Erasmus+

ERASMUS+ Programme, Key Action 1 – Student Mobility for Traineeship
A.A. 2020/2021

UNIPHARMA-GRADUATES PROJECT

Scientific coordinator:
Prof. Daniela De Biase, Sapienza University of Rome

SCIENTIFIC REPORT

01/02/2022 – 31/07/2022

"Improving the Understanding of the Vaginal Microbiota through *De Novo* Structural Protein Prediction and Annotation"

Trainee
Maria Victoria Bussoletti Panizo

Tutor/supervisor
Sean P Kennedy

Institut Pasteur, Paris, France

Date of submission: 30/08/2022

Abstract

There is a strong relationship between the vaginal microbial community and the human host. Five types of microbial communities in the vagina have been identified and, depending on the community present, women show variable susceptibility to diseases and different pregnancy outcomes. While there is some knowledge about the different microbial communities' compositions, the unique features of each species and the understanding of the bacteria-host interactions are hampered by an incomplete functional annotation. In this work a pipeline has been developed, with the application of protein structure prediction techniques, to improve the *de novo* functional annotation of unknown genes within the vaginal microbiota. The results obtained show that *in silico* protein annotation proves to be a valid method to start understanding the coding content of bacterial species and could guide *in vitro* studies towards a more precise and comprehensive annotation of the vaginal microbiota communities.

Introduction

The human body contains more than 100 trillion symbiotic microorganisms[1], collectively referred to as the microbiota. Microbiota research has increasingly gained the interest of the scientific community and is now recognized as having a fundamental role in the development and physiology of the human being. Microbial communities establish a beneficial and finely balanced mutualistic relationship with the human host and are believed to be the first line of defence against infection by competitively excluding invasive non-indigenous organisms [2]. The alteration of microbiota composition or its dysfunction is often associated with diseases. The genetic profile of all the species that compose these microbial communities is called the microbiome. With culture-independent methods and genomic approaches, we are able to characterize and compare different microbial communities and expand the knowledge of microbiome composition, function and role in both health and disease. Using high throughput sequencing (HTS) technologies, such as 16S rDNA sequencing, it is possible to both have information about the genetic content and to quantify the relative abundance of the different species residing in different niches of the human body. The microbiome provides humans with unique and specific enzymes: for example, the gut microbiome is enriched for genes involved in starch, sucrose and other carbohydrates catabolism, complementing a low presence in the human genome [3] . They provide also biochemical pathways that are generally beneficial to the host. Commensal organisms in the microbiota play a role in the competitive exclusion of pathogens, in the development of immune system and in the production of antimicrobial substances [1,4].

The human vagina and the bacterial community that resides within are an example of this beneficial association. Again, the disruption of the equilibrium of the vaginal microbiome is often associated with enhanced risk of acquiring sexually transmitted diseases, bacterial vaginosis, fungal infections and preterm birth [2]. The interaction between the human host and the vaginal microbiota is highly dynamic: the vaginal physiology and the microbiota composition go through major changes over a woman's lifetime, shaped by transitional periods such as puberty, menopause and pregnancy, while daily fluctuations are the results of daily activities and lifestyle [5].

High-throughput 16s rRNA gene sequencing has shown that, in the vagina of reproductive-aged women, it is not possible to identify a standard core microbiome. Instead researchers have identified at least five major types of microbial communities, known as Community State Types (CST)[5]. Vaginal CSTs, unlike others human microbiomes, are generally characterized by a lower microbial diversity, dominated by *Lactobacillus* species [2,6]. The CSTs can be further classified by the exact species of *Lactobacillus* present: CST I is dominated by *Lactobacillus crispatus*, the CST II by *L. gasseri*, the CST III by *L. iners* and the CST V is dominated by *L. jensenii*. The CST IV, on the contrary, is characterized by an higher diversity and contains a diverse array of facultative or strict anaerobic bacteria [7] , such as *Prevotella* and *Gardnenella.* Every woman in her lifespan is susceptible to transition from one CST state type to another, and the composition of the vaginal microbiome can be shaped by genetic factors, such as ethnicity [5], as well as other internal and external factors including menstruation, hormonal contraceptives, antibiotics and sexual activities [2].

While there is some knowledge about the CST and their composition, little is known about the genetic and metabolic features of each dominating species and the specific relationship each species establishes with the host.

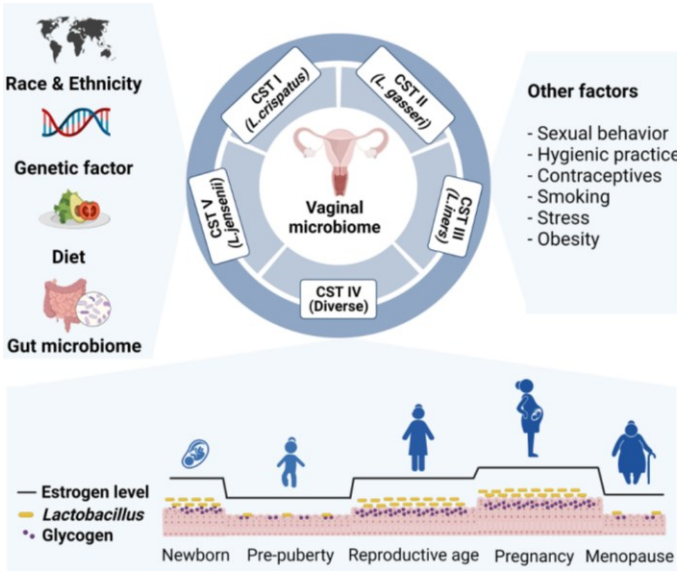

Figure 1: overview of the composition of the different vaginal microbiomes, and the intrinsic and extrinsic factors that affect the vaginal microbial community [8]

Members of the *Lactobacillus* genus are generally identified as the hallmark of a healthy vagina during reproductive age: their production of lactic acid lowers the pH of the vaginal environment (pH ≤ 4.5), which is highly protective against infections or colonization of the vagina by pathogens and non-indigenous microbes [5,9]. Women with a genital microbiota dominated by *Lactobacilli* have a significantly lower risk of acquisition of sexually transmitted diseases and experience improved pregnancy outcomes [9]. It has been shown that lactic acid is able to inhibit *Chlamydia trachomatis* infection and also HSV-2 and HIV, when the pH is lower than 4.0 [5]. *Lactobacilli* are also able to produce bacteriocins, bactericidal compounds with a narrow-spectrum of killing, achieved by enhancing the permeability of the target cell membrane [10]. Another agent produced by this genus is hydrogen peroxide, which synergizes with lactic acid to establish an optimal low-diversity environment [2]. It has been suggested that hydrogen-peroxide producing species are most likely to maintain persistent vaginal colonization [11] and to be protective against acquisition of BV [7], pre-term birth and HIV acquisition [12] : however, this important characteristic seems to be limited to only some *Lactobacillus* species.
The vaginal *Lactobacillus* abundance and thus the establishment of the protective acidic environment are connected to estrogen levels, which are higher in reproductive age and pregnancy [2,9]. Estrogen controls the vaginal epithelial accumulation of glycogen, whose levels strongly correlate with the level of colonization of *Lactobacillus* and the decrease of the vaginal pH [9]. Glycogen is primarily catabolized by human α-amylase into smaller sugars, like maltose and maltotriose, favoring *Lactobacillus* growth, that in turn produces lactic acid and bacteriocins, contributing to the host defense. [5,13].

On the other hand, a lower abundance of *Lactobacillus* and a type IV CST have been associated with an increased risk of bacterial vaginosis (BV), an anaerobic polymicrobial disease. This community state is highly similar to that associated with various sexually transmitted diseases, including HIV, and has also been linked with a higher risk of preterm birth [2,7]. The presence of CST-IV is associated with an higher vaginal pH, predisposing women to genital infections [10], and the triggering of a pro-inflammatory response [5]. The composition of CST-IV is characterized by a diverse array of strict anaerobic bacteria, with *Gardnerella vaginalis* among one of the predominant species [14]. *G. vaginalis* is also the most common microorganism identified from vaginal samples of women with BV [15] , and its presence is a positive criterion of the disease in the Nugent score test, used for the diagnosis of BV based on the relative abundance of some bacterial morphotypes [14]. However, little is known about its role in BV: while the initial hypothesis was it was the only causative agent, more recent studies paint a more complex picture, in which *G. vaginalis* could have

both a direct and indirect role in BV, and may change the host landscape in a way that makes other organisms more likely to colonize or cause disease [16]. In addition, even though *Gardnerella* species are generally regarded as a major cause of BV, they are also detected in the vaginal microbiome of healthy women [17]. While normally this CST manifests as BV [5], there is a significant percentage of healthy and asymptomatic women, especially of Black and Hispanic ethnicity, that have an higher vaginal pH, a vaginal microbiota mainly dominated by these strict anaerobic bacteria and generally lacking in *Lactobacillus* species [2,5,7]. These differences might be driven by host factors and behavioral differences, that may play a role in determining the vaginal microbial community composition. Moreover, this diversity complicates the dynamics of the relationship between the bacterial communities and the host and any determination of what should be considered healthy or not. In fact, in absence of symptomatology, this type of vaginal community might be considered normal and healthy, even though its composition closely resembles those associated with symptomatic BV and it is associated with significantly increased risk of adverse pregnancy outcomes. In addition to that, metagenomic analysis showed that the core genome shared by all *G. vaginalis* isolates consists of only 25% of the total genes in the pangenome [16], suggesting a diversity of outcomes depending on individual colonizations. This opens new questions about the possible presence of *G. vaginalis* strains with different virulency [16] or other unknown protective traits that are associated with this CST.

