerging world of the fungal microbiome. *Trends in Microbiology* **21**, 334-41 (2013).
29. Bianco, L. & Perrotta, G. Methodologies and perspectives of proteomics applied to filamentous fungi: from sample preparation to secretome analysis. *International Journal of Molecular Sciences* **16**, 5803-29 (2015).
30. Macho Fernandez, E., Pot, B. & Grangette, C. Beneficial effect of probiotics in IBD: are peptidogycan and NOD2 the molecular key effectors? *Gut Microbes* **2**, 280-6 (2011).
31. Fernandez, E.M., et al. Anti-inflammatory capacity of selected lactobacilli in experimental colitis is driven by NOD2-mediated recognition of a specific peptidoglycan-derived muropeptide. *Gut* **60**, 1050-9 (2011).
32. Potgieter, M.G., et al. MetaNovo: a probabilistic approach to peptide and polymorphism discovery in complex mass spectrometry datasets. *BioRxiv* **1**, 605550 (2019).
33. Smyth, G., et al. Limma: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor. *Statistics for Biology and Health* 397-420 (2005).
34. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.A. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLoS Computational Biology* **13**, e1005752 (2017).
35. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-9 (2016).
36. Liaw, A. & Wiener, M. “Classification and Regression by randomForest.” *R News* **2**, 18–22 (2002)

37. Pearce, M.M., et al. The female urinary microbiome: a comparison of women with and without urgency urinary incontinence. *MBio* **5**, e01283–14 (2014).
38. Callahan, B. J., et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**, 581-583 (2016)
39. Narayan, N.R., et al. Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. *BMC genomics* **21**, 1-12 (2020).

FIGURE LEGENDS

Figure 1 Clustering of women according to overall metaproteomic profiles. Liquid chromatography-tandem mass spectrometry was used to evaluate the metaproteome in lateral vaginal wall swabs from 113 women from Cape Town, South Africa. Principal component analysis (mixOmics R package) was used to group women based on the log_2-transformed intensity-based absolute quantification (iBAQ) intensities of all proteins identified. Grouping by (**a**) bacterial vaginosis (BV) status, (**b**) sexually transmitted infection (STI) status, (**c**) vaginal pro-inflammatory cytokine concentrations, and (**d**) vaginal chemokine concentrations are shown. Each point represents an individual woman. (**e**) Statistically significant differences in the relative abundance of proteins between BV positive and negative women, women with high versus low chemokine or pro-inflammatory cytokine concentrations, and STI positive versus negative women were determined using the limma package in R. The number of statistically significant differentially abundant proteins per comparison is indicated by circle size, with the taxonomic annotation of the proteins shown on the left size of the graph. PC: Principal component.

Figure 2 Bacterial relative abundance determined using metaproteomics versus 16S rRNA amplicon sequencing by inflammation cytokine profile. Liquid chromatography tandem mass spectrometry was used to evaluate the metaproteome in lateral vaginal wall swabs from women from Cape Town, South Africa. For 16S analysis, the V4 region was amplified and libraries sequenced on an Illumina MiSeq platform. The relative abundance of the top 20 most abundant bacterial taxa identified using (**a**) 16S rRNA amplicon and (**b**) metaproteomics is shown for all participants for whom both 16S rRNA data and metaproteomics data were generated (n=74). OTU: Operational taxonomic unit.

Figure 3 Significantly differentially abundant proteins in women with high versus low inflammatory cytokine profiles. Unsupervised hierarchical clustering of log_2-transformed intensity-based absolute quantification (iBAQ) values for significantly differentially abundant proteins followed by K-means clustering of both rows and columns was conducted using the ComplexHeatmap R package. The list of significantly differentially abundant proteins between women with low (n=23), medium (n=57) and high (n=33) inflammation was obtained using the limma R package. The p-values

were calculated using the moderated t-test after adjusting for confounding variables including age, contraceptives, semen exposure and sexually transmitted infections. Proteins with false discovery rate (FDR) adjusted p-values <0.05 and log_2-transformed fold change $\geq$ 1.2 or $\leq$ -1.2 were considered as significantly differentially abundant and are shown in the heatmap. The bar graph shows log_2-transformed fold changes (LOGFC), with red indicating positive and blue indicating negative fold changes in women with high versus low inflammation. The species annotation and the biological process gene ontology of the proteins are shown on the right of the heatmap. Adjusted p-values are shown by dots, with orange dots indicating FDR adjusted p-values (<0.005). BV: Bacterial vaginosis. NA: Not applicable. BP: Biological Process.

Figure 4 Comparison of bacterial, bacterial protein and bacterial function relative abundance for prediction of genital inflammatory cytokine profiles. Random forest analysis was used to evaluate the accuracy of (**a**) bacterial relative abundance (determined using 16S rRNA amplicon sequencing; n=74), (**b**) bacterial functional predictions based on 16 rRNA data (n=74), (**c**) bacterial protein relative abundance (determined using metaproteomics; n=74) and (**d**) bacterial protein molecular function relative abundance (determined by metaproteomics and aggregation of protein values assigned to a same gene ontology term; n=74) for determining the presence of genital inflammation. The bars and numbers within the bars indicate the importance of each taxon, protein or function based on the Mean Decrease in Gini Value. The size of bars in each panel differs based on the length of the labels (**e**) Each random forest model was iterated 100 times for each of the input datasets separately and the distribution of the out-of-bag error rates for the 100 models were then compared using t-tests. OTU: Operational taxonomic unit.

