Proteome-wide prediction of bacterial carbohydrate-binding proteins as a tool for understanding commensal and pathogen colonisation of the vaginal microbiome

François Bonnardel[1,2,3], Stuart M. Haslam[4,5], Anne Dell[4,5], Ten Feizi[5,6], Yan Liu[5,6], Virginia Tajadura-Ortega[5,6], Yukie Akune[6], Lynne Sykes[5,7,8], Phillip R. Bennett[5,7,8,9], David A. MacIntyre[5,7,9*], Frédérique Lisacek[2,3,10*] and Anne Imberty[1]*

[1] University Grenoble Alpes, CNRS, CERMAV, Grenoble, France

[2]Swiss Institute of Bioinformatics, Geneva, Switzerland

[3]Computer Science Department, UniGe, Geneva, Switzerland

[4]Department of Life Sciences, Imperial College London, London, UK

[5]March of Dimes European Prematurity Research Centre, Imperial College London, UK

[6]Glycosciences Laboratory, Department of Metabolism Digestion and Reproduction, Imperial College London, London, UK

[7]Imperial College Parturition Research Group, Division of the Institute of Reproductive and Developmental Biology, Department of Metabolism Digestion and Reproduction, Imperial College London, London, UK

[8]Queen Charlotte's Hospital, Imperial College Healthcare NHS Trust, London, UK

[9]Tommy's National Centre for Miscarriage Research, Imperial College London, London, UK

[10]Section of Biology, UniGe, Geneva, Switzerland

*Corresponding Authors:

Email : d.macintyre@imperial.ac.uk, frederique.lisacek@sib.swiss, anne.imberty@cermav.cnrs.fr

Keywords : Lectins, glycans, vaginal microbiota, pregnancy, infection, bioinformatics

Abstract

Lectins, such as adhesins and toxins, are carbohydrate-binding proteins that recognise glycans of cells and their secretions. While mediation of microbe-microbe and microbe-host interactions by lectins has long been recognised in the lung and gut, little is known about those in the vagina, where such interactions are implicated in health and various disease states. These include sexually transmitted infections, cervical cancer and poor pregnancy outcomes such as preterm birth. In this study, the curated UniLectin3D database was used to establish a lectin classification based primarily on taxonomy and protein 3D structure. The resulting 109 lectin classes were characterised by specific Hidden Markov Model (HMM) profiles. Screening of microbial genomes in the UniProt and NCBI NR sequence databases resulted in identification of >100 000 predicted bacterial lectins available at unilectin.eu/bacteria. Screening of the complete genomes of 90 isolates from 21 vaginal bacterial species showed that the predicted lectomes (ensemble of predicted lectins) of *Lactobacilli* associated with vaginal health are substantially less diverse than those of pathogens and pathobionts. Both the number of predicted bacterial lectins, and their specificities for carbohydrates correlated with pathogenicity. This study provides new insights into potential mechanisms of commensal and pathogen colonisation of the reproductive tract that underpin health and disease states.

Author Summary

Microbes play an important role in human health and disease. Bacteria use protein receptors called lectins to anchor to specific sugars (i.e. glycans) decorating the surface of proteins and cells. While these have been extensively studied in the mouth and gut, much less is known about how bacteria attach and colonise the lower female reproductive tract. This limits our understanding of how they contribute to sexually transmitted infections, cervical cancer and preterm birth. To address this, we designed and implemented a bioinformatics workflow to identify and classify novel lectins in 21 vaginal bacterial species implicated in reproductive tract health and disease. Our results show that species associated with infection and inflammation produce a larger variety of lectins thus enabling them to potentially bind a wider array of glycans in the vagina. These findings provide new targets for the development of compounds designed to prevent pathogen colonisation or encourage growth of commensal species.

Introduction

Microbiota-host interactions within different ecological niches of the human body are critical determinants of health and disease states [1]. At mucosal surface interfaces, microbial and host cells, as well as non-cellular components of the mucosa, present an exceptionally complex array of attachment and recognition sites for microbiota, many of which are carbohydrate sequences displayed on extensively glycosylated mucin-type glycoproteins rich in O-glycans. The diverse populations of glycans provide recognition sites for microbial adhesins that have the ability to distinguish the various motifs displayed. Bacteria also produce glycosylhydrolases and other enzymes that facilitate the use of secreted mucins as primary carbon sources for energy metabolism [2, 3]. The abilities of microbes to specifically recognise, attach and adhere to cellular and non-cellular surfaces are thus key aspects of commensal and pathogenic colonisation and are mediated by receptors, such as lectins and carbohydrate-binding modules (CBMs) [3-6].

Lectins are ubiquitous proteins of non-immune origin that bind to a variety of carbohydrates without modifying them [7]. Through their interactions with glycoproteins and glycolipids via the oligosaccharides, lectins play crucial roles in cell-cell communication, signalling pathways and immune responses [8]. Bacterial lectins may be incorporated into multiprotein organelles, such as fimbriae (pili) or flagellae and participate in the mediation of host recognition and adhesion [9]. In pathogenic species, lectins may also be toxin subunits targeting a toxic catalytic unit towards subcellular components that display specific glycoconjugates [10]. Soluble lectins are also expressed as virulence factors by opportunistic bacteria [11] and can alter dynamics of glycolipids to induce the internalization of whole bacteria into host cells [12]. Bacterial lectins have also been shown to directly impair immune signalling and repair pathways and are implicated in the formation of biofilms [13].

The role of lectins and their ligands in shaping microbial niches within the human body is increasingly recognised, particularly at mucosal interfaces including the gut [3, 14, 15] and oral cavity [16]. However, much less is known about the role of lectins in shaping microbial niches in the lower female reproductive tract, which play a key role in shaping health and disease throughout a woman's life span [17]. Colonisation in the vagina by *Lactobacillus* species has been consistently considered a hallmark of health [18, 19], whereas *Lactobacillus* deplete, high diversity vaginal microbiomes enriched for potential pathogens are characteristic of bacterial vaginosis and are associated with increased risk of sexually transmitted infections (STIs) acquisition [20, 21], progression of cervical cancer [22] and adverse pregnancy outcomes such as miscarriage and preterm birth [23-26]. A key component of the vaginal mucosa are highly glycosylated mucins that are derived from the mucin-secreting glands of the cervix [27]. Alteration of terminal glycan residues by microbially secreted sialidases and sulphatases modulate the physical and immunological properties of the vaginal mucosa [28]. Vaginal pathogens such as *Gardnerella vaginalis, Trichomonas vaginalis*, *Prevotella* and *Ureaplasma* species are capable of degrading secretory IgA [29-32]. Moreover, specific strains of *Streptococcus agalactiae* (group B streptococcus) secrete hyaluronidases that degrade cervical hyaluronic acid into disaccharide fragments dampening host immune activation through inhibition of Toll-like receptors, which may contribute to preterm birth via ascending infection [33]. *Streptococcus agalactiae* can also implement a negative signalling mechanism known as sialoglycan mimicry to evade detection and phagocytosis by neutrophils; this is through recognition of terminal α2-3-linked sialic acids on the bacteria as 'self' glycans by the neutrophil lectin Siglec-9 [34].

Despite their important role in infection and pathogenicity, the contribution of bacterial lectins to health and disease states is yet to be fully appreciated. This is partly due to the limited annotation and characterisation of the lectins in protein and proteome databases, that precludes predictions of the diversity, structure and function of the lectins. In recent years, this has begun to be addressed through the development of databases for structural and functional glycobiology [35, 36]. Among these, UniLectin3D provides 3D structures of more than 2500 lectins and their complexes with carbohydrates [37] and sites within UniLectin, a platform dedicated to the curation and collection of lectin knowledge. In this study we describe how manual selection of lectin domains in 3D structures permits the identification of lectin classes characterised by fold similarity and minimum thresholds of sequence identity. We show that defined amino acid sequence motifs and profiles characterising each lectin class can be used to screen proteomes and translated genomes to identify unannotated lectins. Comparison of these lectins across different vaginal microbiota strains provides new insights into the potential mechanisms by which commensal and pathogen colonisation associate with physiological and pathological conditions in the lower reproductive tract.