The *Lactobacillus*-dominated CSTs are not necessarily equivalent, as every species has its unique characteristics and adaptations. *L. crispatus*-dominated communities are able to establish a more acidic environment in comparison with the other species [10,14]. This CST is the most stable and has a strong correlation with eliminating and excluding *Gardnerella* [10]. Conversely, *L. iners* fails to produce hydrogen peroxide and is considered among the least protective. *L. iners* is associated with higher pH, as it fails to produce lactic acid in abundance as other *Lactobacillus* strains [5], and can co-occur with *Gardnerella*, along withother BV-associated bacteria and inflammatory processes. These facts, render CST III the most prone to transition to a diseased state [6,18,19]. It is also known that the genome of *L. iners* encodes virulence factors like cytolysins and that can trigger an intermediate immune response [5,20].
It follows that to fully understand the differences between each species and their relationship with the host is it important to gain more knowledge about their specific genetic content: at the current state of knowledge, surprisingly, the majority of genes in the vaginal microbiome is still unannotated. Indeed, many of the unique genetic features of each species do not have known functions. The identification and annotation of each species' unique features is important to identify the different factors that help the specific colonization of a given species, to understand the susceptibility to pathogens and to prevent infectious diseases and adverse outcomes in pregnancy.

The study of the composition of the vaginal microbiota done in this internship is part of the framework of the Innovative Strategies for Perinatal Infection Risk-Reduction (InSPIRe) project. InSPIRe is an ambitious collaboration that includes teams from Paris hospitals, research institutes and the private sector. The primary goal of InSPIRe is to identify biomarkers of perinatal infections and antibiotic resistance to develop a medical device that can detect these markers in less than 15 minutes at the time of delivery. To this aim, vaginal swabs from pregnant women are collected and then the samples are treated for shotgun metagenomic sequencing. It was shown that the vast majority of vaginal communities of pregnant women between 22 and 37 weeks of gestation are dominated by CST I, III, and IV. For this reason, we focused on the functional annotation of *L. crispatus*, *L. iners* and *G. vaginalis*, respectively, to better understand their unique features and adaptation to the vaginal environment.

Aim of the work

The understanding of the bacteria-host interactions is hampered by an incomplete functional annotation. The aim of this work is to take advantage of the proven correlation between protein structure and protein function and apply ground-breaking protein structure prediction techniques to improve the annotation of dominant vaginal microbiota species, in order to better understand their adaptations to and impact on the host vaginal microbiota. The goal is to develop a systematic pipeline that allows the *de novo* functional annotation of predicted ORFs in bacterial genomes.

Results

1. Development of a pipeline to annotate genes with unknown function in the vaginal microbiota

Since structure is more conserved than sequence and there is a correlation between protein structure and its function [21], the annotation based on the protein structure prediction can be considered as an initial step towards a comprehensive annotation and understanding of the vaginal microbiota composition and functions. The goal of this project was to develop a pipeline that leverages the recent breakthroughs in protein structure prediction accuracy, in order to provide functional annotation of unknown genes within the vaginal microbiota. To this aim, a series of Python scripts was specifically developed and adapted in order to predict, starting from the nucleotide sequence, the protein structure with AlphaFold2 [22], and annotate the results thanks to the structural comparison with annotated proteins with Foldseek [23].

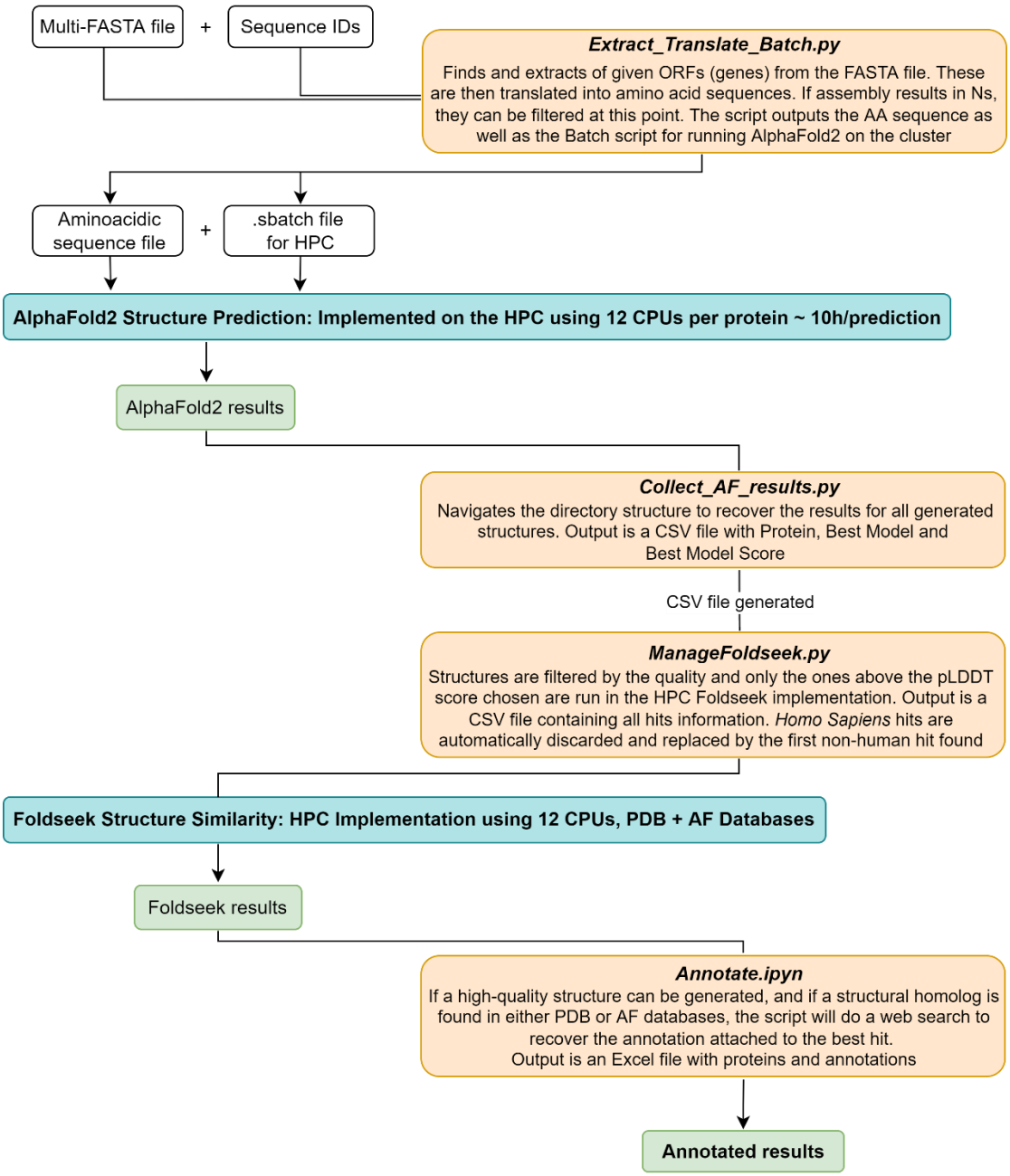

Figure 2: pipeline description. In white, the input and output files, in orange the Python scripts developed and in green the data generated

The genes used were taken from the VIRGO database, a non-redundant gene catalogue of the vaginal microbiota [24]. For the protein structure prediction, we selected unannotated sequences from *Lactobacillus crispatus*, *Lactobacillus iners* and *Gardnerella vaginalis*, with a ORF length between 600 and 2100 bp: this allowed to have at least 200 amino acids-long sequences, which is enough be modelled, and discard all the sequences with more than 700 amino acids, which would have been too demanding in terms of memory usage on the HPC cluster. The sequenceID was exported from our intranet database MetageneDB [25], based on the VIRGO database.

The *Extract_translate_batch.py* script allows extraction of the nucleotide sequences corresponding to each sequenceID from the VIRGO multi-FASTA file and their translation into aminoacidic sequences. At the same time, the script generates, for each translated sequence, a batch file to run AlphaFold2 on the Institut Pasteur HPC cluster. Five models for each protein were generated. The script is also able to handle Ns-containing sequences and filter them by a parameter given by the user. If Ns-containing sequences are chosen to be translated, the triplet is translated as a glycine: this amino acid is the smallest and has often a linker role in tertiary structures, hence, among all the amino acids is the one that could have the least impact on the predicted protein structure.

The predicted structures were then filtered by the *Collect_AF_results.py* script based on their pLDDT score, which is an estimation of the accuracy of the model [26]. Only the predicted structures with a score above 70 pLDDT were further used for the functional prediction and comparison with protein online databases. Empirical testing and published data suggested that, protein structural prediction below this score are expected to have less accurate backbone predictions and less reliable overall modelling [26].

The functional prediction from generated structure was done with Foldseek, a software that enables fast and sensitive protein structure search and comparison [23]. The *ManageFoldseek.py* script allows the use of the Foldseek implementation on the Institut Pasteur HPC cluster and the organization of all the results. This script runs the comparison between the AlphaFold2 PDB-format predicted structures above the threshold pLDDT score and the structures available in the Protein Data Bank (PDB) and UniProt databases. Then, for each predicted protein chooses the best available hit on each database based on the E-value. The structural comparisons' results were organized in two CSV files, one for each database, by sequenceID, number of total hits in the database, the first hit identifier, the E-value of the first hit, the species of the first hit and additional notes. The script has also a built-in function that allows it to filter and discard *Homo Sapiens* hits. This additional species filtering was made to enable a functional annotation with results that come from mostly prokaryotic organisms, coherent with our target organisms.

All the results were then annotated with the *Annotate.ipynb* script. This script retrieves from the PDB/UniProt databases the annotation for each predicted protein best hit. The final result is an Excel file in which are added the category and the main keywords related to each predicted protein's best hit.

2. Biological annotations

The results generated by the Python pipeline were manually reviewed. The focus was on specific categories of interests that could give an insight into the microbiota-host interaction and the unique features of each species. The primary features considered were cell adhesion, membrane proteins, bacteriocin production, metabolism, transcription factors involved in oxidative stress. Other features were also considered depending on the species. Annotation was performed mainly with the hits found in the PDB, as this database contains only experimentally validated structures. For the hits that had no PDB structure or with insufficient information, UniProt annotation was taken into consideration: this database contains other predicted structures and putative domains that might help in the functional annotation. For the most interesting hits an additional structural comparison was done manually using the Matchmaker function in

Chimera, to better visualize the structural similarities and the conservation of specific aminoacidic residues between the predicted structure and PDB structures.