Figure 5 Microbial biological process and cellular component gene ontologies associated with genital inflammatory cytokine profiles. Unsupervised hierarchical clustering of aggregated intensity-based absolute quantification (iBAQ) values for proteins assigned to the same biological process (**a**) or cellular component (**b**) gene ontology (GO). The heatmaps show aggregated GO relative abundance and the bar graphs show fold changes in aggregated log_2-transformed iBAQ values (LOGFC) for each microbial protein with the same biological process (**a**) or cellular component (**b**) GO. Red bars indicate positive and blue bars indicate negative fold changes in women with high

versus low inflammation. False discovery rate adjusted p-values are shown by dots, with red dots indicating low p-values. Red arrows indicate cell wall and membrane processes and components. BV: Bacterial vaginosis. Proinflam cyt: Pro-inflammatory cytokine

Figure 6 Relative abundance of biological process and cellular component gene ontologies between inflammatory versus non-inflammatory *Lactobacillus* isolates *in vitro*. Twenty-two *Lactobacillus* isolates that induced relatively high inflammatory responses and 22 isolates that induced lower inflammatory responses were adjusted to 4.18 x 10^6 CFU/ml in bacterial culture medium and incubated for 24 hours under anaerobic conditions. Following incubation, protein was extracted, digested and liquid chromatography tandem mass spectrometry was conducted. Raw files were processed with MaxQuant against a database including the *Lactobacillus* genus and common contaminants. The intensity-based absolute quantification (iBAQ) values for proteins with the same gene ontologies were aggregated, $\log_{10}$-transformed and compared using Mann-Whitney U test. Box-and-whisker plots show $\log_{10}$-transformed iBAQ values. A false discovery rate step-down procedure was used to adjust for multiple comparisons and adjusted p-values <0.05 were considered statistically significant.

Supplementary Figure 1 Identification of women with low, medium and high levels of pro-inflammatory cytokines and chemokines. The concentrations of 48 cytokines were measured in menstrual cup samples from all study participants (n=113). Of these cytokines, nine were classified as pro-inflammatory [interleukin (IL)-1α, IL-1β, IL-6, IL-12p40, IL-12p70, tumor necrosis factor (TNF)-α, TNF-β, TNF-related apoptosis-inducing ligand (TRAIL), interferon (IFN)-γ] and twelve were classified as chemokines [cutaneous T-cell-attracting chemokine (CTACK), eotaxin, growth regulated oncogene (GRO)-α, IL-8, IL-16, IFN-γ-induced protein (IP)-10, monocyte chemoattractant protein (MCP)-1, MCP-3, monokine induced by IFN-γ (MIG), macrophage inflammatory protein (MIP)-1α, MIP-1β, regulated on activation, normal T cell expressed and secreted (RANTES)]. Unsupervised hierarchical followed by K-means clustering were then used to identify women with low, medium and high pro-inflammatory cytokine or chemokine profiles. The heatmaps show $\log_{10}$-transformed cytokine concentrations.

Supplementary Figure 2 Taxonomic annotation of proteins identified using metaproteomics. Liquid chromatography tandem mass spectrometry was used to evaluate metaproteomic profiles in vaginal swabs from 113 women from Cape Town, South Africa. Proteins were identified using MaxQuant and a custom database generated using *de novo* sequencing to filter the UniProt database. Taxonomy was assigned using Uniprot and relative abundance of each taxon was determined by aggregating the intensity-based absolute quantification (iBAQ) values of all proteins identified for each taxon. (**a**) Eukaryota and bacteria domains are shown. Eukaryota includes the fungi kingdom and metazoan subkingdom, while the bacteria domain includes actinobacteria, firmicutes, fusobacteria, bacteriodetes and gammaproteobacter phyla. (**b**) Distribution of taxa identified is shown as a pie chart. (**C**) Number of proteins detected for genera/species for which the greatest number of proteins were identified. (**D**) Protein relative abundance per genus/species for taxa with the highest relative abundance.

Supplementary Figure 3 Human biological processes, microbial taxa and microbial biological processes in individual women with low, medium and high inflammation. Liquid chromatography-tandem mass spectrometry was used to evaluate the metaproteome in lateral vaginal wall swabs from 113 women from Cape Town, South Africa. Proteins were identified using MaxQuant and a custom database generated using *de novo* sequencing to filter the UniProt database. Gene ontology was determined using UniProt and the intensity-based absolute quantification (iBAQ) values of proteins with the same biological process (BP) gene ontologies were aggregated separately for human proteins and microbial proteins. The top 20 most abundant (**a**) human BPs, (**b**) microbial taxa determined using metaproteomics and (**c**) microbial BPs are shown as stacked bar graphs for individual participants. Participants are grouped according to inflammatory cytokine profiles.

Supplementary Figure 4 Top 50 significantly differentially abundant human biological processes in women with low, and high inflammation. Liquid chromatography-tandem mass spectrometry was used to evaluate the metaproteome in lateral vaginal wall swabs from 113 women from Cape Town, South Africa. Proteins were identified using MaxQuant and a custom database generated using *de novo* sequencing to filter the UniProt database. Gene ontology was determined

using UniProt and the intensity-based absolute quantification (iBAQ) values of human proteins with the same biological process (BP) gene ontologies were aggregated. The significantly differentially abundant pathways were obtained using the moderated t-test implemented in the limma R package. False discovery rate (FDR) adjusted p-values were calculated after controlling for potentially confounding variables including age, contraceptives, semen exposure and sexually transmitted infections. Results were visualized using the REVIGO tool (http://revigo.irb.hr/). (**a**) The top 50 overrepresented human BPs in women with inflammation (FDR p-values <0.05). (**b**) The top 50 underrepresented human BPs in women with inflammation (FDR p-values <0.05).

Supplementary Figure 5 Selection of *Lactobacillus* isolates inducing low versus high inflammatory responses for proteomics analysis. Sixty-four *Lactobacillus* species were isolated from the cervicovaginal secretions of 25 of the study participants. (**a**) Lactobacilli adjusted to 4.18 x 10^6 CFU/ml in antibiotic-free cell culture medium were incubated with vaginal epithelial (VK2) cells for 24 hours before inflammatory cytokine concentrations were quantified using Luminex. Log_{10}-transformed concentrations of cytokines produced by VK2 cells in responses to the isolates are shown as a heatmap. (**b**) Principal component analysis was used to compare the overall inflammatory profiles of the isolates and to select the 22 most inflammatory (coral) and 22 least inflammatory (blue) isolates for liquid chromatography mass spectrometry analysis.