Results

Structural classification of bacterial lectins in Unilectin3D

Structural classification of known lectins curated in the Unilectin3D database (www.unilectin.eu/unilectin3D/) was first performed on the basis of differences in fold, i.e. structure of the protein backbone and then on amino acid sequences at 20% of sequence similarity for lectin classes and at 70% of sequence similarity for lectin families. This led to the identification of 35 different folds and 109 lectin classes (S1 Table) derived from a total of 2483 structural lectin entries that primarily originated from plant and animal sources. However, bacterial lectins from 46 different species accounted for approximately 20% of database entries (495/2483), which were distributed among 19 different folds (Figure 1) and 37 classes (S1 Table).

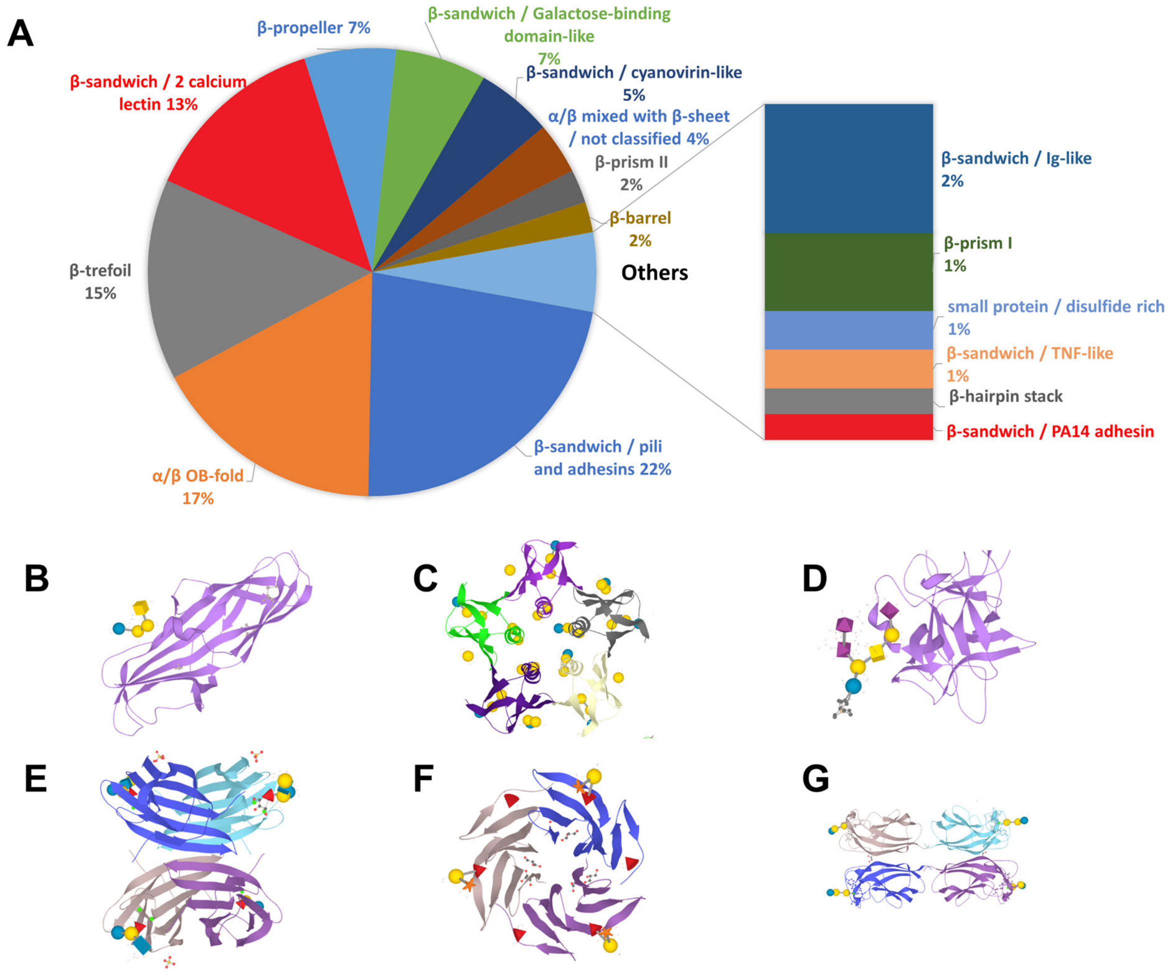

Figure 1: Structural classification of bacterial lectins. (A) Distribution of bacterial lectin folds derived from the UniLectin3D database. From the analysis of fold distribution of bacterial lectin crystal

structures, the six most frequent fold are represented: (B) Pili and adhesins: 1J8R PapG *Escherichia coli*, (C) OB fold: 1BOS SLT-1 / STX-1 *E. coli*, (D) *β*-trefoil: 1FV2 TeNT *Clostridium tetani*, (E) 2 calcium lectin: 1W8F LecB / PA-IIL, RSIIL *Pseudomonas aeruginosa,* (F) *β*-propeller: 2BS6 RSL, BambL *Ralstonia solanacearum*, and (G) Galactose binding domain-like: 2VXJ LecA / PA-IL *Pseudomonas aeruginosa*. 3D structures were generated using LiteMol [38] with monosaccharides in binding sites represented using Symbol Nomenclature for Glycans (SNFG) [39].

The analysis of fold distribution in bacterial lectin crystal structures showed an over-representation of β-sheet containing folds, which were common to adhesins and toxins including previously described pili adhesins, such as FimH in uro-pathogenic *Escherichia coli*, the oligomer-binding (OB) fold of the cholera toxin binding domain, the β-sandwich of LecA and LecB in *Pseudomonas aerigunosa* and the β-trefoil of the recognition domain in clostridial neurotoxins. While the majority of lectin folds were shared between sources of origin, classes of pili adhesins and AB_5 toxins were found to be specific to bacteria.

Prediction of lectin sequences from bacterial proteomes

Alignment of amino acid sequences in each of the 109 identified lectin classes led to the identification of 109 characteristic motifs of conserved residues. Profiles characterising each lectin class were generated with Hidden Markov Models (HMM), which were subsequently used to screen 130 million bacterial protein sequences from the UniProt database and over 168 million bacterial protein sequences from the NCBI RefSeq database derived from over 100 000 bacterial species. The TIM fold (named after triosephosphate isomerase) and Variable Lymphocyte Receptor folds are highly frequent in the resulting predictions. The TIM lectin class may arise from its high occurrence in hydrolases. Consequently, both of these lectin classes were excluded from whole proteome predictions. This resulted in the selection of 100 671 sequences as putative lectins in 10126 distinct bacterial species (reduced to 46 322 sequences in 6 425 distinct bacterial species when applying a score of 0.25). A web interface dedicated to the exploration of these bacterial lectin candidates is available at www.unilectin.eu/bacteria/.