Category	*Lactobacillus crispatus*	*Lactobacillus iners*	*Gardnerella vaginalis*
Antibiotic		1	
Antimicrobial protein	1		
Antitoxin	1	2	3
ATP-binding protein	1		
Biosynthetic protein	5	1	1
Carbohydrate binding protein	1		
Cell adhesion	9	20	170
Cell cycle		1	8
Cell invasion		4	7
Chaperone		2	2
Choline-binding protein	2	1	
Coupling protein	1		
DNA/RNA binding protein	30	21	38
Electron Transport		1	
Helicase	1		
Heme-Binding Protein	1		
Hydrolases	28	27	80
Immune System	2	6	20
Isomerase	3	1	5
Ligase		3	6
Lipid Binding Protein	3	1	
Lipid Transport			6
Membrane Protein	4	5	11
Metal Binding Protein	1	2	2
Metal Transport			5
Nucleic Acid Recognition			2
Oxidoreductase	4	5	15
Protein Binding	2	4	4
Protein Fibril	2		
Protein Transport	4		
Recombination	2	2	2
Replication	5	3	
Signaling Protein	2	4	4
Structural Protein	5	4	14
Sugar Binding Protein	1	2	3
Toxin		5	9
Transcription Factors	8	3	3
Transferases	17	16	85
Transport Protein	12	4	38
Unknown Function	7	4	14
Viral Protein	12	26	30
Virus	6	4	4

Table 1: PDB categories for all the hits found in all 3 species. Categories uniquely found in some species are highlighted in different colors, depending on the species considered. In green are highlighted categories found exclusively in L. crispatus, in blue categories found in L. iners and G. vaginalis excluding L. crispatus, in yellow categories found only in G. vaginalis

Below are reported some biological examples of the most interesting hits found for each species. Additional information can be found in the supplementary material. Before introducing the results, it is important to highlight that the aminoacidic identities between the unknown genes and the reference genes are almost all under 30% and would not generally yield any relevant hits using primary sequence alignments tools: this further underlines the advantage of using structural predictions for genome annotation.

2.1 *Lactobacillus crispatus and Lactobacillus iners*

	Lactobacillus crispatus	***Lactobacillus iners***
Number of total sequences modelled	450	523
Predicted structures with at least one hit in PDB	187	187
Predicted structures with at least one hit in UniProt	271	220
Predicted structures with no hits in either database	162	263
Predicted structures with only *Homo Sapiens* hits	4	3

Table 2: data generated for L. crispatus and L. iners with the de novo functional annotation pipeline

2.1.1 Cell adhesion

Bacterial adherence to the host epithelial surface is often mediated by exopolysaccharides, S-layer proteins and other extracellular appendages. In the gut microbiota it was shown that some *Lactobacilli* are able to adhere via cell surface proteins with mucus-binding capacity [27]. Previous studies have suggested that *Lactobacilli* have a role in the maintaining of an healthy vagina by their adhesion to vaginal epithelial cells and/or mucus: this can promote colonization and exclusion and growth prevention of pathogenic microorganisms [27].

In *L. crispatus* we have identified putative cell adhesion proteins that could play a fundamental role in explaining the predominance of this species in the women vaginal microbiota.
A good hit was found with a Mucus Binding Protein (MUB) from *Limosilactobacillus reuteri subsp. suis*, a multi-repeat cell-surface adhesin involved in the interaction with mucus and colonization of the digestive tract. The binding of full-length MUB to mucus is via multiple interactions involving terminal sialylated mucin glycans. Each repeat consists of tandemly arranged Ig- and mucin-binding protein (MucBP) modules [28]. MubBP repeats are thought to be specific to Lactic Acid Bacteria (LABs) and mediate the host-microbe interaction. This particular organization of MUB repeats provides a structural explanation for the mechanisms in which *Lactobacilli* have adapted to their host niche by maximizing interactions with the mucus receptors, potentiating the retention of bacteria within the mucus layer [28].

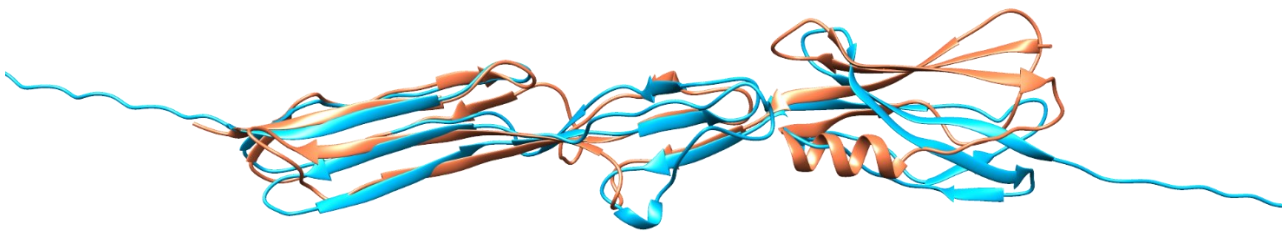

Figure 3: structural comparison made with MatchMaker between sequence V1237124 (blue) and 4MT5 (orange), a Mucus-Binding Protein from Limosilactobacillus reuteri subsp. suis. The aminoacidic identity between the two structures is 29.35%

We also identified several potential S-layer proteins, proteins that are commonly found on the bacterial surface. The predicted structures match with CbpA, an S-layer choline-binding protein from *Ligilactobacillus salivarius str. ren*. It is reported that CbpA acts as a multifunctional adhesin that is able to cleave the host extracellular matrix and to adhere to the human gut, allowing the colonization and interaction between the bacterium and the host [29].

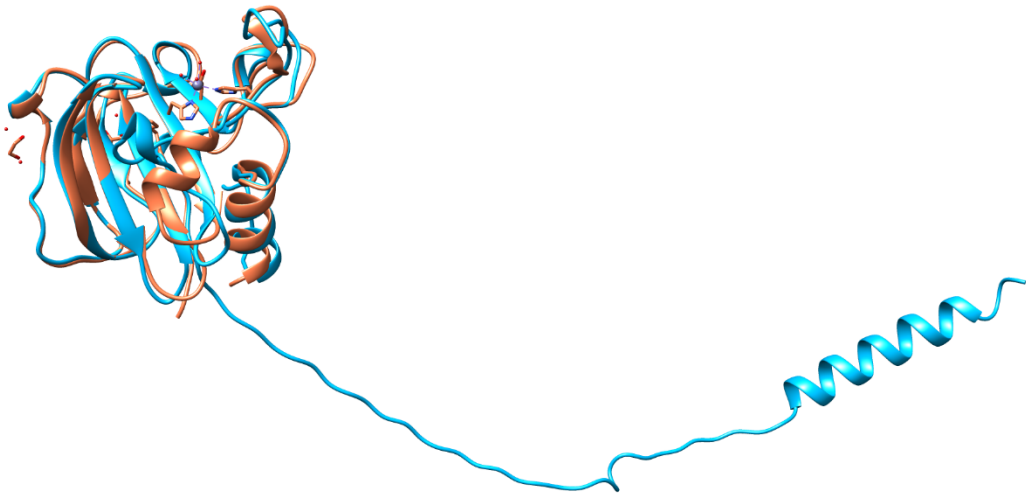

Figure 4: structural comparison made with MatchMaker between sequence V1033564 (blue) and 5GT1 (orange), a choline-binding protein from Ligilactobacillus salivarius str. ren. The aminoacidic identity between the two structures is 32.76%

While the majority of the proteins found with the homology search in *L. crispatus* came mostly from other beneficial gut species, a higher percentage of *L. iners* hits were with proteins coming from uropathogenic species. This could be a possible indicator of why *L. iners* can be found in association with potentially pathogenic bacteria and it is the CST most prone to the transition into a diseased state.

One of the hits was with UafA, the Uro-adherence factor A coming from *Staphylococcus saprophyticus*, a gram-positive bacteria involved in urinary tract infections. UafA is the only cell-wall anchored protein in this bacterium, necessary to survive in the human urinary tract. The functional region, composed by 3 domains, binds the host and has also an erythrocyte binding activity [30].

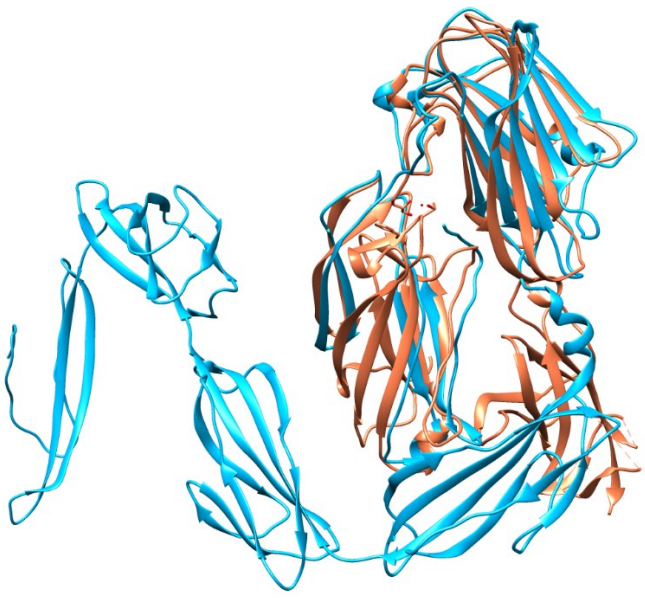

Figure 5: structural comparison made with MatchMaker between sequence V1195929 (blue) and 3IRZ (orange), a cell adhesion protein from Staphylococcus saprophyticus. The aminoacidic identity between the two structures is 16.32%

Other predicted structures might be involved in adhesion, as they resemble the domains from a cell surface protein from *Streptococcus*. These domains are known as SHIRT (Streptococcus High Identity Repeats in Tandem), and they are present in a recently categorized class of "Periscope proteins". These proteins are composed by tandem arrays of highly similar folded domains. The protein length can vary depending on the nature of the interaction with the host. It is suggested that this length variability given by these SHIRT domains can play an important role in regulating bacterial interaction with the hosts. All the predicted structures contain a highly conserved SHIRT domain [31].