Supplementary Figure 6 Quality control analysis. All 113 lateral vaginal wall samples were evaluated in 11 batches over a period of 3 months. A quality control lateral vaginal wall swab sample pool was evaluated at least twice with every batch. (**a**) The raw intensities of all proteins detected in each pool are shown by dots, with the same proteins connected by lines. (**b**) The log_{10}-transforrmed intensity-based absolute quantification (iBAQ) values of each pool are shown, with lines indicating medians and error bars indicating range. (**c**) Principal component analysis was used to compare the overall raw intensities of each participant sample (n=113), grouped and colored according to batch number.

FIGURE 1

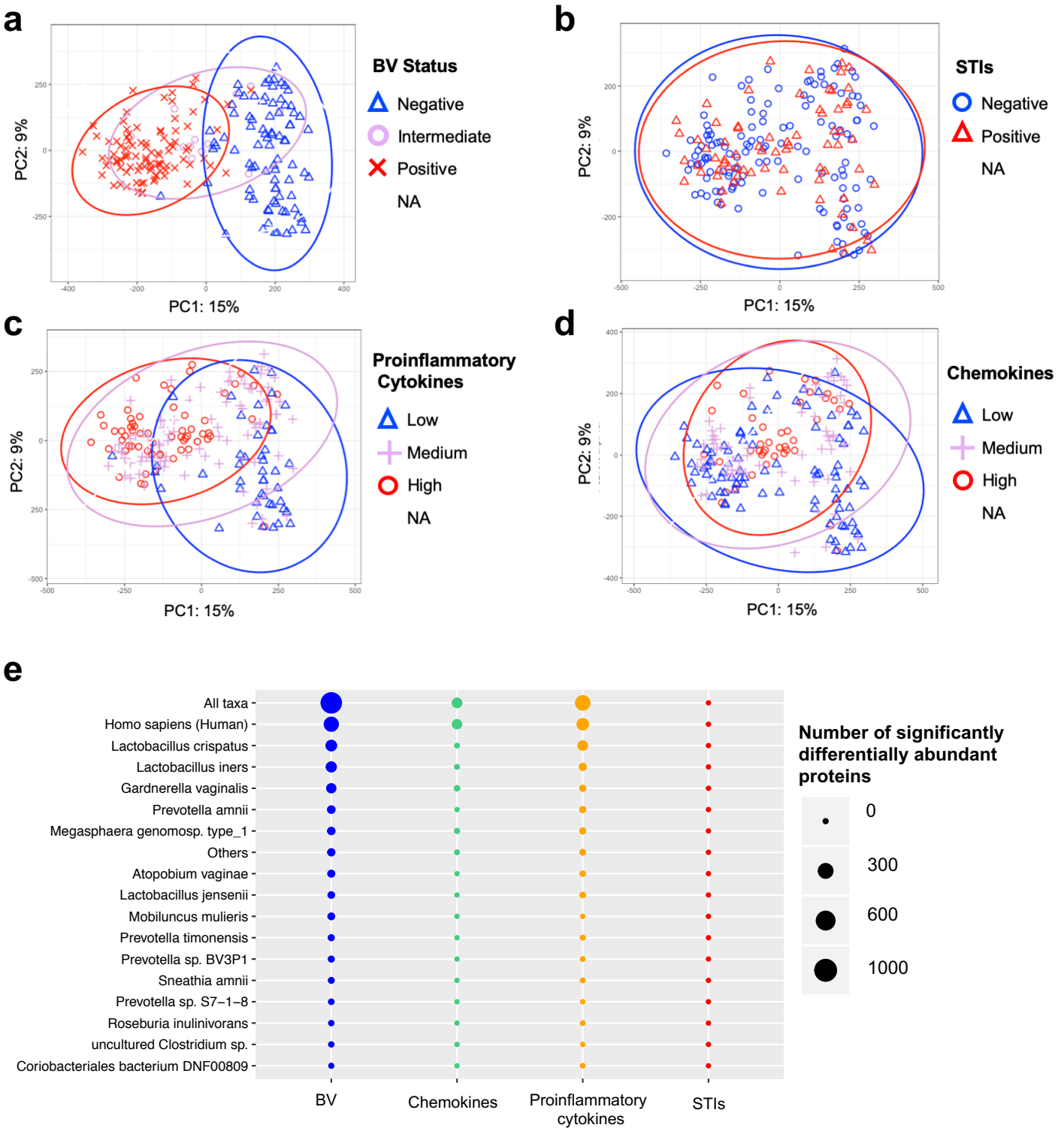

a
PC2: 9%
PC1: 15%
BV Status
Negative
Intermediate
Positive
NA
b
PC2: 9%
PC1: 15%
STIs
Negative
Positive
NA
c
PC2: 9%
PC1: 15%
Proinflammatory Cytokines
Low
Medium
High
NA
d
PC2: 9%
PC1: 15%
Chemokines
Low
Medium
High
NA
e
All taxa
Homo sapiens (Human)
Lactobacillus crispatus
Lactobacillus iners
Gardnerella vaginalis
Prevotella amnii
Megasphaera genomosp. type_1
Others
Atopobium vaginae
Lactobacillus jensenii
Mobiluncus mulieris
Prevotella timonensis
Prevotella sp. BV3P1
Sneathia amnii
Prevotella sp. S7−1−8
Roseburia inulinivorans
uncultured Clostridium sp.
Coriobacteriales bacterium DNF00809
BV
Chemokines
Proinflammatory cytokines
STIs
Number of significantly differentially abundant proteins
0
300
600
1000

