Genetic elements orchestrating *Lactobacillus crispatus* glycogen metabolism in the vagina

Rosanne Hertzberger¹, Ali May²,³, Gertjan Kramer⁴, Isabelle van Vondelen¹, Douwe Molenaar¹ and Remco Kort¹,² *

¹Department of Molecular Cell Biology, Vrije Universiteit Amsterdam, De Boelelaan 1108, 1081 HZ Amsterdam, The Netherlands; r.y.hertzberger@vu.nl; d.molenaar@vu.nl

²ARTIS-Micropia, Plantage Kerklaan 38 – 40, 1018 CZ Amsterdam, The Netherlands

³Biodata & Translation Innovation, Royal DSM, Alexander Fleminglaan 1, 2613 AX Delft, The Netherlands, ali.may@dsm.com

⁴Laboratory for Mass Spectrometry of Biomolecules, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands; g.kramer@uva.nl

*Correspondence: r.kort@vu.nl*
Abstract

Glycogen in the female lower reproductive tract is a major carbon source for vaginal colonization and acidification by common vaginal Lactobacillus species, such as Lactobacillus crispatus. Previously we identified the pullulanase gene pulA in Lactobacillus crispatus, correlating with its ability to autonomously utilize glycogen for growth. Here we further characterize genetic variation and differential regulation of pulA affecting the presence of its gene product on the outer surface layer. We show that alpha-glucan degrading activity dissipates when Lactobacillus crispatus is grown on glucose, maltose and maltotriose, in agreement with carbon catabolite repression elements flanking the pulA gene. Proteome analysis of the S-layer confirmed that the pullulanase protein is highly abundant in an S-layer enriched fraction, but not in a strain with a defective pullulanase variant or in a pullulanase-sufficient strain grown on glucose. In addition, we provide evidence that Lactobacillus crispatus pulA mutants are relevant in vivo, as they are commonly observed in metagenome datasets of human vaginal microbial communities. Analysis of the largest publicly available human vaginal metagenome dataset indicates that 15 out of 272 samples, containing a Lactobacillus crispatus pulA gene, contain a defective variant of this gene. Another 23 out of 272 samples show large deletions or transposon insertions. Taken together, these results demonstrate that both environmental as well as genetic factors explain variation of Lactobacillus crispatus alpha-glucosidases in the vaginal environment.

Keywords: Lactobacillus crispatus; pullulanase; glycogen

1. Introduction

The vaginal environment of reproductive-age women is unique amongst vertebrates in several aspects: it has a low pH due to high levels of lactate (~100-150 mM) [1] it has abundant glycogen levels [2] and its microbial community is dominated by Lactobacillus species [3,4]. In recent years more evidence has emerged that the vaginal microbiome is associated with reproductive and sexual health. Women with a Lactobacillus crispatus microbiome are at lower risk of vaginal infections and sexually communicable infections [5–7]. In addition, vaginal Lactobacillus crispatus is inversely associated with more serious health outcomes such as subfertility [8,9], preterm birth [10] and persistent HPV infection [11,12].

We previously demonstrated that most vaginal L. crispatus isolates are capable of autonomous growth on glycogen. A disrupted N-terminal signal peptide sequence in the pullulanase gene, either by small mutations or structural variants, coincided with the inability of growing on glycogen [13].

A recent key study confirmed alpha-glucosidase activity of pullulanase genes from different vaginal bacteria by heterologous expression. It showed that the Lactobacillus crispatus pullulanase degrades glycogen, pullulan and amylose retaining activity at pH values that are common in a Lactobacillus-dominated vaginal environment [14]. L. crispatus strains isolated from low-glycogen environments, such as the human or poultry gut, more often lacked the pullulanase gene indicating niche-specificity [15].

An open question in the field is the role of vaginal glycogen stores in bacterial acidification and the origin of vaginal glycogen degrading enzymes. Vaginal samples show alpha-glucosidase activity, which was postulated to be mostly of human origin on the basis of antibody [16,17] and proteomic analysis [18]. At the same time results from two studies have indicated that alpha-glucosidases in a subset of vaginal samples may be of bacterial origin.
on the basis of optimal pH and substrate and product profiling [14,19]. However, a metaproteomic study of vaginal lavages from women with a crisatus-dominated microbiome, failed to detect any L. crisatus pullulanase protein, which poses questions about host-related or genetic factors influencing its presence [18].

Here we present results showing substantial disruption and down-regulation of the Lactobacillus crisatus pullulanase gene. We use a mass spectrometric approach to demonstrate the presence of the Lactobacillus crisatus pullulanase enzyme in an S-layer enriched fraction, its downregulation in glucose-, maltose- and maltotriose-grown cultures and its absence in a strain with a defective pulA gene. Analysis of a vaginal metagenomic dataset demonstrates the presence of a significant number of mutations and structural variants expected to fully disrupt pullulanase functionality in vivo, indicating that both environmental as well as genetic factors may refrain vaginal Lactobacillus from metabolizing glycogen.

2. Results

2.1 Substrate utilization of pullulanase sufficient and deficient Lactobacillus crisatus strains.

Previously we reported a collection of Lactobacillus crisatus strains with variable growth on glycogen associated with the presence of an intact amino-terminal signal peptide for protein secretion in the pullulanase gene product (PulA) [13]. Here we further explore the substrate range of a Lactobacillus crisatus deficient (RL09) and sufficient (RL10) strain.