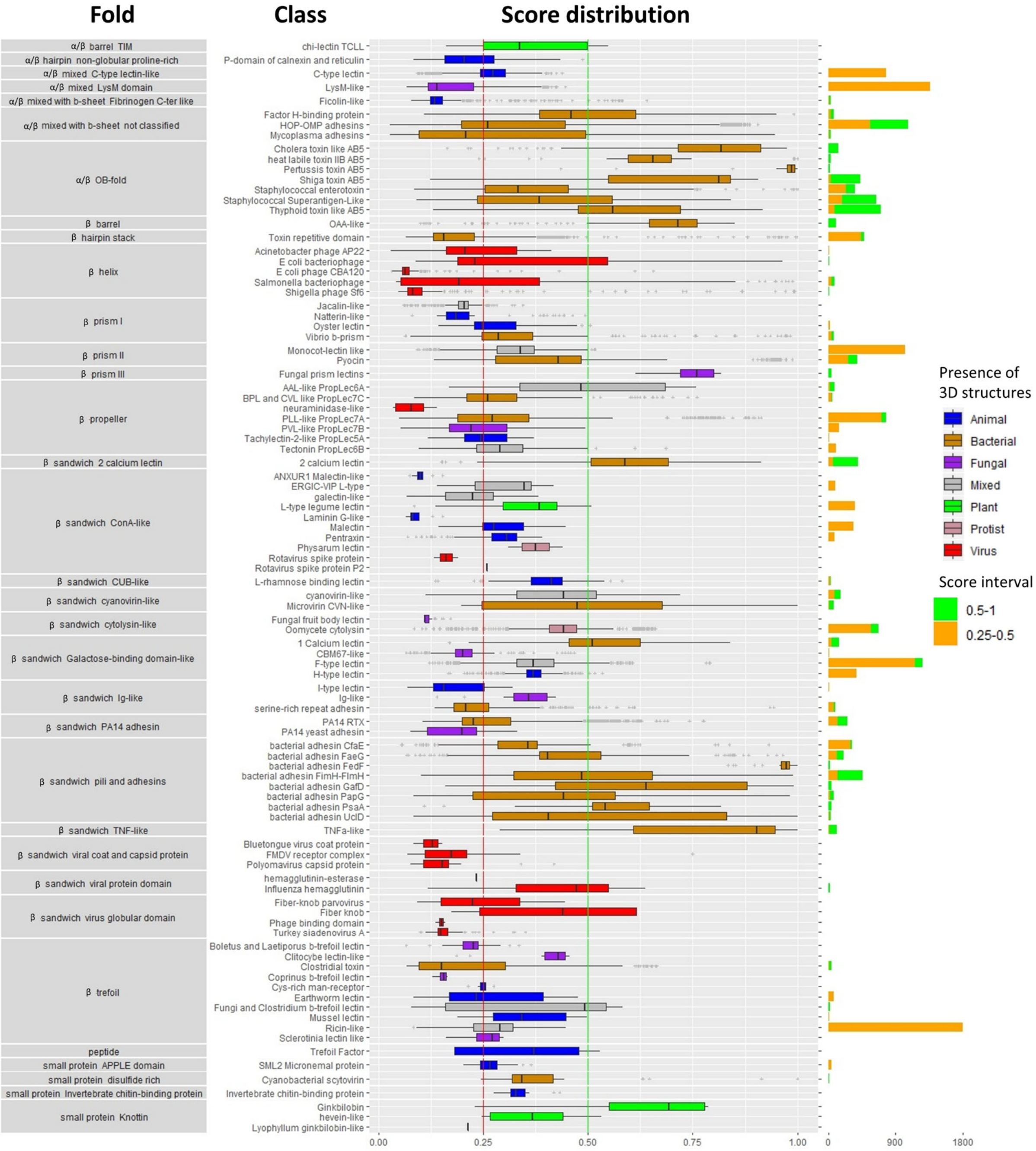

Figure 2: Distribution of structural fold-types within predicted lectin classes derived from 21 different bacterial genomes. Distributions of the predicted lectin classes are presented as horizontal box and whisker plots coloured on the basis of genome origin. The whisker plot represents the minimum, maximum, median, first quartile and third quartile in each class. Values approaching 1 are indicative of high sequence similarity to the reference motif. The predicted lectins in [0.25-0.5] and [0.5-1] score intervals are presented as bar graphs. The total number of predicted lectins in each class is listed in S1 Table.

Although 481 3D-structures of bacterial lectins were categorized into 37 classes, the screening results indicated that the putative lectins are predicted to occur in 97 out of the 107 identified classes (with a cutoff of 25% of sequence similarity with the reference) (S1 Table). Putative lectin sequences identified in each class, together with the distribution of the prediction scores to the original HMM motif, are presented in Figure 2. The fold distribution of predicted lectins differed from that obtained when using 3D structures generated from the UniLectin3D database with several classes comparatively over-represented, including the Ricin-like (β trefoil), the LysM domains (LysM fold), and the F-type lectins (β sandwich galactose binding domain like) (S1 Table). Each lectin domain is predicted by selecting the best fitted HMM. A score reflecting the sequence similarity is computed as the difference between the predicted lectin domain and the reference conserved motif. Lectins with the highest prediction scores per class were, as expected, of bacterial origin and included adhesins, AB5 toxins and calcium-dependent soluble lectins. However, the β-prism III fungal lectin was also found to have a high prediction score indicative of genetic exchange between bacteria and fungi. The majority of low scoring predictions (<0.25) reflective of low sequence similarity, were identified in viruses with the exception of the Influenza hemagglutinin, which contains a high abundance of sequences for the characteristic domain although not all are carbohydrate-binding. Lectins with mid-range (0.25-0.5) prediction scores were evenly distributed across multiple genome sources.

Identification and characterisation of vaginal microbiota lectins

We next obtained publicly available genome data for 90 vaginal bacterial strains classified on the basis of potential pathogenicity within the vaginal niche and having a known association with states of health or disease (S2 Table). Comparison of the lectomes, i.e. the predicted ensemble of lectins, highlighted major differences across species with pathobionts generally harbouring a higher diversity of lectin classes compared to commensals (Figure 3). Considering the low number of identified lectins, the TIM lectin and the Variable lymphocyte receptor classes were kept, despite a low probability of lectin activity. For example, the only predicted lectin consistently identified across *L. crispatus* isolates was LysM, a common domain involved in cell wall attachment in many different bacteria. Consistent with this, the LysM domain was predicted from the majority of examined vaginal microbial genomes but interestingly, was absent from *L. iners* and most *Provotella* strains.