FIGURE 2

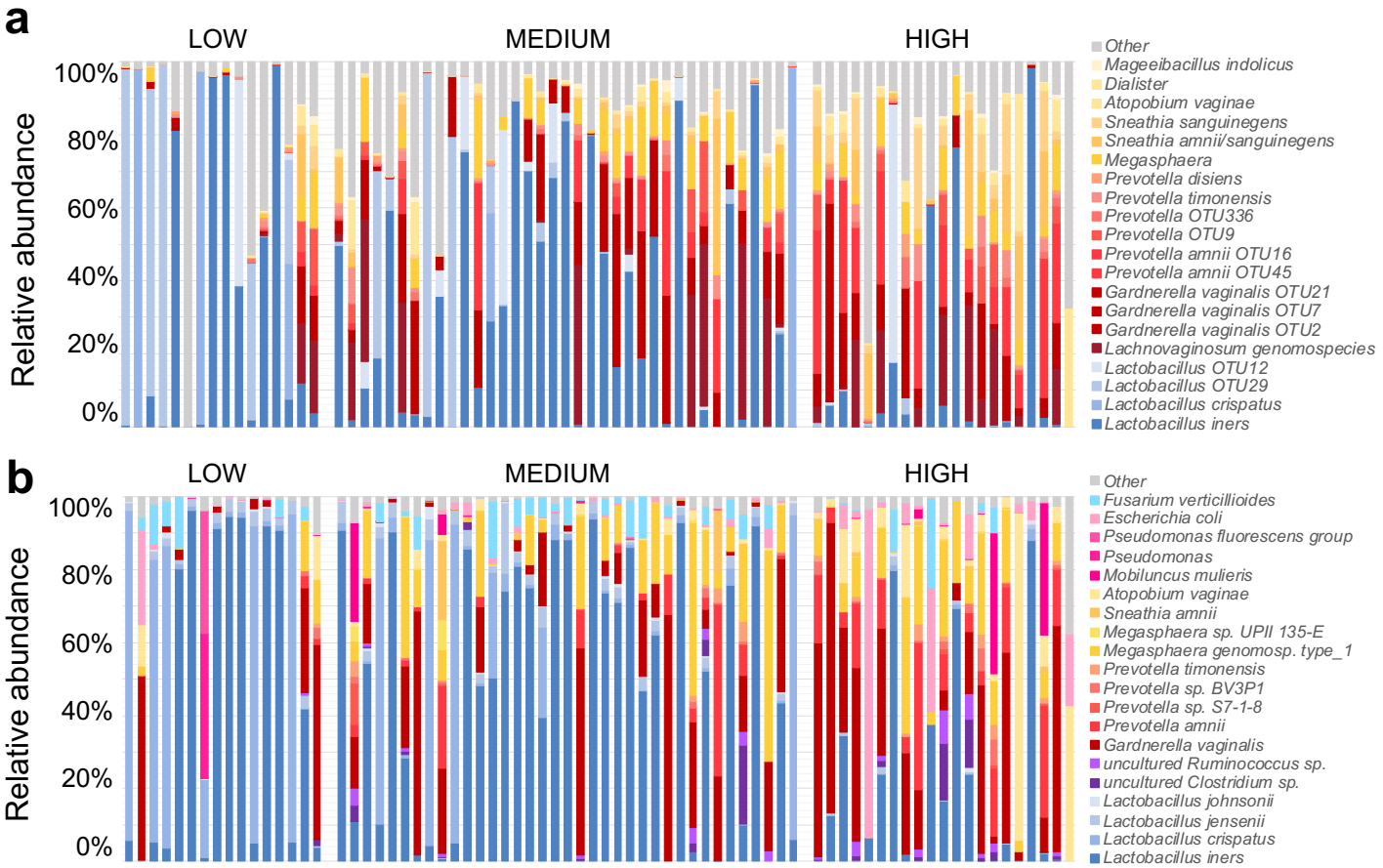

a
LOW
MEDIUM
HIGH
Relative abundance
100%
80%
60%
40%
20%
0%
Other
Mageeibacillus indolicus
Dialister
Atopobium vaginae
Sneathia sanguinegens
Sneathia amnii/sanguinegens
Megasphaera
Prevotella disiens
Prevotella timonensis
Prevotella OTU336
Prevotella OTU9
Prevotella amnii OTU16
Prevotella amnii OTU45
Gardnerella vaginalis OTU21
Gardnerella vaginalis OTU7
Gardnerella vaginalis OTU2
Lachnovaginosum genomospecies
Lactobacillus OTU12
Lactobacillus OTU29
Lactobacillus crispatus
Lactobacillus iners
b
LOW
MEDIUM
HIGH
Relative abundance
100%
80%
60%
40%
20%
0%
Other
Fusarium verticillioides
Escherichia coli
Pseudomonas fluorescens group
Pseudomonas
Mobiluncus mulieris
Atopobium vaginae
Sneathia amnii
Megasphaera sp. UPII 135-E
Megasphaera genomosp. type_1
Prevotella timonensis
Prevotella sp. BV3P1
Prevotella sp. S7-1-8
Prevotella amnii
Gardnerella vaginalis
uncultured Ruminococcus sp.
uncultured Clostridium sp.
Lactobacillus johnsonii
Lactobacillus jensenii
Lactobacillus crispatus
Lactobacillus iners