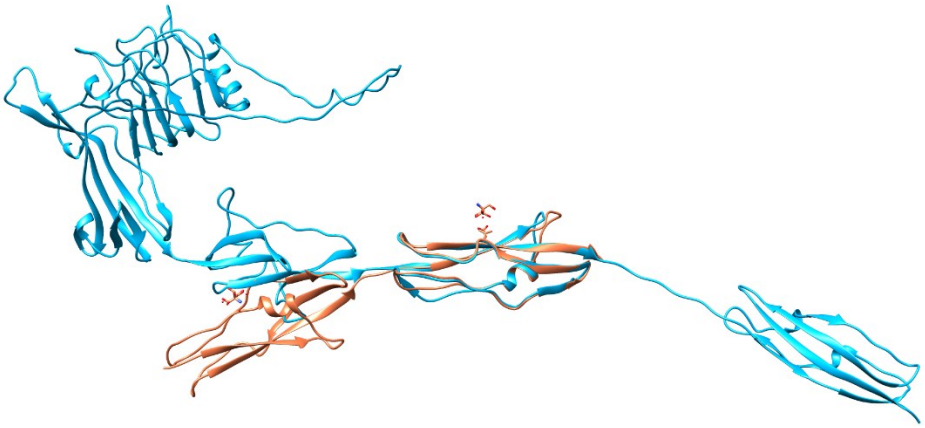

Figure 6: structural comparison made with MatchMaker between sequence V1506282 (blue) and 7AVH (orange), a cell surface protein containing a SHIRT domain from Streptococcus. The aminoacidic identity between the two structures is 27.54%

2.1.2 Bacteriocins

Bacteriocins are a diverse group of ribosomally synthetized antimicrobial peptides produced by bacteria [32]. These compounds are active against other bacteria, while the producer has a specific immunity mechanism. Previous studies have proven that some *Lactobacillus* strains obtained from human vaginal samples are able to produce active bacteriocins [33], and there is evidence that bacteriocin production may impact the ability of the producer strain to compete within complex microbial communities and positively influence the health of the host, by directly inhibiting the invasion of competing strains or pathogens.

In *L. crispatus*, a strong homology was found with NisC, a Lantibiotic cyclase from *Lactococcus lactis,* another member of the gut microbiota. It is one of the enzymes involved in the synthesis of Nisin, one of the best known bacteriocins, effective against a range of gram-positive bacteria. This enzyme contains a characteristic thioether rings that are essential for the biological activity of the produced bacteriocin [34].

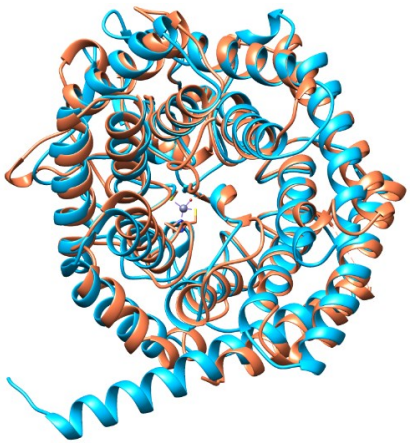

Figure 7: structural comparison made with MatchMaker between sequence V1806592 (blue) and 2G0D (orange), the NisC lantibiotic cyclase from L. lactis. The aminoacidic identity between the two structures is 24.21%

Another match with a bacteriocin-related protein among the *L. crispatus* structures was with the sheath protein of an R-type diffocin produced by *Clostridium difficile*. Diffocins are high-molecular-weight phage tail-like bacteriocins that act as molecular puncture devices, able to specifically penetrate the cell envelope of other *C. difficile* strains to dissipate the membrane potential and kill the attacked bacterium. The structure of R-type bacteriocins is highly conserved and their killing capacity has narrow strain specificity. Interestingly, they have been proposed as novel therapeutics to treat infections cause by antibiotic-resistant bacteria [35].

Figure 8: structural comparison made with MatchMaker between sequence V1884532 (blue) and 6GKW (orange), an R-type diffocin from C. difficile. The aminoacidic identity between the two structures is 14.04%

In both *L. crispatus* and *L. iners*, were found structures matching with the core of TruD, a heterocyclase enzyme that comes from a cyanobacteria, *Prochloron sp. 06037A*. This enzyme is capable of heterocyclizating cysteins to form thiazolins, a 5-ring compound commonly found in microcins, drugs and toxins [36].

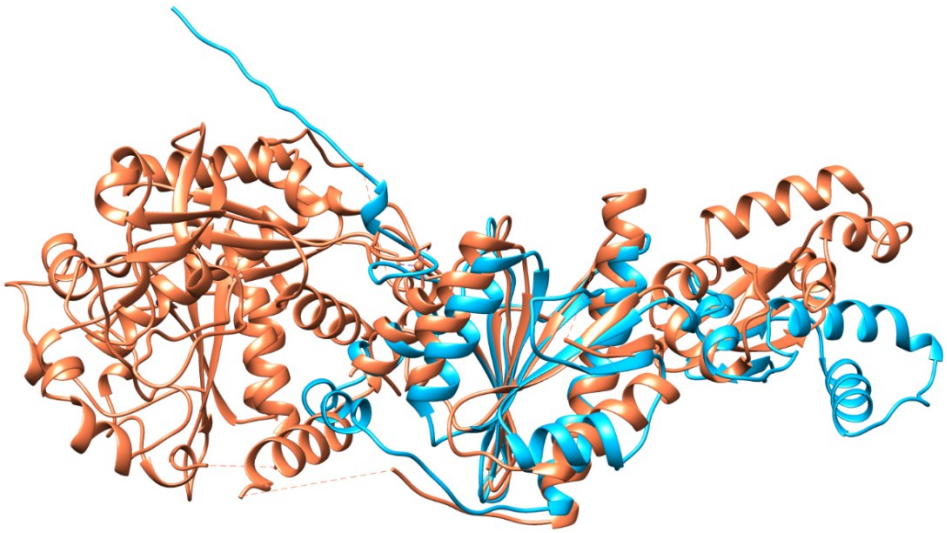

Figure 9: structural comparison made with MatchMaker between sequence V1547326 (blue) from L. iners and 4BS9 (orange), a TruD cyclase from a cyanobacteria. The aminoacidic identity between the two structures is 15.25%

There were also identified in both *L. crispatus* and *L. iners* two chains of a Microcin synthetase from *Escherichia coli str. K-12 substr. MG1655.* This biosynthetic enzyme is composed by an octameric complex and produces microcin B17, a bacterial topoisomerase inhibitor by converting serine and cysteine residues on selected peptides to oxazoles and thiazoles [37]. A hypothesis is that it might be also involved in the formation of some *Lactobacillus* unique bacteriocins.

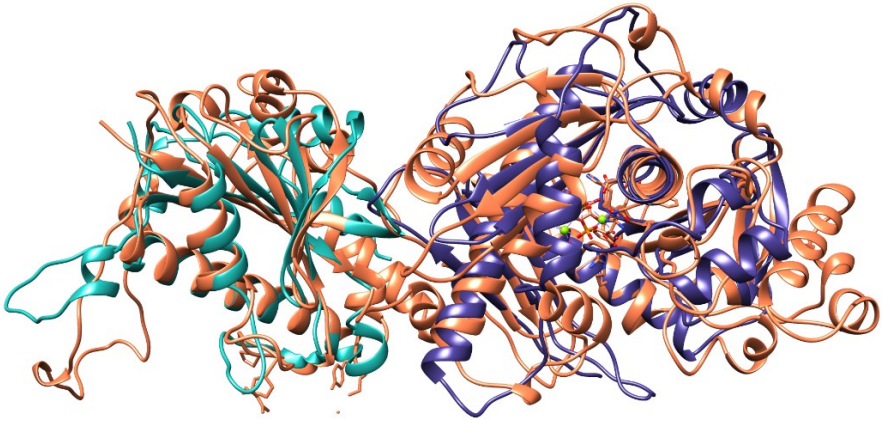

Figure 10: structural comparison made with MatchMaker between sequence V1046759 from L.iners (green), sequence V1808260 from L. crispatus (purple) and chains C and D from 6GRG (orange), a Microcin synthetase from E.coli. The aminoacidic identity between V1046759 and 6GRG is 15.24%, and between V1808260 and 6GRG is 17.27%

2.1.3 Transcription factors

Some *L. crispatus* strains are able to produce hydrogen peroxide, one of the reactive oxygen species (ROS). ROS can react with lipids, proteins and nucleic acids causing oxidative cell damage. It is reported that *Lactobacilli* possess mechanism of protection against their own hydrogen peroxide [38], and here it is reported the possible presence of a transcriptional factor controlling the antioxidant response. There is a good structural match between our predicted model and one of the chains of OxyR from *Pseudomonas aeruginosa PAO1* [39]. OxyR is a multimeric transcriptional factor in bacteria that controls the antioxidant response. It activates a wide range of genes involved in the antioxidative defense. OxyR is capable of sensing low concentrations of H_2O_2 within the cell, binding hydrogen peroxide. Subsequent oxidation causes a structural change in OxyR that allows the binding between the transcription factor and DNA, allowing the expression of ROS response genes [39]. Here the predicted model has a good overlap with one of the OxyR tetramer subunits, and the overlap difference with the N-terminal hinge might be caused by the structural change caused by the binding with hydrogen peroxide, present in the experimentally validated structure.