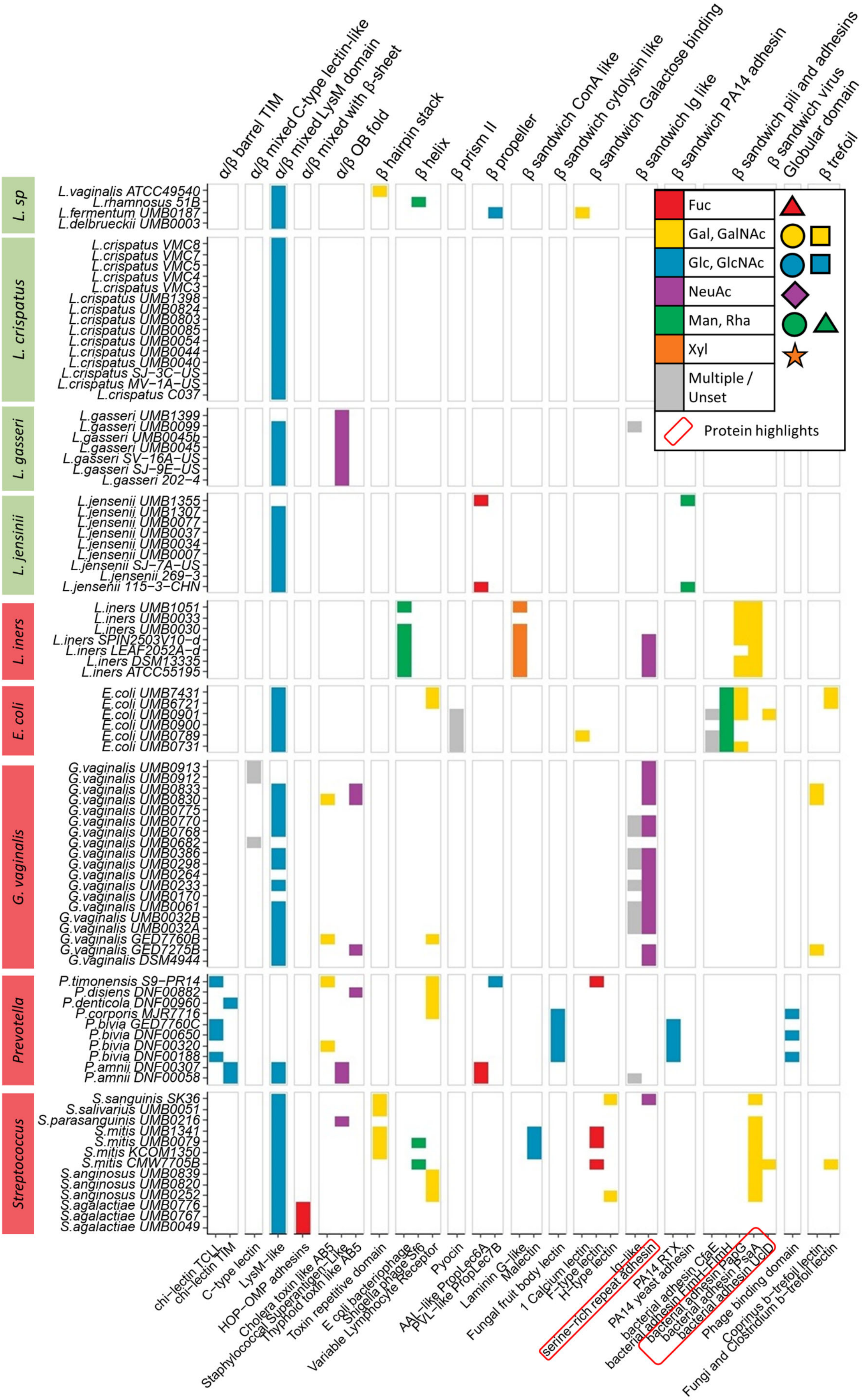
α/β barrel TIM
α/β mixed C-type lectin-like
α/β mixed LysM domain
α/β mixed with β-sheet
α/β OB fold
β hairpin stack
β helix
β prism II
β propeller
β sandwich ConA like
β sandwich cytolysin like
β sandwich Galactose binding
β sandwich Ig like
β sandwich PA14 adhesin
β sandwich pili and adhesins
β sandwich virus
Globular domain
β trefoil
L. sp
L.vaginalis ATCC49540
L.rhamnosus 51B
L.fermentum UMB0187
L.delbrueckii UMB0003
L. crispatus
L.crispatus VMC8
L.crispatus VMC7
L.crispatus VMC5
L.crispatus VMC4
L.crispatus VMC3
L.crispatus UMB1398
L.crispatus UMB0824
L.crispatus UMB0803
L.crispatus UMB0085
L.crispatus UMB0054
L.crispatus UMB0044
L.crispatus UMB0040
L.crispatus SJ−3C−US
L.crispatus MV−1A−US
L.crispatus C037
L. gasseri
L.gasseri UMB1399
L.gasseri UMB0099
L.gasseri UMB0045b
L.gasseri UMB0045
L.gasseri SV−16A−US
L.gasseri SJ−9E−US
L.gasseri 202−4
L. jensinii
L.jensenii UMB1355
L.jensenii UMB1307
L.jensenii UMB0077
L.jensenii UMB0037
L.jensenii UMB0034
L.jensenii UMB0007
L.jensenii SJ−7A−US
L.jensenii 269−3
L.jensenii 115−3−CHN
L. iners
L.iners UMB1051
L.iners UMB0033
L.iners UMB0030
L.iners SPIN2503V10−d
L.iners LEAF2052A−d
L.iners DSM13335
L.iners ATCC55195
E. coli
E.coli UMB7431
E.coli UMB6721
E.coli UMB0901
E.coli UMB0900
E.coli UMB0789
E.coli UMB0731
G. vaginalis
G.vaginalis UMB0913
G.vaginalis UMB0912
G.vaginalis UMB0833
G.vaginalis UMB0830
G.vaginalis UMB0775
G.vaginalis UMB0770
G.vaginalis UMB0768
G.vaginalis UMB0682
G.vaginalis UMB0386
G.vaginalis UMB0298
G.vaginalis UMB0264
G.vaginalis UMB0233
G.vaginalis UMB0170
G.vaginalis UMB0061
G.vaginalis UMB0032B
G.vaginalis UMB0032A
G.vaginalis GED7760B
G.vaginalis GED7275B
G.vaginalis DSM4944
Prevotella
P.timonensis S9−PR14
P.disiens DNF00882
P.denticola DNF00960
P.corporis MJR7716
P.bivia GED7760C
P.bivia DNF00650
P.bivia DNF00320
P.bivia DNF00188
P.amnii DNF00307
P.amnii DNF00058
Streptococcus
S.sanguinis SK36
S.salivarius UMB0051
S.parasanguinis UMB0216
S.mitis UMB1341
S.mitis UMB0079
S.mitis KCOM1350
S.mitis CMW7705B
S.anginosus UMB0839
S.anginosus UMB0820
S.anginosus UMB0252
S.agalactiae UMB0776
S.agalactiae UMB0767
S.agalactiae UMB0049
chi−lectin TCLL
chi−lectin TIM
C−type lectin
LysM−like
HOP−OMP adhesins
Cholera toxin like AB5
Staphylococcal Superantigen−Like
Typhoid toxin like AB5
Toxin repetitive domain
E coli bacteriophage
Shigella phage Sf6
Variable Lymphocyte Receptor
Pyocin
AAL−like PropLec6A
PVL−like PropLec7B
Laminin G−like
Malectin
Fungal fruit body lectin
1 Calcium lectin
F−type lectin
H−type lectin
Ig−like
serine−rich repeat adhesin
PA14 RTX
PA14 yeast adhesin
bacterial adhesin CfaE
bacterial adhesin FimH−FimH
bacterial adhesin PapG
bacterial adhesin PsaA
bacterial adhesin UclD
Phage binding domain
Coprinus b−trefoil lectin
Fungi and Clostridium b−trefoil lectin
Fuc
Gal, GalNAc
Glc, GlcNAc
NeuAc
Man, Rha
Xyl
Multiple / Unset
Protein highlights

Figure 3. Heatmap of predicted lectomes from different vaginal commensal and pathobiont bacterial species classified by fold and class. Green species label represent commensal species and red species labels represent pathobiont species. Colours within each class of lectin reflects its main glycan specificity characterised by binding monosaccharides using standardised Symbol Nomenclature for Glycans (SNFG) (https://www.ncbi.nlm.nih.gov/glycans/snfg.html). The lectin class circled in red are further discussed in the results due to their particular presence in *L.iners.*

Predicted lectins of *L. iners* could be mapped to five different classes: *E.coli* bacteriophage β-helix, laminin G-like, adhesin domain of two type 1 pili PapG and PsaA (chaperon-usher-assembled, CUP) and the adhesin domain of serine rich repeat protein (SRRP), which was also prominently observed in *G. vaginalis* species and in a *Streptococcus sanguinis* strain. Up to 10 different lectin classes were predicted from other *Streptococcus* species although *S. agalactiae* (also known as Group B Streptococcus), which is a pathogen known to cause sepsis, pneumonia and meningitis in newborn babies, was the only vaginal species predicted to produce Outer Membrane Protein (OMP) adhesins.

Identification and characterisation of vaginal microbiota carbohydrate binding modules

The screening strategy was extended to the prediction of carbohydrate binding modules (CBMs), small domains that are generally associated with carbohydrate modifying enzymes, often involved in microbial digestion of mucin glycans. A few of particular interest including CBM34, CBM41 and CBM48, which are specific for glucose containing polysaccharides (e.g. amylose, glycogen) and generally act as binding modules for amylases and related enzymes, were predicted consistently across almost all vaginal species (Figure 4).

While the majority of CBMs have been characterised as enzyme-associated domains in plant polysaccharides, two human-specific CBMs were observed in the dataset (S3 Table). The first is CBM40 considered as sialic acid-specific since it has been identified in association with a bacterial sialidase [40]. In the dataset analysed here, it is predicted to occur only in *L. iners* pathobiont species, *S. mitis* and some *Prevotella* species. Considering the earlier observation regarding the predicted SRR adhesin domain, the sialic acid binding ability appears to correlate mainly with lectins and CBMs present in the lectomes of pathobiont bacteria. The second domain of interest is CBM47, shown to be fucose-specific in the lectin regulatory domain of a cholesterol-dependent cytolysin present in some *S. mitis* strains (Feil, Lawrence et al. 2012). It shares structure and sequence similarity with the F-lectins from fishes [41]. In our study, this fucose-binding module is identified in *S. mitis* as well as in some pathobionts, i.e. *Provotella* and *L. iners*. Furthermore, the *L. iners* lectome contained two predicted adhesins, PapG and PsaA, as well as CBM60, which bind to galactose epitope occurring on human Gb3 gangliosides [42].