FIGURE 3

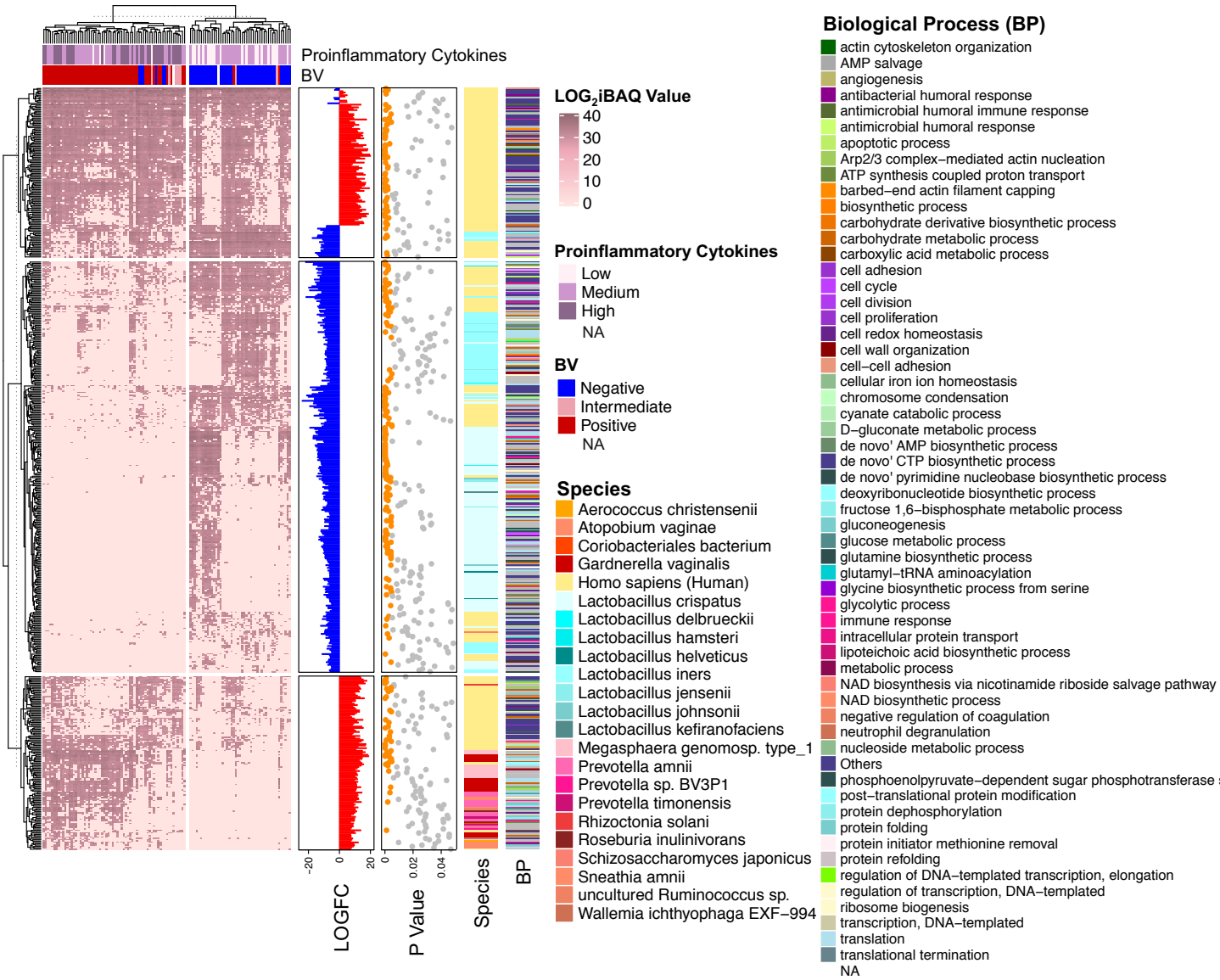

Proinflammatory Cytokines
BV
LOGFC
P Value
Species
BP
LOG2iBAQ Value
40
30
20
10
0
Proinflammatory Cytokines
Low
Medium
High
NA
BV
Negative
Intermediate
Positive
NA
Species
Aerococcus christensenii
Atopobium vaginae
Coriobacteriales bacterium
Gardnerella vaginalis
Homo sapiens (Human)
Lactobacillus crispatus
Lactobacillus delbrueckii
Lactobacillus hamsteri
Lactobacillus helveticus
Lactobacillus iners
Lactobacillus jensenii
Lactobacillus johnsonii
Lactobacillus kefiranofaciens
Megasphaera genomosp. type_1
Prevotella amnii
Prevotella sp. BV3P1
Prevotella timonensis
Rhizoctonia solani
Roseburia inulinivorans
Schizosaccharomyces japonicus
Sneathia amnii
uncultured Ruminococcus sp.
Wallemia ichthyophaga EXF−994
Biological Process (BP)
actin cytoskeleton organization
AMP salvage
angiogenesis
antibacterial humoral response
antimicrobial humoral immune response
antimicrobial humoral response
apoptotic process
Arp2/3 complex−mediated actin nucleation
ATP synthesis coupled proton transport
barbed−end actin filament capping
biosynthetic process
carbohydrate derivative biosynthetic process
carbohydrate metabolic process
carboxylic acid metabolic process
cell adhesion
cell cycle
cell division
cell proliferation
cell redox homeostasis
cell wall organization
cell−cell adhesion
cellular iron ion homeostasis
chromosome condensation
cyanate catabolic process
D−gluconate metabolic process
de novo' AMP biosynthetic process
de novo' CTP biosynthetic process
de novo' pyrimidine nucleobase biosynthetic process
deoxyribonucleotide biosynthetic process
fructose 1,6−bisphosphate metabolic process
gluconeogenesis
glucose metabolic process
glutamine biosynthetic process
glutamyl−tRNA aminoacylation
glycine biosynthetic process from serine
glycolytic process
immune response
intracellular protein transport
lipoteichoic acid biosynthetic process
metabolic process
NAD biosynthesis via nicotinamide riboside salvage pathway
NAD biosynthetic process
negative regulation of coagulation
neutrophil degranulation
nucleoside metabolic process
Others
phosphoenolpyruvate−dependent sugar phosphotransferase
post−translational protein modification
protein dephosphorylation
protein folding
protein initiator methionine removal
protein refolding
regulation of DNA−templated transcription, elongation
regulation of transcription, DNA−templated
ribosome biogenesis
transcription, DNA−templated
translation
translational termination
NA