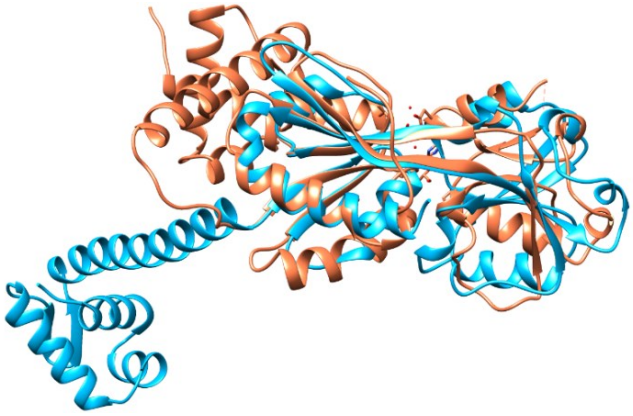

Figure 11: structural comparison made with MatchMaker between sequence V1237198 (blue) and the chain F of 4X6G (orange), a hydrogen peroxide-related transcription factor from P. aeruginosa PAO1. The aminoacidic identity between the two structures is 20.56%

Interestingly, in *L. iners* it was identified a hydrogen-peroxide related enzyme, a thioredoxin from *L. lactis*. Thioredoxins are proteins involved in various cellular redox processes, and their presence is important in catalase-negative bacteria, as LAB, because they are fundamental to protect the cell from oxidative stress. The presence of thioredoxin in *L. iners* does not exclude its presence also in other vaginal bacterial species, and it could explain how the different species can survive in presence of hydrogen-peroxide producing *L. crispatus* strains and eventually take over them and cause a shift in the CST composition.

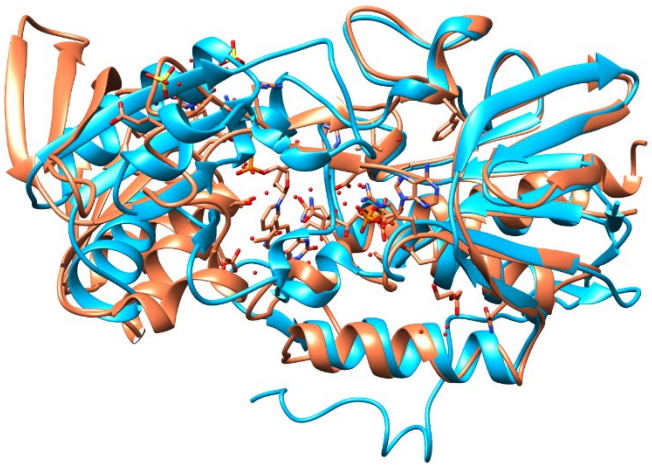

Figure 12: structural comparison made with MatchMaker between sequence V1310512 (blue) and 5MH4 (orange), a thioredoxine from L. lactis. The aminoacidic identity between the two structures is 19.49%

2.1.4 Antibiotic resistance

Antibiotic treatments may induce the emergence of resistant bacteria in the vaginal microbiota. Both in *L. crispatus* and in *L. iners* antibiotic-resistance related proteins were found but in different categories.
In *L. crispatus* only polyspecific membrane-bound transport protein were identified. Here it is reported LmrP, a multidrug transporter from *Lactococcus lactis*, a major facilitator superfamily (MSD) member, that has a broad target specificity [40]. This protein is closely related to MdfA, another drug transporter found both in *L. crispatus* and in *G. vaginalis* (reported later).

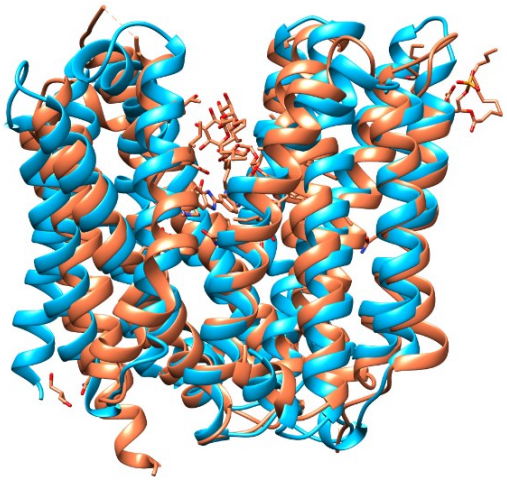

Figure 13: structural comparison made with MatchMaker between sequence V1828794 (blue) and 6T1Z (orange), a multidrug transporter from L. lactis. The aminoacidic identity between the two structures is 17.16%

On the contrary, in *L. iners* one hit found was with a biosynthetic enzyme. There is a good structural homology with the C-terminus region of a Cephalosporine esterase from *Streptomyces clavuligerus*. This region has a "beta-lactamase-like" folding, and even though the reference protein didn't show a penicillin-binding activity, the predicted structure could be involved in the development of antibiotic resistance features in *L. iners* [41].

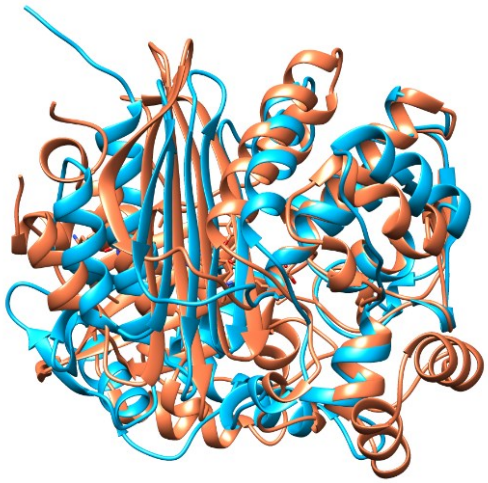

Figure 14: structural comparison made with MatchMaker between sequence V1008909 (blue) and the chain B of 2XF3 (orange), an esterase from S. clavuligerus. The aminoacidic identity between the two structures is 15.21%

2.2 *Gardnerella vaginalis*

	Gardnerella vaginalis
Number of total sequences modelled	1469
Predicted structures with at least one hit in PDB	593
Predicted structures with at least one hit in UniProt	809
Predicted structures with no hits in both databases	604
Predicted structures with only Homo Sapiens hits	1

Table 3: data generated for G. vaginalis with the de novo functional annotation pipeline

In the functional annotation done for both *Lactobacillus* species, a considerable number of sequences were excluded from the AlphaFold2 modelling, because of the presence of undefined nucleotides (Ns) in the sequencing data. Most of the times the total number of undefined nucleotides could be considered negligible, especially when compared to the total sequence length. For this reason, for the functional annotation of *G. vaginalis,* the *Translate_Extract_Batch.py* script was adjusted, allowing the translation and therefore the modelling of sequences that contain undefined nucleotides (N). The triplets that contained at least one N were translated as glycine, as it is a small amino acid often added as a synthetic linker owing to its expected lower impact on the overall structure. This is also a chance to see if even with the introduction of an unknown amino acid in certain positions AlphaFold2 is still able to generate a high-quality structure. In consequence of that, the total number of modelled sequences was higher in comparison with the other species' results. The information concerning the sequences containing Ns is available in the supplementary material. In the biological results, no bacteriocins or oxidative stress response proteins were found in *G. vaginalis*: this suggests that this species might have a different way to interact with the host and compete with the other species in the vaginal environment.

2.2.1 Cell adhesion

Most of the proteins found belonged to the cell adhesion category. The structures that may be involved in cell adhesion in *G. vaginalis* have similarities with the ones experimentally identified in pathogenic and invasive species. This is consistent with the information found in literature and could also explain why it is a dominant species in some healthy women and also during pregnancy.

Pilus proteins are found on the cell surface of prokaryotic cells. They generally have a role in movement but are also involved in adhesion and colonization, a reason why they are considered as a key virulence characteristic. Forty modelled sequences showed a strong structural similarity with SpaD, a major pilin coming from *Corynebacterium diphtheriae*. The surprising data about this hit is that all the modelled sequences showed an E-value comparison with the reference hit on the PDB lower that E^{-20}, confirming a consistently significant result. SpaD is the major pilin protein that forms the polymeric backbone of one of the three types of pili expressed by this human pathogen. They generally have highly variable Ig-like domains and present variation in the number and position of loops and helices, which presumably reflect the bacteria's response to immune pressure and to different host environments. SpaD has two chains and has been shown to be able to bind epithelial cells: thus, it is important for Gram-positive bacteria to adhere and colonize the host [42]. Here it is shown the structural comparison between one of our predicted structures and the chain B of SpaD.
A considerable number of sequences from *G. vaginalis* had a strong structural similarity with FimP (not shown here), a fimbrial protein from *Actonomyces oris*, a bacteria involved in the formation of dental plaque. The structure of FimP shares both structure and topology with SpaA, another pilin surface protein from *C. diphteriae*. This suggests that in *G. vaginalis* there is a higher number of surface proteins of this kind [43].

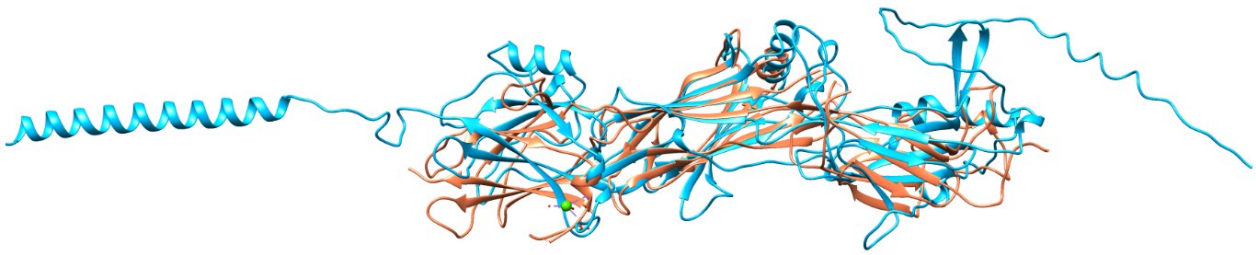

Figure 15: structural comparison made with MatchMaker between sequence V1151576 (blue) and the chain B of 4HSS (orange) from SpaD, a pilus protein from C. diphtheriae. The aminoacidic identity between the two structures is 26.68%

Another cell adhesion protein found is an adhesin capable to bind sialic acid, found also among *L. iners* hits. This adhesin comes from *Streptococcus sanguignis* and is capable of recognizing sialylated glycans on saliva, platelets, and plasma glycoproteins. The presence of repeated tandem domains suggest more than one type of binding mechanism and could have a role in helping pathogens adapt to different host receptors [44]. Is it possible to note that even though the structural superimposition between the reference structure and the predicted structure is not extensive, the single domains present in both structures have a similar folding.