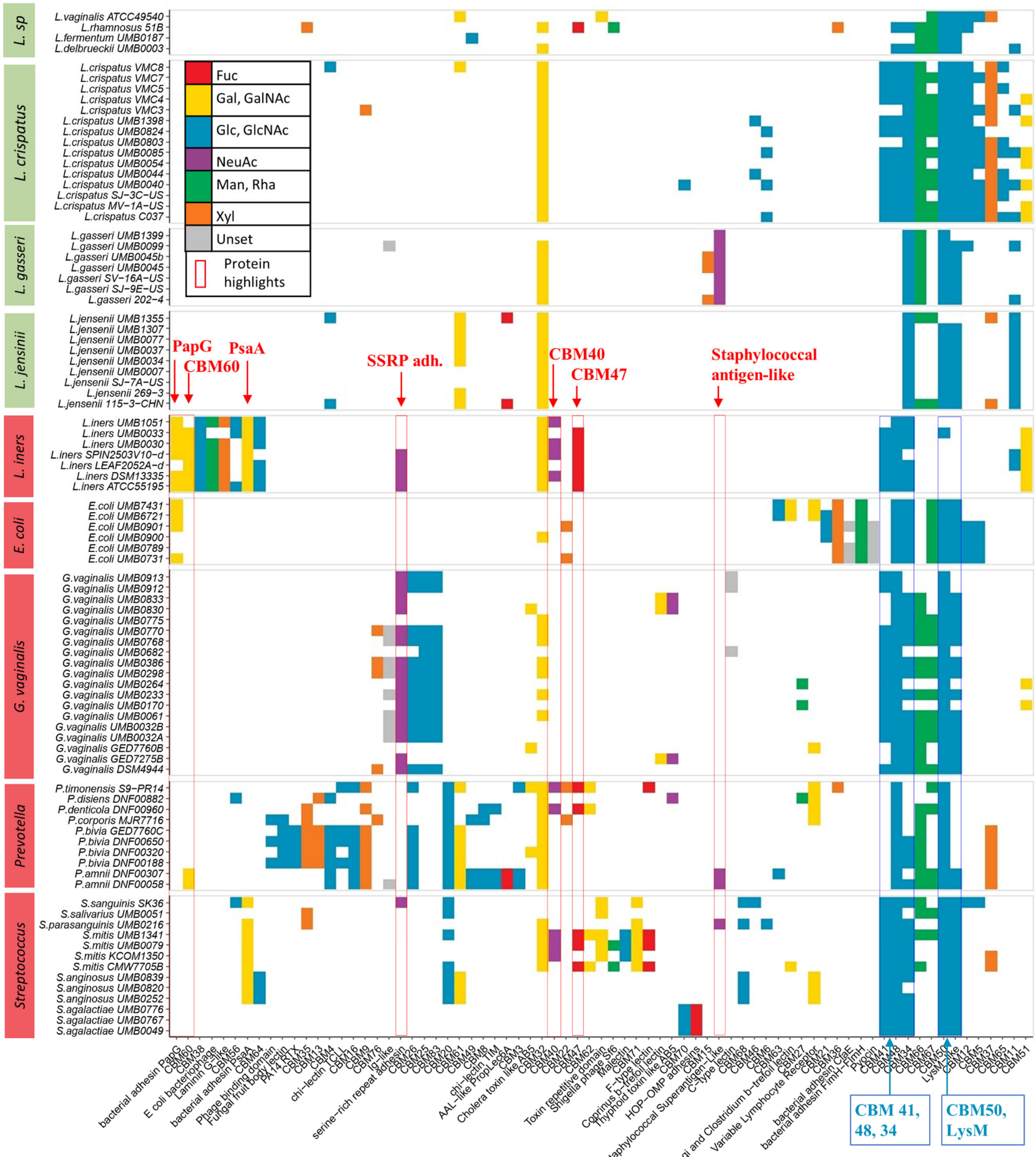

Figure 4. Heatmap of predicted lectin and CBM domains from different vaginal commensal and pathobiont bacterial species arranged by domain composition similarity. Colours within each class of lectin reflect its main glycan specificity (SNFG nomenclature). The domains highlighted are further discussed in the results due to their particular presence in *L. iners.* The addition of the CBM domains strengthens the distinction between commensal and pathobiont bacteria.

Further comparison of the predicted lectin and CBM profiles of vaginal commensals and pathobionts were obtained by performing unsupervised hierarchical clustering on a Euclidean distance matrix of the number of proteins per species for each lectin and CBM domain (Figure 5). The resulting hierarchical radial plot using predicted lectins only, showed a clear clustering of the majority of *Lactobacillus* species, with further sub-clustering at species level observable, with the exception of *L.iners* strains, which clustered more closely with other pathobiont species including *Prevotella* and *Streptococcus* species. *G.vaginalis* also did not cluster in a single group. The inclusion of predicted CBMs in the clustering led to improved discrimination between commensal and pathobiont species and led to species-specific clustering of the majority of isolates.

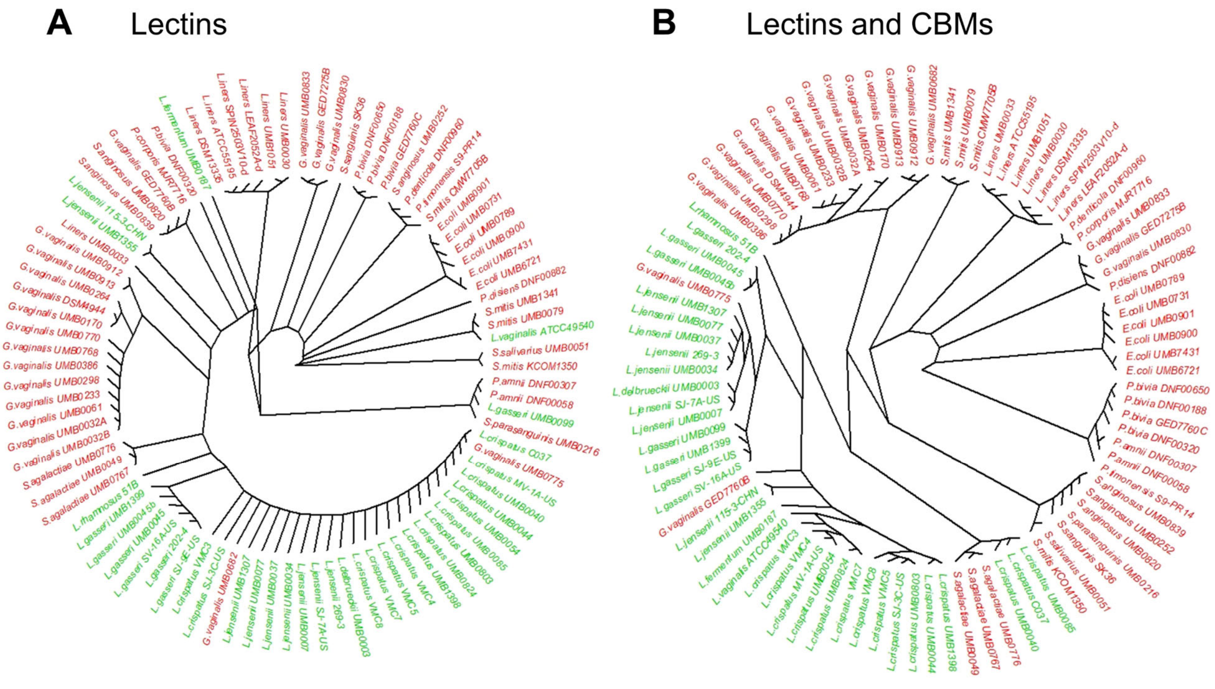

Figure 5. Hierarchical radial tree of (A) predicted lectin classes only or (B) lectin classes and predicted CBMs in vaginal commensal (green) and pathobiont (red) bacteria. LysM and CBM50 are excluded from the dataset to generate the hierarchical radial tree. While the majority of

Lactobacillus species clustered closely to each other, indicating similar putative lectomes, the lectome of *L.iners* isolates more closely resembled that of pathobionts

Discussion

The contribution of bacterial lectins to health and disease remains poorly understood. This is in part because their structural and functional complexity and the limited annotation of bacterial lectins in protein and proteome databases has prevented the development of predictive models of structure, diversity and function. Here, we begin to address this through manual selection of lectin domains in 3D structures obtained from the recently curated Unilectin3D database, followed by the prediction of lectin classes based upon fold similarity and minimum thresholds of sequence identity. This strategy led to the identification of more than 35 different structural folds and 109 predicted lectin classes, of which 19 folds and 37 classes were of bacterial origin. These were particularly rich in β-sheet containing folds, which have previously been recognised as key structural characteristics of lectins from non-bacterial origin [43]. Moreover, predicted classes of pili adhesins and AB5 toxins were found to be exclusive to bacteria. While other lectin classes also appeared to be exclusively predicted in bacteria, these results are likely to be influenced by the fact that to date, many structurally characterised and curated lectins represent those of highest abundance in readily culturable bacteria.