FIGURE 4

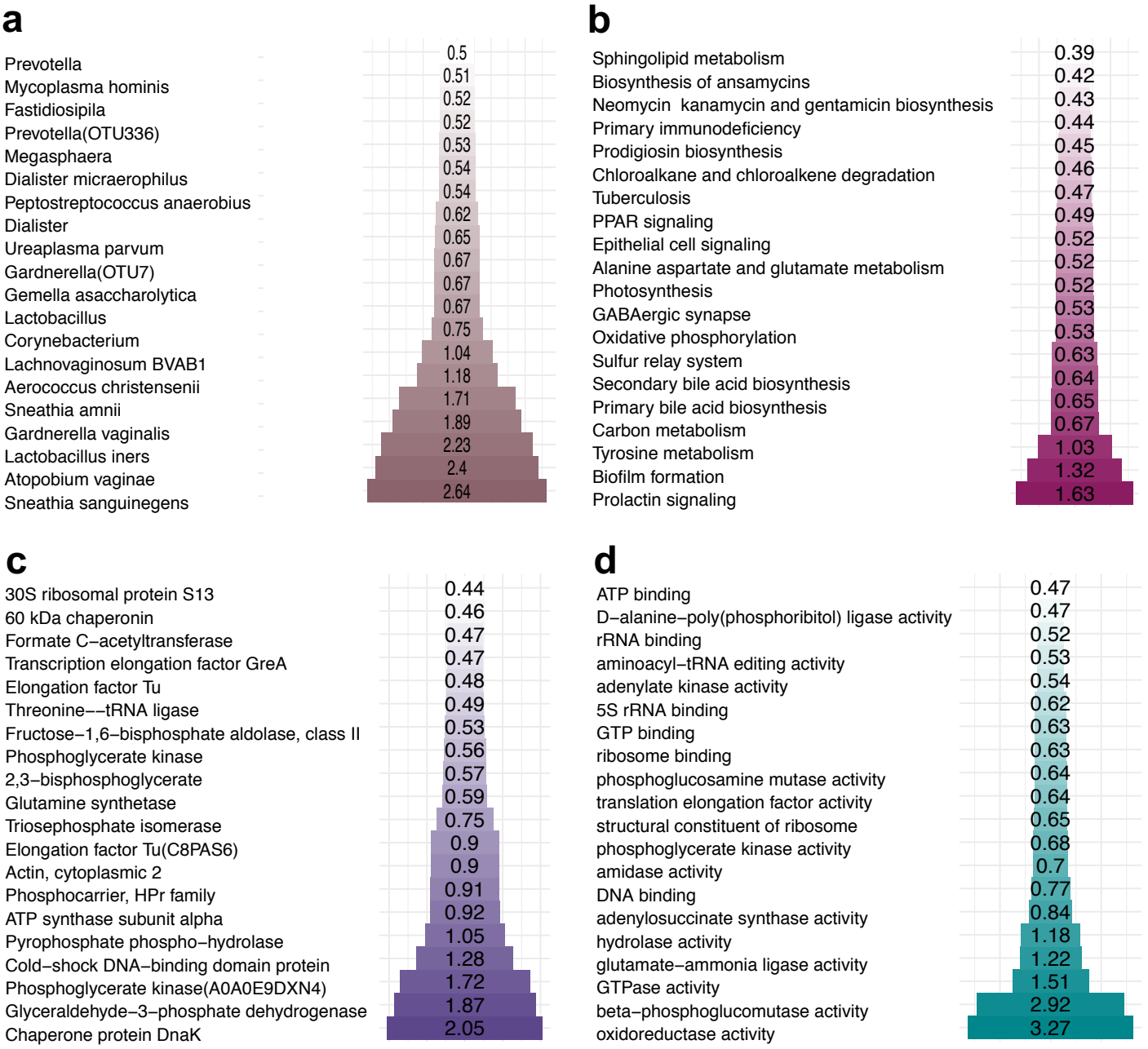

a
Prevotella 0.5
Mycoplasma hominis 0.51
Fastidiosipila 0.52
Prevotella(OTU336) 0.52
Megasphaera 0.53
Dialister micraerophilus 0.54
Peptostreptococcus anaerobius 0.54
Dialister 0.62
Ureaplasma parvum 0.65
Gardnerella(OTU7) 0.67
Gemella asaccharolytica 0.67
Lactobacillus 0.67
Corynebacterium 0.75
Lachnovaginosum BVAB1 1.04
Aerococcus christensenii 1.18
Sneathia amnii 1.71
Gardnerella vaginalis 1.89
Lactobacillus iners 2.23
Atopobium vaginae 2.4
Sneathia sanguinegens 2.64
b
Sphingolipid metabolism 0.39
Biosynthesis of ansamycins 0.42
Neomycin kanamycin and gentamicin biosynthesis 0.43
Primary immunodeficiency 0.44
Prodigiosin biosynthesis 0.45
Chloroalkane and chloroalkene degradation 0.46
Tuberculosis 0.47
PPAR signaling 0.49
Epithelial cell signaling 0.52
Alanine aspartate and glutamate metabolism 0.52
Photosynthesis 0.52
GABAergic synapse 0.53
Oxidative phosphorylation 0.53
Sulfur relay system 0.63
Secondary bile acid biosynthesis 0.64
Primary bile acid biosynthesis 0.65
Carbon metabolism 0.67
Tyrosine metabolism 1.03
Biofilm formation 1.32
Prolactin signaling 1.63
c
30S ribosomal protein S13 0.44
60 kDa chaperonin 0.46
Formate C−acetyltransferase 0.47
Transcription elongation factor GreA 0.47
Elongation factor Tu 0.48
Threonine−−tRNA ligase 0.49
Fructose−1,6−bisphosphate aldolase, class II 0.53
Phosphoglycerate kinase 0.56
2,3−bisphosphoglycerate 0.57
Glutamine synthetase 0.59
Triosephosphate isomerase 0.75
Elongation factor Tu(C8PAS6) 0.9
Actin, cytoplasmic 2 0.9
Phosphocarrier, HPr family 0.91
ATP synthase subunit alpha 0.92
Pyrophosphate phospho−hydrolase 1.05
Cold−shock DNA−binding domain protein 1.28
Phosphoglycerate kinase(A0A0E9DXN4) 1.72
Glyceraldehyde−3−phosphate dehydrogenase 1.87
Chaperone protein DnaK 2.05
d
ATP binding 0.47
D−alanine−poly(phosphoribitol) ligase activity 0.47
rRNA binding 0.52
aminoacyl−tRNA editing activity 0.53
adenylate kinase activity 0.54
5S rRNA binding 0.62
GTP binding 0.63
ribosome binding 0.63
phosphoglucosamine mutase activity 0.64
translation elongation factor activity 0.64
structural constituent of ribosome 0.65
phosphoglycerate kinase activity 0.68
amidase activity 0.7
DNA binding 0.77
adenylosuccinate synthase activity 0.84
hydrolase activity 1.18
glutamate−ammonia ligase activity 1.22
GTPase activity 1.51
beta−phosphoglucomutase activity 2.92
oxidoreductase activity 3.27
e
p=9.4E-25
p=6.4E-101
p=1.7E-94
p=0.79
OOB estimate of error rate (%)
35
40
45
50
OTU
OTU-based function
Protein
Protein-based function