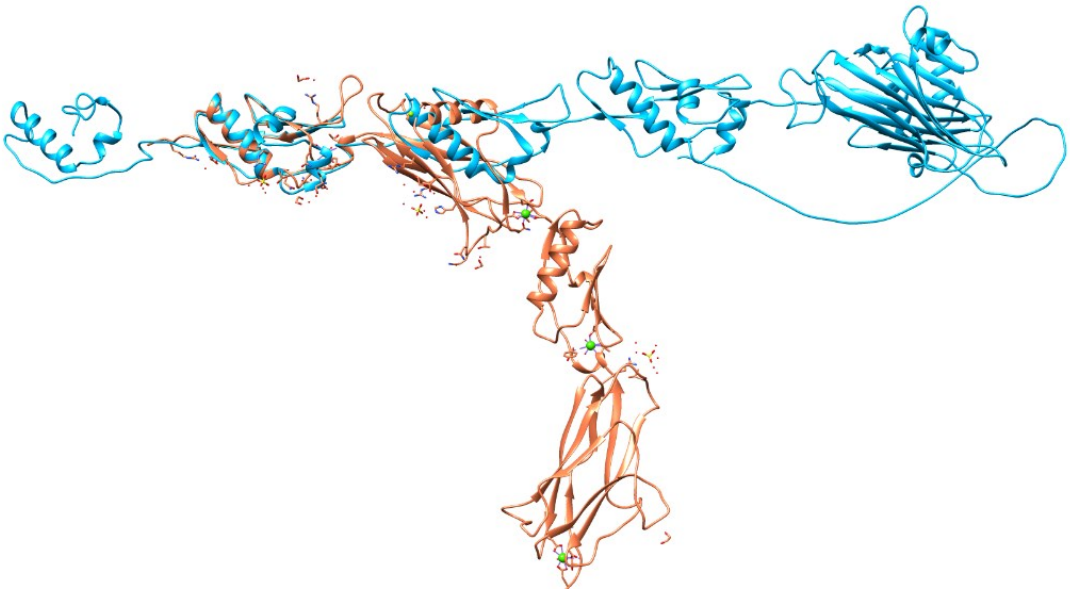

Figure 16: structural comparison made with MatchMaker between sequence V1265848 (blue) and the chain A of 6VS7 (orange), an adhesin from S. sanguignis. The aminoacidic identity between the two structures is 19.07%

To conclude, a good domain conservation was found with a surface protein from *Streptococcus agalactiae,* containing tandemly arrayed consecutive Rib domains. Rib domains are commonly found in invasive strains, and the variability in Rib domain number would result in differential projection of an N-terminal host-colonization domain from the bacterial surface. This confirms the strong presence in *G. vaginalis* of invasion-related proteins that may confer a strong adhesion to the vaginal surface environment.

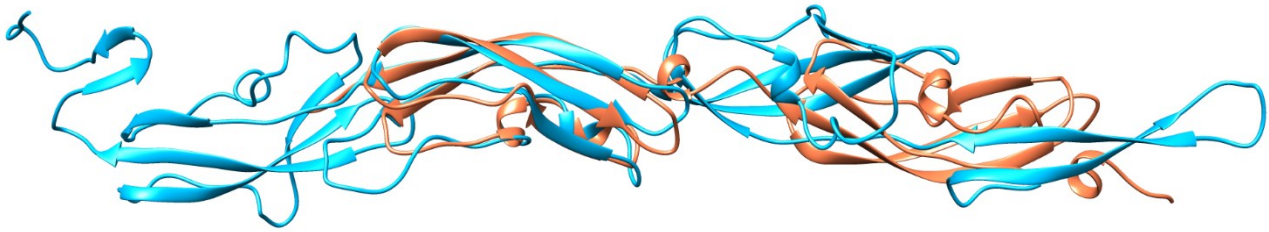

Figure 17: structural comparison made with MatchMaker between sequence V114661 (blue) and 6S5Y (orange), a cell adhesion protein from S. agalactie. The aminoacidic identity between the two structures is 29.54%

2.2.2 Antibiotic resistance

Several sequences have yielded a structural hit with the *Escherichia coli* MdfA transporter in complex with chloramphenicol, an antibiotic used to treat a wide range of infections. This is an antiporter from the major facilitator superfamily (MSD), a group of multidrug-resistance transporters. Interestingly, this transporter was found also in *L. crispatus* (reported but not shown), and could represent an adaptative response to extensive antibiotic treatments [45].

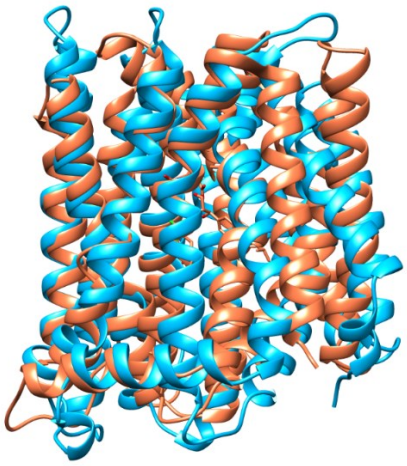

Figure 18: structural comparison made with MatchMaker between sequence V1172209 (blue) and 4ZOW (orange), a multidrug transporter from E. coli. The aminoacidic identity between the two structures is 16.88%

Finally, also a hit with an enzyme related to antibiotic resistance was found. It is the macrolide 2'-phosphotransferase MphH from *Brachybacterium faecium*, an anaerobic bacterium commonly found in soils and animals. MphH confers resistance to azithromycin, and the precited structure has a good structural alignment with the experimental structure in complex with the macrolide antibiotic [46]. This result suggests the presence of also in *G. vaginalis* of enzymes involved in antibiotic resistance pathways that could be the effect of several antibiotic treatments.

Figure 19: structural comparison made with MatchMaker between sequence V1442790 (blue) and chain A of 5UXD (orange), which is an antibiotic resistance-related enzyme in complex with azithromycin from B. faecium. The aminoacidic identity between the two structures is 17.67%

Discussion

The results of this work show that computational methods can be a valuable aid to better understanding the underlying biology of microbial communities. The functional annotation based on structural prediction has proven to be a powerful method, especially when considering at the aminoacidic identity between the unknown targets and the reference proteins: while the mean sequence identity is lower than 30%, the structure identity is always superior, demonstrating that *de novo* protein structure prediction techniques and comparison can extract heretofore unexploited structural information that can be further used as a guideline for the experimental validation.

In the results it was possible to observe some key differences between the three species targeted. The cell adhesion in *L. crispatus* seems to be mostly similar to other beneficial bacterial species, while in *L. iners* and *G. vaginalis* it resembles more those belonging to more pathogenic species. This is surprising especially for *L. iners*, that belongs to the same genus as *L. crispatus*, and it could partially explain its association with BV-associated bacteria and inflammatory processes, and the predisposition of this CST to shift into a diseased state. The presence of antibiotic resistance-related enzymes and transporters in all three species could be the result of antibiotic treatments. The development of this features could also play a role in determining the predominant CST type in women and affect the microbiota composition of the new-born child.
In *G. vaginalis*, in contrast with the *Lactobacillus* genus results, no bacteriocins and oxidative-stress related proteins were found, suggesting that this species might have different features, not yet discovered, to colonize, compete and eventually overcome the other bacterial species.

In conclusion, in this work it is presented a functional pipeline that improves the *de novo* annotation of unknown genes of dominant vaginal microbiota species applying protein structure prediction techniques. The subsequent functional annotation has allowed both an initial understanding of the coding content of each species analyzed and, most importantly, defined and strongly reduced the number of targets to potentially test in a following experimental validation. The future perspective of this work are the *in vitro* study and characterization of the most interesting target proteins found, which will give the opportunity to further understand and characterize the unique features of each species and the relationship between the vaginal microbiota and the human host.

Materials and methods

1. Genomes

The genomes used for the functional annotation were taken from the VIRGO database, a non-redundant gene catalogue of the vaginal microbiota [24]. VIRGO genes were viewed and sorted in the MetageneDB [25] database, developed in our group (intranet only). The sequences chosen for the protein structure predictions came from *Lactobacillus crispatus*, *Lactobacillus iners* and *Gardnerella vaginalis* and were filtered for ORF lengths between 600 and 2100 bp. The sequenceID of unannotated genes was exported from MetageneDB [25].

2. Python pipeline

The scripts used in the pipeline were developed with Python 3.7.
All the scripts and data produced can be found on GitHub.

2.1 Protein structure prediction with AlphaFold2

Extract_translate_batch.py: this script can find and extract the given sequenceIDs from multi-FASTA files, translates them into aminoacidic sequences and generates a sbatch file for each sequence. The outputs are a FASTA file for each sequence containing the translated aminoacidic sequence and a sbatch file to run AlphaFold2 on our HPC cluster. The arguments needed to run the script are a multi-FASTA file with all genome ORFs, a single column CSV containing the sequenceIDs chosen, an output directory for the FASTA aminoacidic sequences and the sbatch files. The output directory for the structures that will be generated by AlfaFold2 is also added, in order to pass this information in the generated sbatch files. The script is also able to handle Ns-containing sequences and filter them by a parameter given by the user. If Ns-containing sequences are chosen to be translated, the ambiguous triplet is translated as a glycine.

The protein structure prediction was done with the AlphaFold2 2.1.1 [22] implementation on the Institut Pasteur HPC cluster. We used the CPU instead of the GPU:
OPENMM_PLATFORM='CPU'
OPENMM_CPU_THREADS=12
ALPHAFOLD_JACKHMMER_N_CPU=12
ALPHAFOLD_HHBLITS_N_CPU=12

2.2 Results filtering

Collect_AF_results.py: this script collects the results from the AlphaFold2 output directories for each individual protein and generates a summary CSV file. The arguments needed to run the script are the AlphaFold2 output directory and the desired path and name of the CSV output file. The CSV file is organized by the sequenceID of the predicted protein, best model number (from 1 to 5 in this case), pLDDT score of the best model.