Subsequent prediction of lectin sequences from the bacterial proteomes identified could be used as a basis for future identification of therapeutic molecules to specifically target pathogenic bacteria. They could also have possible interesting specificity to other glycans, but glycan array screenings are required to have further information. 3D structure crystallisation of new possible lectins is also required for a better understanding of their variation in the glycan recognition site.

Given the increased awareness of the importance of the vaginal microbiome in shaping reproductive tract health outcomes, we next undertook comparative analyses of predicted lectins derived from vaginal commensal and pathogenic bacterial isolates. Our analysis, based on 109 structurally characterised lectin classes, suggests that the common commensal species, *L. crispatus* and *L. gasseri* only produces LysM, a ubiquitous domain present in almost all bacteria and involved in binding peptidoglycan with an N-acetylglucosamine specificity [44]. CBM50 is the other denomination of LysM and is therefore also widespread. CAZy annotations confirm it is involved in binding N-acetylglucosamine residues in bacterial peptidoglycans and in chitin. The number of CBMs identified in these species was also very low and corresponded mainly to domains associated with nutrient-degrading glycosylhydrolases. These results suggest that *Lactobacillus* species associated with optimal vaginal microbiome compositions appear to be comparatively ill-equipped for binding mucins. It is important to note that this observation may be biased because the analysis only involved structurally characterised lectins. Further, a limited number of other "mucin adhesion factors" have been described in *Lactobacilli* [45, 46], but except for the fimbriae domain in *L. rhamnosus* (Nishiyama, Ueno et al. 2016), these are in general described as moonlighting proteins, i.e. with adhesion properties being only a side activity in addition to their main function. A shift from *Lactobacillus* species dominance of the vaginal niche towards increased bacterial diversity and enrichment of pathobionts is a signature of vaginal dysbiosis, which has been associated with a range of pathology states including increased risk of sexually transmitted infections [21] and various poor pregnancy outcomes including miscarriage [23], prelabour premature rupture of the fetal membranes [24, 47] and preterm birth [25, 26, 48, 49]. We demonstrated here that the strategy for binding mucins appears to be more evolved in vaginal pathobionts than in commensals, with the former producing a much larger variety of lectins and CBMs. Consistent with our findings, different species of *Streptococci* have been previously shown to produce a large number of lectin domains that form integral parts of toxins, adhesins and pilins [50].

While *Lactobacillus* species are considered hallmarks of optimal vaginal health, *L. iners* is considered a marker of a "transitional microbiome" at the crossroads of vaginal health and disease [51, 52]. The predicted lectomes of the various *L. iners* strains screened were found to contain a significantly larger number of lectin domains than those in other *Lactobacilli,* and the same observation stands when analysing CBMs. This is somewhat surprising considering that *L. iners* has a much smaller genome than other *Lactobacilli* [51]. Several of these identified domains are glycan-specific for glycans present on human mucins such as sialic-binding domain from SRPPs, galactose-specific pilin domain, as well as fucose-binding CBMs usually associated with *Streptococci*. This similarity between *L. iners* and pathogens is in agreement with the previous identification of inerolysin, a pore-forming toxin from *L. iners* also found in *Gardnerella* [53]. Moreover, sequences with similarity to fimbrial proteins PapG from *E. coli*, and Psa/Myf from *Yersinia pestis* were identified in almost all strains of *L. iners.* Interestingly, these two adhesins have similar specificity towards α-galactosylated epitopes [54].

The lectome expansion that appears to correlate with the transition towards species involved in vaginal dysbiosis, raises the question of associated changes in vaginal glycans, and particularly in glyco-epitopes present on mucins. Mucin glycans have been more characterized in gut and lung, and it has been demonstrated that glycosylation is altered in case of inflammation. For example, in cystic fibrosis patients, inflammation results in an increase in fucosylation and sialylation, favouring the attachment of opportunistic pathogens such as *Pseudomonas aeruginosa*, which in turn stimulates the inflammatory process [55]. Such glycan-based processes may occur in the vagina and a deeper characterisation of mucin glycosylation in this context is needed.

While the mechanisms underpinning dynamic shifts in vaginal microbial structure and composition remain to be fully elucidated, our study provides important new insights into lectin profiles of commensal and pathogen colonisation of the reproductive tract that are associated with health and disease states.

The screening tools described and used in the present study can be run on any sequence data and reveal currently concealed information on the content and the role of the lectome. Results show clearly the emergence of characteristic patterns indicative of pathological states. This may guide the development of new strategies for novel therapeutics designed to manipulate adhesion and attachment of microbes to promote optimal colonisation of the lower reproductive tract.

Materials and Methods

Definition of signature profiles for lectins

A new lectin classification has been recently defined based on structural data and is available in the UniLectin3D database (https://unilectin.eu/unilectin3D/). The classification is built on three levels: 1) the fold level directly derived from the protein three-dimensional structure that describes the fold adopted by the whole lectin domain (β-helix, β-propeller and others). The nomenclature on fold are adopted from the reference structural-based databases, CATH [56] and SCOPe [57] and previous reports on structural classification of lectins [58]; 2) The class level defined by sequence similarity with a 20% cut-off between different classes, i.e., lectin sequences in one class are at least 20% similar to one another; 3) The family level defined at a minimum of 70% of sequence identity. The values of cut-offs were set in agreement with definitions in the CATH database for the class level, and empirically for the family level in order to maximise the consistency of each family. The classification is therefore organized in 35 folds, 109 classes, and 350 families.

For each of the 109 lectin classes, UniLectin3D sequences were aligned with the Muscle software [59] to construct a characteristic motif of conserved residues. Sequence redundancy was automatically removed. Manual inspection of characteristic lectin domains led to creating a list of disqualifying domains such as peptide tags in order to manage future systematic removal. Conserved regions from the multiple alignments were then fed to a Hidden Markov Modelling tool to generate profiles characterising each lectin class. The HMMER-hmmbuild tool [60] was used to align each lectin class multiple sequence alignment against protein sequence datasets, with the sym_frac parameter at 0.8 to avoid isolated regions in the conserved motifs.

Prediction of bacterial lectins in protein databases

Bacterial sequences recorded in UniProtKB [61] and in non-redundant NCBI were processed with HMMER-hmmsearch, with default parameters and a p-value below 10^{-2}, to run profiles obtained with HMMER-hmmbuild. Parameters include the BLOSUM62 score matrix for amino acid substitutions (Eddy 2004). Further filtering was applied to multiple strains of the same species with almost identical proteins and only a few different amino acids due to natural mutation, sequencing errors, or protein prediction errors. Post-processing involved keeping only one representative protein for all redundant proteins (with 100 consecutive amino acids that are identical). Predicted domains with less than 15 amino acids are considered as small fragments.

Each sequence match output by the HMMER toolset is evaluated with a quality score that has no upper boundary. Furthermore, because each family profile is generated independently of one another, quality scores are not comparable across motifs used for the prediction. This makes it impossible to use a single cut-off for all lectin classes. Additionally, in the case of tandem repeat domains, the quality score is proportional to the number of repeats and artificially promotes sequences with repeated domains. To address these scoring issues, a prediction score for each database hit was defined to give the similarity between the predicted domain and the reference lectin motif. The amino acid sequence alignment generated by HMMER during the search is further evaluated: at each position of the alignment, a cumulative counter is incremented by 1 if amino acids are identical, else by a normalised BLOSUM62 substitution score. The final value of the counter divided by the domain length (i.e., the total number of positions) results in a value between 0 to 1 that defines the prediction/similarity score. A predicted lectin may belong to several classes, independently of the prediction score. The prediction/similarity score is mainly destined to order the information to be displayed on the UniLectin platform for each predicted lectin. HMMER p-value threshold (better defined then HMMER score) applied before remains the most reliable parameter for trusting a candidate lectin.

For each predicted protein, associated annotations are extracted and loaded from UniProt and from the NCBI. This includes the taxonomy details of the protein and the corresponding ID of the NCBI taxonomy database. Proteins considered as obsolete in the latest releases of UniProt or in the NCBI, with no associated metadata, are removed.