FIGURE 5

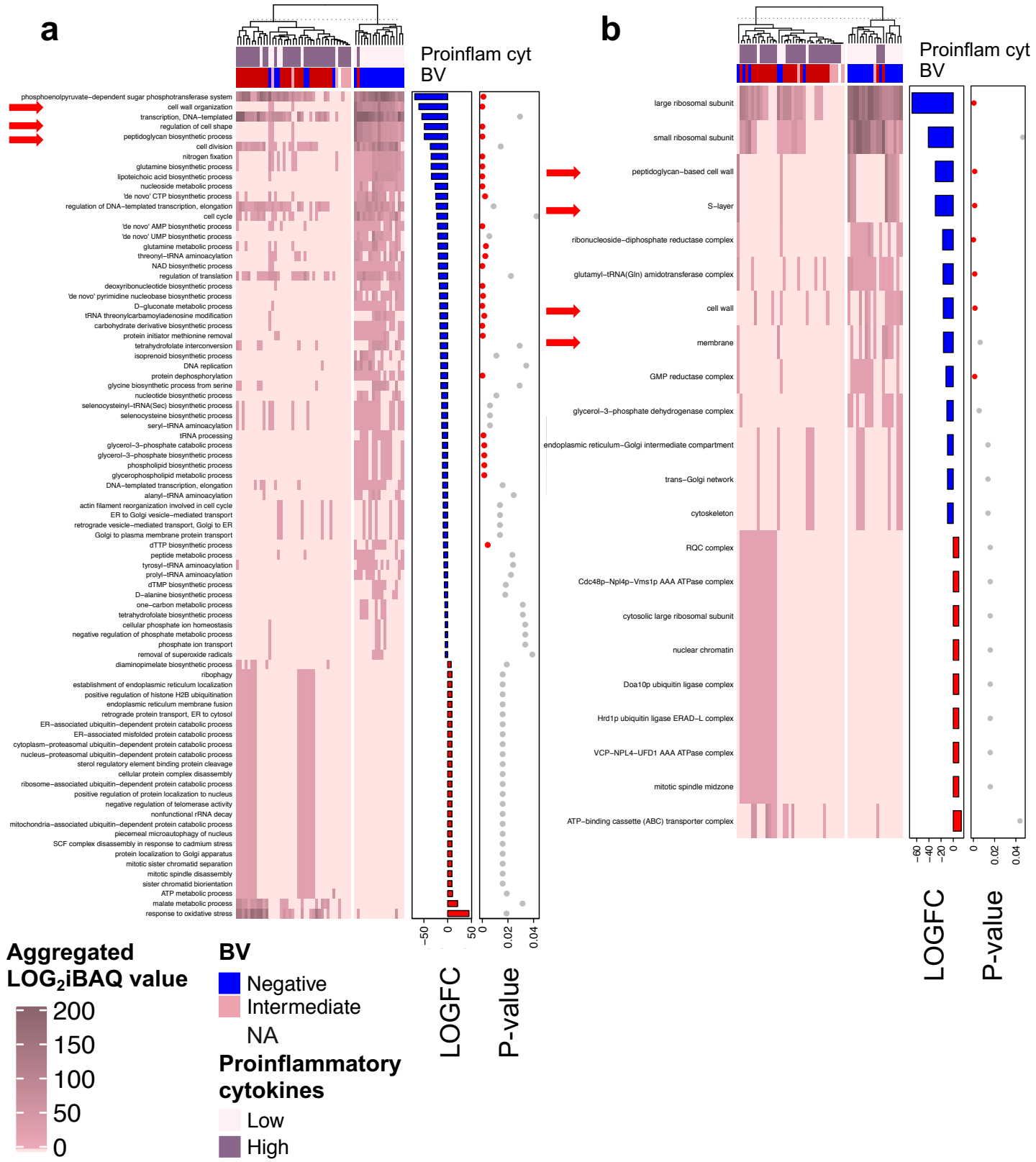
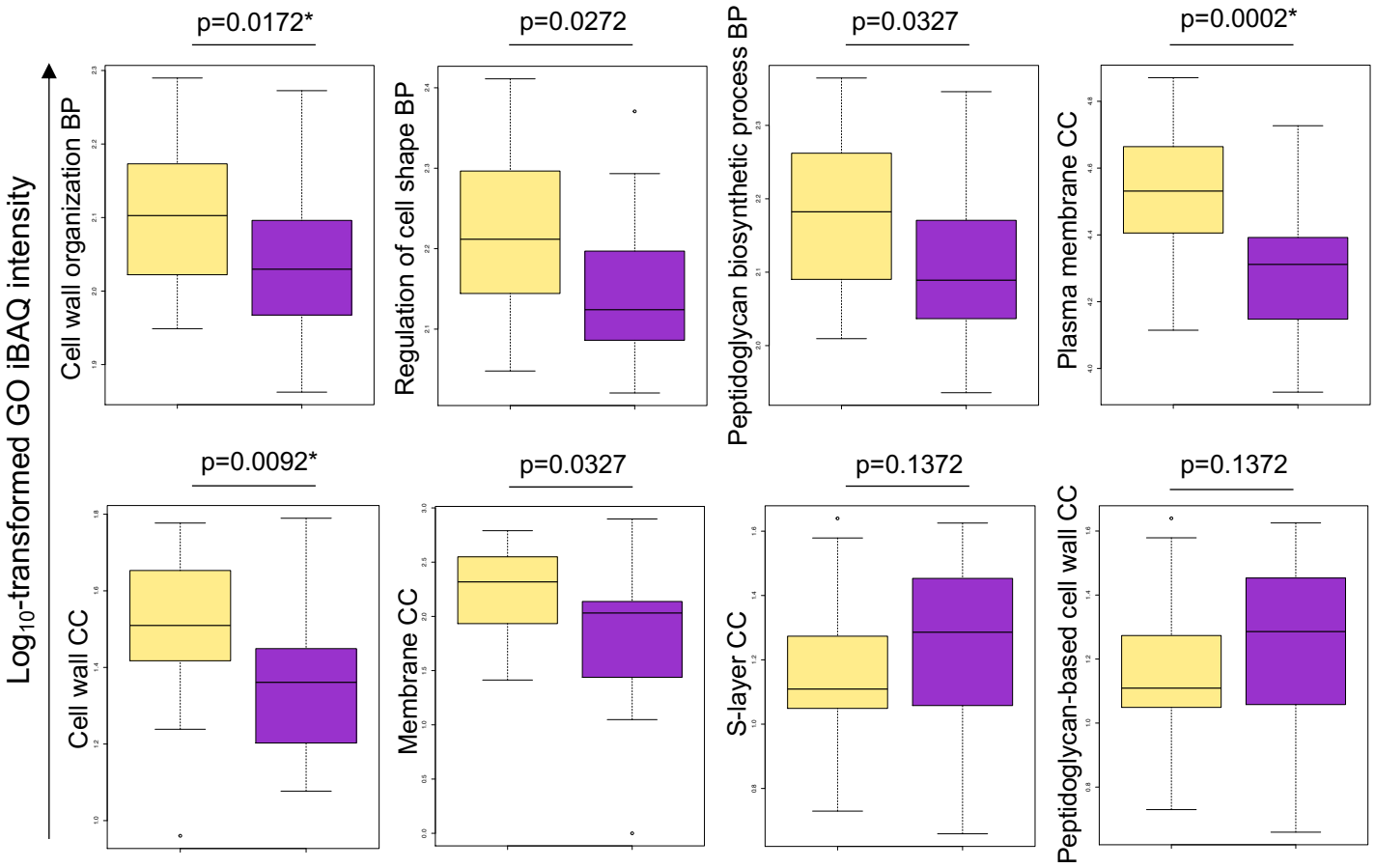