2.3 Functional prediction with Foldseek

ManageFoldseek.py: This script takes as input the CSV output file of *Collect_AF_results.py* and runs the structural comparison on the Foldseek [23] implementation on HPC clusters. Foldseek performs the comparison between the predicted structure and the available structures on the PDB and the UniProt databases. The arguments needed to run the script are the AlphaFold2 root results directory, the database directory, the CSV output file from *Collect_AF_results.py* path, the output directory to store the Foldseek search results, the directory where the predicted structures in PDB format will be copied and the pLDDT cut-off value chosen by the user. In the first part of the script, the best out of the original five predicted relaxed structures which also have a pLDDT score above the cut-off chosen are copied into a separate

directory set by the user. Only this subset of structures is analyzed by structural comparison against the chosen database(s). The output result of the Foldkseek search is a m8-format file.
The second part of the script handles the m8-format file and organizes the results in a CSV corresponding to the database(s) used. Each comparison is organized by sequenceID, number of total hits in the database, the first hit, the e-value of the first hit, the species of the first hit and additional notes. The script also has a built-in function that allows it to discard *Homo Sapiens* hits, going to the first non-*Homo sapiens* hit available, or if not possible, highlights the presence of only *Homo Sapiens* hits in the additional notes.

2.4 Results annotation

Annotate.ipynb: this script handles the CSV output file(s) of the previous script and is able to retrieve specifically from the PDB and UniProt databases the annotation attached to the best hit. The inputs needed are the paths of the *ManageFoldseek.py* CSV results files. The output is an Excel file, organized in three sheets (one for the PDB annotations, one for the UniProt annotation and another one for the combined annotations) Additional information is also compiled including annotation (categories) and the keywords related to best hits.

3. Structural comparison and percentage of aminoacidic identity on Chimera

To further analyze and validate the results obtained with the Python pipeline a structural comparison between the predicted structure with AlphaFold2 and the best hit on the PDB was done using Chimera 1.16 and its tool “Matchmaker”. For the comparison, the reference structure set was always the PDB structure, the default settings were used with the flagged “Show pairwise alignment” option.
The identity was assessed using the ClustalOmega alignment tool in the MultAlignViewer, by using the “Realign sequences” tool. The default settings were used. Then the percentage of identity was assessed with the “Percent identity” info on the realigned sequences window, using the default settings.

Supplementary material

Lactobacillus crispatus

Sequence(s)	PDB ID	Annotation	AA identity %	pLDDT	E-value
V1237124	4MT5	MUB: Mucus Binding Protein from *Limosilactobacillus reuteri* (Cell Adhesion, Protein Binding)	29.35%	87.72	$7.333E^{-11}$
V1033564 V1033753	5GT1	CbpA: Choline-binding protein from *Ligilactobacillus salivarus* (Cell Adhesion, Choline-binding protein) [also found in *L. iners*]	32.76%	88.08	$1.765E^{-14}$
V1806592	2G0D	NisC: Lantibiotic cyclase from *Lactococcus lactis* (Bacteriocin, Biosynthetic protein)	24.21%	92.79	$9.878E^{-15}$
V1884532	6GKW	R-type diffocin from *Clostridium difficile* (Bacteriocin, Structural protein)	14.04%	80.67	$3.639E^{-6}$
V1033523	4BS9	TruD: cyanobactin heterocyclase from *Prochloron sp.* (Bacteriocin, Hydrolase) [also found in *L. iners*]	16.24%	88.19	$2.233E^{-10}$
V1143191 V1808261	6GRG	Microcin synthetase chain C from *Escherichia coli* (Bacteriocin, Biosynthetic protein)	27.42%	89.21	$2.935E^{-16}$
V1808259 V1808260	6GRG	Microcin synthetase chain D from *Escherichia coli* (Bacteriocin, Biosynthetic protein) [also found in *L. iners*]	17.27%	85.89	$3.359E^{-08}$
V1163777 V1237198	4X6G	OxyR chain F from *Pseudomonas aeruginosa* (Transcription factor)	20.56%	93.38	$1.038E^{-17}$
V1828794	6T1Z	LmrP : polyspecific transport protein from *Lactococcus lactis* (Antibiotic resistance, Transport Protein) [also found in *G. vaginalis*]	17.16%	91.93	$6.472E^{-09}$
V1807941 V1143192	4ZOW	MdfA: multidrug transporter in complex with chloramphenicol from *Escherichia coli* (Antibiotic resistance, Transport protein) [also found in *G. vaginalis*]	16.37%	91.35	$1.442E^{-08}$

Table 4: relevant L. crispatus results. In the table is shown the VIRGO database sequence(s) ID, the Protein Data Bank (PDB ID) of the best hit, the biological annotation and other additional information (if available), the aminoacidic identity between the first sequenceID structure and PDB structure, the pLDDT score of the best AlphaFold2 generated structure of the first sequenceID, and the Foldseek structural comparison E-value of the first sequenceID

Lactobacillus iners

Sequence(s)	PDB ID	Annotation	AA identity %	pLDDT	E-value
V1195929	3IRZ	UafA: Uro-adherence factor from *Staphylococcus saprophyticus* (Cell Adhesion)	16.32%	89.67	$5.169E^{-14}$
V1506282 V1314836 V1420705 V1500278	7AVH	SHIRT (Streptococcus High Identity Repeats in Tandem) domains 3-4 from a cell surface protein from *Streptococcus* (Cell Adhesion) [also found in *G. vaginalis*]	27.54%	84.90	$1.109E^{-08}$
V1035134	5GT1	CbpA: Choline-binding protein from *Ligilactobacillus salivarus* (Cell Adhesion, Choline-binding protein) [also found in *L. crispatus*]	36.78%	81.75	$3.898E^{-13}$
V1880119 V1445426	6VS7	Sialic acid binding region from *Streptococcus sanguignis* (Cell Adhesion) [also found in *G. vaginalis*]	21.76%	87.93	$5.917E^{-14}$
V1310512	5MH4	Thioredoxin reductase from *Lactococcus lactis* (Oxidoreductase)	19.49%	91.85	$3.896E^{-17}$
V1547326	4BS9	TruD: cyanobactin heterocyclase from *Prochloron sp.* (Bacteriocin, Hydrolase) [also found in *L. crispatus*]	15.25%	89.38	$3.346E^{-11}$
V1046759	6GRG	Microcin synthetase chain C from *Escherichia coli* (Bacteriocin, Biosynthetic protein) [also found in *L. crispatus*]	15.24%	92.51	$4.204E^{-08}$
V1008909	2XF3	Cephalosporine esterase chain B from *Streptomyces clavuligerus* (Antibiotic resistance, Hydrolase)	15.21%	93.06	$3.697E^{-15}$

Table 5: relevant L. iners results. In the table is shown the VIRGO database sequence(s) ID, the Protein Data Bank (PDB ID) of the best hit, the biological annotation and other additional information (if available), the aminoacidic identity between the first sequenceID structure and PDB structure, the pLDDT score of the best AlphaFold2 generated structure of the first sequenceID, and the Foldseek structural comparison E-value of the first sequenceID

Gardnerella vaginalis

Sequence(s)	Ns %	PDB ID	Annotation	AA identity %	pLDDT	E-value
V1151576 V1010633 V1011329 V1036177 V1036344 V1051511 V1054450 V1070309 V1081122 V1099329 V1099572 V1102932 V1103156 V1103591 V1105705 V1116704 V1137006 V1138767 V1139526 V1145773 V1151928 V1152048 V1152183 V1152244 V1152295 V1154184 V1155369 V1156061 V1167326 V1171121 V1171839 V1172319 V1172422 V1175888 V1180994 V1190746 V1191089 V1191419 V1191820 V1198643	N/A	4HSS	Major Pilin SpaD chain B from *Corynebacterium diphtheriae* (Cell Adhesion)	26.68%	88.52	$8.063E^{-28}$
V1008624 V1014161 V1036358 V1051534 V1051822 V1070435 V1072094 V1081477 V1085966 V1098962 V1099526 V1103443 V1103497 V1103594 V1106264	N/A	3UXF (FimP) 3HR6 (SpaA)	FimP: fimbrial protein from *Actinomyces oris* (Cell Adhesion) SpaA: major pilin from *Corynebacterium diphtheriae*	22.77% 21.56%	87.06	$8.034E^{-28}$ (3UXF) $3.008E^{-22}$ (3HR6)

V1136956 V1138953 V1139070 V1151597 V1152143 V1156777 V1156780 V1158508 V1166825 V1170707 V1171043 V1175164 V1176869 V1191492 V1195031 V1195166 V1198688 V1199718 V1199896						
V1054309	N/A	7AVH	SHIRT (Streptococcus High Identity Repeats in Tandem) domains 3-4 from a cell surface protein from *Streptococcus* (Cell Adhesion) [also found in *L. iners*]	24.55%	84.65	$5.359E^{-09}$
V1265848 V1117267 V1152146 V1037691 V1395126 V1315205 V1383900	0.5 %	6VS7	Sialic acid binding region chain A from *Streptococcus sanguignis* (Cell Adhesion) [also found in *L. iners*]	19.07%	87.52	$1.636E^{-10}$
V1265896 V1347811 V1265894	2.3 %			21.52%	88.94	$2.775E^{-08}$
V1146621 V1037871 V1141746 V1146749 V1364246 V1364652 V1364147		6S5Y	Rib2R: tandemly arrayed consecutive Rib domains chain F from *Streptococcus agalactiae*	29.54%	90.53	$1.104E^{-13}$
V1364511	0.1 %			31.90%	91.10	$1.602E^{-12}$
V1054366 V1244703 V1380848 V1167014 V1430781	N/A	6T1Z	LmrP: polyspecific transport protein from *Lactococcus lactis* (Antibiotic resistance, Transport Protein) [also found in *L. crispatus*]	17.27%	88.11	$2.713E^{-08}$
V1172209 V1176154 V1079381 V1083136 V1246319 V1238925 V1356867 V1115070 V1242436 V1195087	N/A	4ZOW	MdfA: multidrug transporter in complex with chloramphenicol from *Escherichia coli* (Antibiotic resistance, Transport protein) [also found in *L. crispatus*]	16.88%	89.75	$3.303E^{-09}$

V1442790 V1313365	N/A	5UXD	MphH: Macrolide 2'-phosphotransferase in complex with azithromycin chain A from *Brachybacterium faecium* (Antibiotic resistance, Transferase)	17.67%	82.76	$2.322e^{-12}$

Table 6: relevant G. vaginalis results. In the table is shown the VIRGO database sequence(s) ID, the number of N nucleotides in the VIRGO sequences (if available), Protein Data Bank (PDB) ID of the best hit, the biological annotation and other additional information (if available), the aminoacidic identity between the first sequenceID structure and PDB structure, the pLDDT score of the best AlphaFold2 generated structure of the first sequenceID and of the models built from Ns-containing sequences, and the Foldseek structural comparison E-value of the first sequenceID and of the Ns-containing sequences

Bibliography

1. Wang, B., Yao, M., Lv, L., Ling, Z. & Li, L. The Human Microbiota in Health and Disease. *Engineering* **3**, 71–82 (2017).