Prediction of lectins and CBMs in the vaginal microbiome

The subset of bacteria corresponding to the vaginal microbiome (S2Table) was identified from genome database annotations, such as those found in the Bioproject www.ncbi.nlm.nih.gov/bioproject/PRJNA316969 and from a published list of bacteria [62]. Bacteria belonging to different species of *Lactobacilli*, *Gardnerella*, *Prevotella*, *E. coli* and Group B *Streptococc*i were selected and classified into commensals or pathobionts on the basis of their potential pathogenicity within the vaginal niche [63], and their association with states of health and disease including bacterial vaginosis, preterm birth and risk of acquisition of sexually transmitted infections [18, 19, 21, 25, 47, 49, 52, 63].

The proteome of each strain was downloaded from the NCBI assembly database [64]. The corresponding sequences were processed to detect lectins and CBMs with the same method of prediction involving the 109 lectin profiles generated as described above. HMMER-hmmsearch was run to identify the lectome of each strain's proteome with default parameters and a p-value below 10^{-2} with no further filtering. Proteins producing good quality alignments (HMM score > 50) with HMMER during the analysis of amino acid sequences were directly tagged as lectin domains. For lesser quality alignments the "Align Sequences Protein BLAST" component of the BlastP tool (ref) was used with default parameters to align a predicted domain against the closest reference lectin with a defined 3D structure. Manual quality checks, especially focused on the glycan binding pocket, were carried out to verify the amino acid conservation and ensure the quality of the predicted lectin.

HMM profiles of Carbohydrate-binding modules (CBMs)were extracted from dbCAN2, a web server for the identification of carbohydrate-active enzymes [65]. The HMM profiles provided by dbCAN2 are based on CAZy CBM sequence data [66]. These profiles were used to identify 1777 proteins from the predicted proteomes of the vaginal commensals and pathobionts. Following removal of high frequency influenza-like predicted lectins and CBD domains occurring in less than three strains, the resulting data was grouped by domain clustering to reflect compositional similarities. The remaining CBMs were associated with their matching glycans and additional information (S3 Table).

To reinforce the results influenza-like predicted lectins are removed (the high frequency of this domain is misleading, as mentioned earlier) and the lectin and CBM domains occurring in less than three strains were filtered out (removing 20 lectin classes and 15 CBM domains for a total of 50 proteins).

Statistical software

Predicted lectins in the HMMER output format were formatted into a tabulated matrix flat file by a python parser and loaded in R for statistical analysis. The following libraries were used:

1. Graphics were generated with R libraries of the Comprehensive R Archive Network (CRAN) including the *d3heatmap* package for heatmaps
2. Hierarchical clustering: The Ward's minimum variance method part of the *hclust* R package was used to process a Euclidean distance matrix of the number of predicted proteins per species for each domain
3. GGplot2 and the APE (Analyses of Phylogenetics and Evolution) package for the hierarchical tree. In this case, prior clustering was applied to the data with the complete linkage method of the *hclust* R package. A Euclidean distance matrix of the number of predicted proteins per species for each domain was input.

For the sake of simplicity, lectins occurring in at least two strains are represented and the Influenza domain is filtered out for the lectin heatmap; and in at least 3 strains for the lectin and CBM heatmap. When lectins and CBMs are represented together the domains present in at least three strains are considered. The lectin and CBM specificity for glycans was manually recovered using UniLectin3D database and CAZy database annotations. Only predicted bacterial lectins with a score greater than 0.25 are kept.

Acknowledgments

The authors acknowledge support by the ANR PIA Glyco@Alps (ANR-15-IDEX-02),Alliance Campus Rhodanien Co-Funds (http://campusrhodanien.unige-cofunds.ch), Labex Arcane/CBH-EUR-GS (ANR-17-EURE-0003) and the March of Dimes.

References

1. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012;13(4):260-70.
2. Thornton DJ, Rousseau K, McGuckin MA. Structure and function of the polymeric mucins in airways mucus. Annu Rev Physiol. 2008;70:459-86.
3. Tailford LE, Crost EH, Kavanaugh D, Juge N. Mucin glycan foraging in the human gut microbiome. Front Genet. 2015;6:81.
4. Corfield AP. The Interaction of the Gut Microbiota with the Mucus Barrier in Health and Disease in Human. Microorganisms. 2018;6(3).
5. Etzold S, Juge N. Structural insights into bacterial recognition of intestinal mucins. Curr Opin Struct Biol. 2014;28:23-31.
6. Ficko-Blean E, Boraston AB. Insights into the recognition of the human glycome by microbial carbohydrate-binding modules. Curr Opin Struct Biol. 2012;22(5):570-7.
7. Lis H, Sharon N. Lectins: Carbohydrate-specific proteins that mediate cellular recognition. Chem Rev. 1998;98(2):637-74.
8. Lepenies B, Lang R. Editorial: Lectins and Their Ligands in Shaping Immune Responses. Front Immunol. 2019;10:2379.
9. Moonens K, Remaut H. Evolution and structural dynamics of bacterial glycan binding adhesins. Curr Opin Struct Biol. 2017;44:48-58.
10. Merritt EA, Hol WGJ. AB5 toxins. Curr Opin Struct Biol. 1995;5:165-71.
11. Imberty A, Mitchell EP, Wimmerová M. Structural basis for high affinity glycan recognition by bacterial and fungal lectins. Curr Opin Struct Biol. 2005;15:525-34.
12. Eierhoff T, Bastian B, Thuenauer R, Madl J, Audfray A, Aigal S, et al. A lipid zipper triggers bacterial invasion. Proc Natl Acad Sci U S A. 2014;111:12895-900.
13. Fazli M, Almblad H, Rybtke ML, Givskov M, Eberl L, Tolker-Nielsen T. Regulation of biofilm formation in Pseudomonas and Burkholderia species. Environ Microbiol. 2014;16(7):1961-81.
14. Iliev ID, Funari VA, Taylor KD, Nguyen Q, Reyes CN, Strom SP, et al. Interactions between commensal fungi and the C-type lectin receptor Dectin-1 influence colitis. Science. 2012;336(6086):1314-7.
15. Pang X, Xiao X, Liu Y, Zhang R, Liu J, Liu Q, et al. Mosquito C-type lectins maintain gut microbiome homeostasis. Nat Microbiol. 2016;1:16023.
16. Cross BW, Ruhl S. Glycan recognition at the saliva - oral microbiome interface. Cell Immunol. 2018;333:19-33.
17. MacIntyre DA, Sykes L, Bennett PR. The human female urogenital microbiome: complexity in normality. Emerging Topics in Life Sciences. 2017;1(4):363-72.
18. Ma B, Forney LJ, Ravel J. Vaginal microbiome: rethinking health and disease. Annu Rev Microbiol. 2012;66:371-89.
19. van de Wijgert JH, Borgdorff H, Verhelst R, Crucitti T, Francis S, Verstraelen H, et al. The vaginal microbiota: what have we learned after a decade of molecular characterization? PLoS One. 2014;9(8):e105998.

20. Reimers LL, Mehta SD, Massad LS, Burk RD, Xie X, Ravel J, et al. The Cervicovaginal Microbiota and Its Associations With Human Papillomavirus Detection in HIV-Infected and HIV-Uninfected Women. J Infect Dis. 2016;214(9):1361-9.

21. Borgdorff H, Tsivtsivadze E, Verhelst R, Marzorati M, Jurriaans S, Ndayisaba GF, et al. Lactobacillus-dominated cervicovaginal microbiota associated with reduced HIV/STI prevalence and genital HIV viral load in African women. ISME J. 2014;8(9):1781-93.

22. Mitra A, MacIntyre DA, Ntritsos G, Smith A, Tsilidis KK, Marchesi JR, et al. The vaginal microbiota associates with the regression of untreated cervical intraepithelial neoplasia 2 lesions. Nat Commun. 2020;11(1):1999.

23. Al-Memar M, Bobdiwala S, Fourie H, Manino R, Lee YS, Smith A, et al. The association between vaginal bacterial composition and miscarriage: a nested case-control study. BJOG. 2019.

24. Brown RG, Al-Memar M, Marchesi JR, Lee YS, Smith A, Chan D, et al. Establishment of vaginal microbiota composition in early pregnancy and its association with subsequent preterm prelabor rupture of the fetal membranes. Transl Res. 2019;207:30-43.