a
Proinflam cyt
BV
phosphoenolpyruvate−dependent sugar phosphotransferase system
cell wall organization
transcription, DNA−templated
regulation of cell shape
peptidoglycan biosynthetic process
cell division
nitrogen fixation
glutamine biosynthetic process
lipoteichoic acid biosynthetic process
nucleoside metabolic process
'de novo' CTP biosynthetic process
regulation of DNA−templated transcription, elongation
cell cycle
'de novo' AMP biosynthetic process
'de novo' UMP biosynthetic process
glutamine metabolic process
threonyl−tRNA aminoacylation
NAD biosynthetic process
regulation of translation
deoxyribonucleotide biosynthetic process
'de novo' pyrimidine nucleobase biosynthetic process
D−gluconate metabolic process
tRNA threonylcarbamoyladenosine modification
carbohydrate derivative biosynthetic process
protein initiator methionine removal
tetrahydrofolate interconversion
isoprenoid biosynthetic process
DNA replication
protein dephosphorylation
glycine biosynthetic process from serine
nucleotide biosynthetic process
selenocysteinyl−tRNA(Sec) biosynthetic process
selenocysteine biosynthetic process
seryl−tRNA aminoacylation
tRNA processing
glycerol−3−phosphate catabolic process
glycerol−3−phosphate biosynthetic process
phospholipid biosynthetic process
glycerophospholipid metabolic process
DNA−templated transcription, elongation
alanyl−tRNA aminoacylation
actin filament reorganization involved in cell cycle
ER to Golgi vesicle−mediated transport
retrograde vesicle−mediated transport, Golgi to ER
Golgi to plasma membrane protein transport
dTTP biosynthetic process
peptide metabolic process
tyrosyl−tRNA aminoacylation
prolyl−tRNA aminoacylation
dTMP biosynthetic process
D−alanine biosynthetic process
one−carbon metabolic process
tetrahydrofolate biosynthetic process
cellular phosphate ion homeostasis
negative regulation of phosphate metabolic process
phosphate ion transport
removal of superoxide radicals
diaminopimelate biosynthetic process
ribophagy
establishment of endoplasmic reticulum localization
positive regulation of histone H2B ubiquitination
endoplasmic reticulum membrane fusion
retrograde protein transport, ER to cytosol
ER−associated ubiquitin−dependent protein catabolic process
ER−associated misfolded protein catabolic process
cytoplasm−proteasomal ubiquitin−dependent protein catabolic process
nucleus−proteasomal ubiquitin−dependent protein catabolic process
sterol regulatory element binding protein cleavage
cellular protein complex disassembly
ribosome−associated ubiquitin−dependent protein catabolic process
positive regulation of protein localization to nucleus
negative regulation of telomerase activity
nonfunctional rRNA decay
mitochondria−associated ubiquitin−dependent protein catabolic process
piecemeal microautophagy of nucleus
SCF complex disassembly in response to cadmium stress
protein localization to Golgi apparatus
mitotic sister chromatid separation
mitotic spindle disassembly
sister chromatid biorientation
ATP metabolic process
malate metabolic process
response to oxidative stress
−50
0
50
0
0.02
0.04
LOGFC
P-value
b
Proinflam cyt
BV
large ribosomal subunit
small ribosomal subunit
peptidoglycan−based cell wall
S−layer
ribonucleoside−diphosphate reductase complex
glutamyl−tRNA(Gln) amidotransferase complex
cell wall
membrane
GMP reductase complex
glycerol−3−phosphate dehydrogenase complex
endoplasmic reticulum−Golgi intermediate compartment
trans−Golgi network
cytoskeleton
RQC complex
Cdc48p−Npl4p−Vms1p AAA ATPase complex
cytosolic large ribosomal subunit
nuclear chromatin
Doa10p ubiquitin ligase complex
Hrd1p ubiquitin ligase ERAD−L complex
VCP−NPL4−UFD1 AAA ATPase complex
mitotic spindle midzone
ATP−binding cassette (ABC) transporter complex
−60
−40
−20
0
0
0.02
0.04
LOGFC
P-value
Aggregated
LOG2iBAQ value
200
150
100
50
0
BV
Negative
Intermediate
NA
Proinflammatory
cytokines
Low
High

FIGURE 6



Lactobacillus isolates inducing low inflammatory responses (n=22)
Lactobacillus isolates inducing high inflammatory responses (n=22)