2. Gupta, S., Kakkar, V. & Bhushan, I. Crosstalk between Vaginal Microbiome and Female Health: A review. *Microb. Pathog.* **136**, (2019).

3. Kaoutari, A. El, Armougom, F. & Henrissat, B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota Carbohydrate-active enzymes View project Rickettsia genomics View project. (2013) doi:10.1038/nrmicro3050.

4. Kennedy, M. S. & Chang, E. B. The microbiome: Composition and locations. in *Progress in Molecular Biology and Translational Science* vol. 176 1–42 (2020).

5. Smith, S. B. & Ravel, J. The vaginal microbiota, host defence and reproductive physiology. *J. Physiol.* **595**, 451–463 (2017).

6. De Seta, F., Campisciano, G., Zanotta, N., Ricci, G. & Comar, M. The vaginal community state types microbiome-immune network as key factor for bacterial vaginosis and aerobic vaginitis. *Front. Microbiol.* **10**, 2451 (2019).

7. Ma, B., Forney, L. J. & Ravel, J. Vaginal Microbiome: Rethinking Health and Disease. *Annu. Rev. Microbiol* **66**, 371–389 (2012).

8. Kwon, M. S. & Lee, H. K. Host and Microbiome Interplay Shapes the Vaginal Microenvironment. **13**, 1–13 (2022).

9. Mirmonsef, P. *et al.* Free Glycogen in Vaginal Fluids Is Associated with Lactobacillus Colonization and Low Vaginal pH. *PLoS One* **9**, e102467 (2014).

10. Gupta, P., Singh, M. P. & Goyal, K. Diversity of Vaginal Microbiome in Pregnancy: Deciphering the Obscurity. *Front. Public Heal.* **8**, 326 (2020).

11. Vallor, A. C., Antonio, M. A. D., Hawes, S. E. & Hillier, S. L. Factors associated with acquisition of, or persistent colonization by, vaginal lactobacilli: role of hydrogen peroxide production. *J. Infect. Dis.* **184**, 1431–1436 (2001).

12. Mitchell, C., Fredricks, D., Agnew, K. & Hitti, J. Hydrogen peroxide-producing lactobacilli are associated with lower levels of vaginal interleukin-1β, independent of bacterial vaginosis. *Sex. Transm. Dis.* **42**, 358–363 (2015).

13. Spear, G. T. *et al.* Human α-amylase present in lower-genital-tract mucosal fluid processes glycogen to support vaginal colonization by Lactobacillus. *J. Infect. Dis.* **210**, 1019–1028 (2014).

14. Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4680–4687 (2011).

15. Chen, X., Lu, Y., Chen, T. & Li, R. The Female Vaginal Microbiome in Health and Bacterial Vaginosis. *Front. Cell. Infect. Microbiol.* **11**, (2021).

16. Morrill, S., Gilbert, N. M. & Lewis, A. L. Gardnerella vaginalis as a Cause of Bacterial Vaginosis: Appraisal of the Evidence From in vivo Models. *Front. Cell. Infect. Microbiol.* **10**, 168 (2020).

17. Qin, H. & Xiao, B. Research Progress on the Correlation Between Gardnerella Typing and Bacterial Vaginosis. *Front. Cell. Infect. Microbiol.* **12**, 324 (2022).

18. Brooks, J. P. *et al.* Changes in vaginal community state types reflect major shifts in the microbiome. (2017) doi:10.1080/16512235.2017.1303265.

19. Ahire, J. J. *et al.* In Vitro Assessment of Lactobacillus crispatus UBLCp01, Lactobacillus gasseri UBLG36, and Lactobacillus johnsonii UBLJ01 as a Potential Vaginal Probiotic Candidate. *Probiotics Antimicrob. Proteins* (2021) doi:10.1007/s12602-021-09838-9.

20. Mendes-Soares, H., Suzuki, H., Hickey, R. J. & Forneya, L. J. Comparative functional genomics of Lactobacillus spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. *J. Bacteriol.* **196**, 1458–1470 (2014).

21. Hvidsten, T. R. *et al.* A Comprehensive Analysis of the Structure-Function Relationship in Proteins Based on Local Structure Similarity. *PLoS One* **4**, (2009).

22. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nat. 2021 5967873* **596**, 583–589 (2021).

23. Kempen, M. van *et al.* Foldseek: fast and accurate protein structure search. *bioRxiv* 2022.02.07.479398 (2022) doi:10.1101/2022.02.07.479398.

24. Ma, B. *et al.* A comprehensive non-redundant gene catalog reveals extensive within-community intraspecies diversity in the human vagina. *Nat. Commun.* **11**, (2020).

25. MetageneDB. https://metagenedb-dev.pasteur.cloud/.

26. AlphaFold2 documentation. https://alphafold.ebi.ac.uk/faq.

27. Zeng, Z., Zuo, F. & Marcotte, H. Putative adhesion factors in vaginal Lactobacillus gasseri DSM 14869: Functional characterization. *Appl. Environ. Microbiol.* **85**, 800–819 (2019).

28. Etzold, S. *et al.* Structural basis for adaptation of lactobacilli to gastrointestinal mucus. *Environ. Microbiol.* **16**, 888–903 (2014).

29. Wang, R. *et al.* The Adhesion of Lactobacillus salivarius REN to a Human Intestinal Epithelial Cell Line Requires S-layer Proteins. *Sci. Reports 2017 71* **7**, 1–10 (2017).

30. Matsuoka, E. *et al.* Crystal structure of the functional region of Uro-adherence factor A from Staphylococcus saprophyticus reveals participation of the B domain in ligand binding. (2010) doi:10.1002/pro.573.

31. Whelan, F. *et al.* Periscope Proteins are variable-length regulators of bacterial cell surface interactions. *Proc. Natl. Acad. Sci. U. S. A.* **118**, 2101349118 (2021).

32. Dobson, A., Cotter, P. D., Paul Ross, R. & Hill, C. Bacteriocin production: A probiotic trait? *Appl. Environ. Microbiol.* **78**, 1–6 (2012).

33. Stoyancheva, G., Marzotto, M., Dellaglio, F. & Torriani, S. Bacteriocin production and gene sequencing analysis from vaginal Lactobacillus strains. *Arch. Microbiol.* **196**, 645–653 (2014).

34. Li, B. *et al.* Structure and mechanism of the lantibiotic cyclase involved in nisin biosynthesis. *Science (80-.).* **311**, 1464–1467 (2006).

35. Jahn, D. *et al.* Crystal Structures of R-Type Bacteriocin Sheath and Tube Proteins CD1363 and CD1364 From Clostridium difficile in the Pre-assembled State. (2018) doi:10.3389/fmicb.2018.01750.

36. Koehnke, J. *et al.* The cyanobactin heterocyclase enzyme: A processive adenylase that operates with a defined order of reaction. *Angew. Chemie - Int. Ed.* **52**, 13991–13996 (2013).

37. Ghilarov, D. *et al.* Architecture of Microcin B17 Synthetase: An Octameric Protein Complex Converting a Ribosomally Synthesized Peptide into a DNA Gyrase Poison. *Mol. Cell* **73**, 749-762.e5 (2019).

38. Strus, M., Brzychczy-Włoch, M., Gosiewski, T., Kochan, P. & Heczko, P. B. The in vitro effect of hydrogen peroxide onvaginal microbial communities. *FEMS Immunol. Med. Microbiol.* **48**, 56–63 (2006).

39. Jo, I. *et al.* Structural details of the OxyR peroxide-sensing mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6443–6448 (2015).

40. Debruycker, V. *et al.* An embedded lipid in the multidrug transporter LmrP suggests a mechanism for polyspecificity. *Nat. Struct. Mol. Biol. 2020 279* **27**, 829–835 (2020).

41. Valegård, K. *et al.* Structural and mechanistic studies of the orf12 gene product from the clavulanic acid biosynthesis pathway. *Acta Crystallogr. D. Biol. Crystallogr.* **69**, 1567–1579 (2013).

42. Kang, H. J. *et al.* Biological Crystallography A slow-forming isopeptide bond in the structure of the major pilin SpaD from Corynebacterium diphtheriae has implications for pilus assembly. doi:10.1107/S1399004714001400.

43. Persson, K., Esberg, A. & Claesson, R. The Pilin Protein FimP from Actinomyces oris: Crystal Structure and Sequence Analyses. *PLoS One* **7**, 48364 (2012).

44. Stubbs, H. E. *et al.* Tandem sialoglycan-binding modules in a Streptococcus sanguinis serine-rich repeat adhesin create target dependent avidity effects. *J. Biol. Chem.* **295**, 14737–14749 (2020).

45. Heng, J. *et al.* Substrate-bound structure of the E. coli multidrug resistance transporter MdfA. *Cell Res.* **25**, 1060–1073 (2015).

46. Pawlowski, A. C. *et al.* The evolution of substrate discrimination in macrolide antibiotic resistance enzymes. *Nat. Commun. 2017 91* **9**, 1–12 (2018).