25. Fettweis JM, Serrano MG, Brooks JP, Edwards DJ, Girerd PH, Parikh HI, et al. The vaginal microbiome and preterm birth. Nat Med. 2019;25(6):1012-21.

26. Kindinger LM, MacIntyre DA, Lee YS, Marchesi JR, Smith A, McDonald JA, et al. Relationship between vaginal microbial dysbiosis, inflammation, and pregnancy outcomes in cervical cerclage. Sci Transl Med. 2016;8(350):350ra102.

27. Gipson IK. Mucins of the human endocervix. Front Biosci. 2001;6:D1245-55.

28. Wiggins R, Hicks SJ, Soothill PW, Millar MR, Corfield AP. Mucinases and sialidases: their role in the pathogenesis of sexually transmitted infections in the female genital tract. Sexually Transmitted Infections. 2001;77(6):402-8.

29. Kilian M, Reinholdt J, Lomholt H, Poulsen K, Frandsen EVG. Biological significance of IgA1 proteases in bacterial colonization and pathogenesis: Critical evaluation of experimental evidence. Apmis. 1996;104(5):321-38.

30. Robertson JA, Stemler ME, Stemke GW. Immunoglobulin-a Protease Activity of Ureaplasma-Urealyticum. Journal of Clinical Microbiology. 1984;19(2):255-8.

31. Coombs GH, North MJ. An Analysis of the Proteinases of Trichomonas-Vaginalis by Polyacrylamide-Gel Electrophoresis. Parasitology. 1983;86(Feb):1-6.

32. Cauci S, Monte R, Driussi S, Lanzafame P, Quadrifoglio F. Impairment of the mucosal immune system: IgA and IgM cleavage detected in vaginal washings of a subgroup of patients with bacterial vaginosis. J Infect Dis. 1998;178(6):1698-706.

33. Vornhagen J, Quach P, Boldenow E, Merillat S, Whidbey C, Ngo LY, et al. Bacterial Hyaluronidase Promotes Ascending GBS Infection and Preterm Birth. MBio. 2016;7(3).

34. Carlin AF, Uchiyama S, Chang YC, Lewis AL, Nizet V, Varki A. Molecular mimicry of host sialylated glycans allows a bacterial pathogen to engage neutrophil Siglec-9 and dampen the innate immune response. Blood. 2009;113(14):3333-6.

35. Mariethoz J, Alocci D, Gastaldello A, Horlacher O, Gasteiger E, Rojas-Macias M, et al. Glycomics@ExPASy: Bridging the gap. Mol Cell Proteomics. 2018.

36. Mariethoz J, Khatib K, Alocci D, Campbell MP, Karlsson NG, Packer NH, et al. SugarBindDB, a resource of glycan-mediated host-pathogen interactions. Nucleic Acids Research. 2016;44(D1):D1243-50.

37. Bonnardel F, Mariethoz J, Salentin S, Robin X, Schroeder M, Perez S, et al. UniLectin3D, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands. Nucleic Acids Research. 2019;47:D1236–D44.

38. Sehnal D, Deshpande M, Varekova RS, Mir S, Berka K, Midlik A, et al. LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. Nature Methods. 2017;14(12):1121-2.

39. Neelamegham S, Aoki-Kinoshita K, Bolton E, Frank M, Lisacek F, Lutteke T, et al. Updates to the Symbol Nomenclature for Glycans guidelines. Glycobiology. 2019;29(9):620-4.

40. Boraston AB, Ficko-Blean E, Healey M. Carbohydrate recognition by a large sialidase toxin from Clostridium perfringens. Biochemistry. 2007;46(40):11352-60.

41. Vasta GR, Amzel LM, Bianchet MA, Cammarata M, Feng C, Saito K. F-Type Lectins: A Highly Diversified Family of Fucose-Binding Proteins with a Unique Sequence Motif and Structural Fold, Involved in Self/Non-Self-Recognition. Front Immunol. 2017;8:1648.

42. Montanier C, Flint JE, Bolam DN, Xie H, Liu Z, Rogowski A, et al. Circular permutation provides an evolutionary link between two families of calcium-dependent carbohydrate binding modules. J Biol Chem. 2010;285(41):31742-54.

43. Loris R. Principles of structures of animal and plant lectins. Biochim Biophys Acta. 2002;1572(2-3):198-208.

44. Mesnage S, Dellarole M, Baxter NJ, Rouget JB, Dimitrov JD, Wang N, et al. Molecular basis for bacterial peptidoglycan recognition by LysM domains. Nat Commun. 2014;5:4269.

45. Nishiyama K, Sugiyama M, Mukai T. Adhesion Properties of Lactic Acid Bacteria on Intestinal Mucin. Microorganisms. 2016;4(3).

46. Velez MP, De Keersmaecker SC, Vanderleyden J. Adherence factors of Lactobacillus in the human gastrointestinal tract. FEMS Microbiol Lett. 2007;276(2):140-8.

47. Brown RG, Marchesi JR, Lee YS, Smith A, Lehne B, Kindinger LM, et al. Vaginal dysbiosis increases risk of preterm fetal membrane rupture, neonatal sepsis and is exacerbated by erythromycin. BMC Med. 2018;16(1):9.

48. Vaneechoutte M. The human vaginal microbial community. Res Microbiol. 2017;168(9-10):811-25.

49. Kindinger LM, Bennett PR, Lee YS, Marchesi JR, Smith A, Cacciatore S, et al. The interaction between vaginal microbiota, cervical length, and vaginal progesterone treatment for preterm birth risk. Microbiome. 2017;5(1):6.

50. Moschioni M, Pansegrau W, Barocchi MA. Adhesion determinants of the Streptococcus species. Microb Biotechnol. 2010;3(4):370-88.

51. Macklaim JM, Gloor GB, Anukam KC, Cribby S, Reid G. At the crossroads of vaginal health and disease, the genome sequence of Lactobacillus iners AB-1. Proc Natl Acad Sci U S A. 2011;108 Suppl 1:4688-95.

52. Petrova MI, Reid G, Vaneechoutte M, Lebeer S. Lactobacillus iners: Friend or Foe? Trends Microbiol. 2017;25(3):182-91.

53. Rampersaud R, Planet PJ, Randis TM, Kulkarni R, Aguilar JL, Lehrer RI, et al. Inerolysin, a cholesterol-dependent cytolysin produced by Lactobacillus iners. J Bacteriol. 2011;193(5):1034-41.

54. Kline KA, Falker S, Dahlberg S, Normark S, Henriques-Normark B. Bacterial adhesins in host-microbe interactions. Cell Host Microbe. 2009;5(6):580-92.

55. Cott C, Thuenauer R, Landi A, Kühn K, Juillot S, Imberty A, et al. *Pseudomonas aeruginosa* lectin LecB inhibits tissue repair processes by triggering beta-catenin degradation. BBA - Molecular Cell Research. 2016;1863:1106-18.

56. Dawson NL, Sillitoe I, Lees JG, Lam SD, Orengo CA. CATH-Gene3D: Generation of the resource and its use in obtaining structural and functional annotations for protein sequences. Methods Mol Biol. 2017;1558:79-110.

57. Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. 2020;48(D1):D376-D82.

58. Fujimoto Z, Tateno H, Hirabayashi J. Lectin structures: classification based on the 3-D structures. Methods Mol Biol. 2014;1200:579-606.

59. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792-7.

60. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res. 2018;46(W1):W200-W4.

61. UniProt C. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506-D15.

62. Thomas-White K, Forster SC, Kumar N, Van Kuiken M, Putonti C, Stares MD, et al. Culturing of female bladder bacteria reveals an interconnected urogenital microbiota. Nat Commun. 2018;9(1):1557.

63. van de Wijgert J, Verwijs MC, Gill AC, Borgdorff H, van der Veer C, Mayaud P. Pathobionts in the Vaginal Microbiota: Individual Participant Data Meta-Analysis of Three Sequencing Studies. Front Cell Infect Microbiol. 2020;10:129.

64. Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, et al. Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Res. 2016;44(D1):D73-80.

65. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46(W1):W95-W101.

66. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research. 2014;42(Database issue):D490-5.

Supporting information captions

S1 Table : List of lectin classes identified from Unilectin3D and used in the classification

S2 Table. List of the species and strains used in the study

S3 Table. CBMs of interest for the present study with associated glycan specificity

.