To verify the identified deletion of two nucleotides we used a PCR-amplified 300 bp DNA fragment encoding the PulA N-terminal region (including 42 nucleotides upstream the start site) of strains RL09 and RL10 and resequenced with Sanger dideoxy sequencing. This confirmed the presence of the deletion of two nucleotides in the signal peptide sequence of RL09 (hereafter referred to as pulA⁻ compared to RL10 (hereafter referred to as pulA⁺ (Appendix Figure 1A).

Previously we found pulA⁻ strain RL09 unable to grow on glycogen. We reassessed growth and acidification of this strain and pulA⁺ strain RL10 on NYCIII medium without glucose supplemented with 11 different carbohydrates to analyze other differences between the strains in metabolic profile. Optical density and medium pH after 24 hours of anaerobic growth are shown in Figures 1A and B, respectively.
Figure 1. Carbohydrate utilization by Lactobacillus crispatus strains pulA\(^{-}\) (RL09) and pulA\(^{+}\) (RL10). A Optical density (600nm) and B pH after serial propagation in NYCIII medium supplemented with different carbohydrates. The bars represent the mean of 3 biological replicates, the standard deviations are shown as error bars. * p < 0.05, ** p < 0.005 (multiple t-test with false discovery rate).

Both L. crispatus strains were equally capable of growth on all carbon sources except for lactose, raffinose, cellobiose and lactulose, which is in disagreement with previous findings [20] Growth was accompanied by acidification of the medium to a pH of 3.9, corresponding to the in vivo vaginal pH [1] and the pKa of lactic acid. Glycogen, both of oyster and bovine origin, was the only carbohydrate resulting in a clear difference between the strains: the pullulanase-deficient strain grew to an optical density of \(~0.5\) and acidified to a pH of 6.0
(Oyster glycogen) and 5.5 (Bovine glycogen) possibly due the presence of smaller maltodextrins and glucose after autoclaving the glycogen. The growth of the pullulanase-sufficient strain on glycogen was comparable to that on glucose, acidifying the medium to a pH of 3.9. Regarding growth on the 11 carbohydrates tested, we conclude that pulA+ and pulA− strains do not substantially differ in their metabolic capacities apart from the ability to grow on glycogen.

2.2 Starch degradation activity of pullulanase sufficient and deficient Lactobacillus crispatus strains on various carbon sources.

Next we measured enzymatic activity of the pullulanase enzyme to understand its regulation and to find pullulanase-expressing conditions that would allow for a direct proteomic comparison between the two strains. Genetic elements in the locus of the pulA gene are strongly indicative of carbon catabolite repression, similar to the regulation of pullulanase in other lactobacilli [21]. The promoter region of the gene has a cre (catabolite responsive element) like palindromic sequence (TGTTATCGATAACA). The catabolite responsive element contains a well-known binding site for the global regulator CcpA (carbon catabolite protein A) suppressing metabolic pathways for alternative carbohydrates in the presence of glucose in various Lactobacillus species [22–24]. Downstream of the pulA gene we identified two open reading frames with homology to the LacI family of repressors, containing a DNA binding domain in one open reading frame (1-110) and an effector domain (91-219) in the other, possibly involved in catabolite repression through binding of breakdown products of glycogen (see below).

To verify the predicted regulation by carbon source we grew the pullulanase-sufficient strain L. crispatus to stationary phase on NYCIII medium supplemented with glycogen, maltotriose, maltose, glucose or galactose and measured alpha-glucosidase activity in the pellet and supernatant. We included galactose since lactic acid bacteria generally show preference of glucose over galactose [25,26] and we hypothesize that galactose does not repress pulA expression to the same extent as glucose.

After 48 hours of growth, spent medium and bacterial cells were mixed with starch solution to detect starch degradation in a standard starch-iodine assay. Starch was used as a substitute for glycogen, since it can be easily detected using iodine and has an amylopectin moiety with alpha1-4 and alpha1-6 linkages comparable to glycogen. Total starch degradation for a period of 24 hours at 37 degrees was used as a measure of alpha-glucosidase activity in the sample. Starch degradation was observed in pellet (Figure 3) as well as spent growth media (Appendix figure A2) of L. crispatus pulA+ strain RL10 after growth on glycogen and galactose, but not after growth on glucose, maltose and maltotriose, confirming the repression of this gene under these conditions.
Figure 3 Cell-bound starch degrading activity of Lactobacillus crispatus pullulanase sufficient and deficient strains grown on NYC-, glucose, maltose, maltotriose, glycogen (only pulA+) and galactose. Cells were cultured in NYClIII medium supplemented with these carbohydrates and after centrifugation and resuspension of the pellet incubated in a starch solution (7.5 g/L). Asterisks indicate p-values calculated by Mann-Whitney of * p<0.05, ** p<0.005, *** p<0.0005. Hashtag indicates no data due to the absence of growth of this strain on glycogen.

Alpha-glucosidase activity in cells and spent medium of the pulA+ strain was comparable to the pulA− strain on all substrates except for glycogen since the pulA− strain did not sufficiently grow on glycogen to measure starch degradation. Bacterial cells or supernatants of the pulA− strain grown on glucose, maltose, maltotriose or galactose did not degrade any starch conforming the link between the defective pullulanase enzyme and glycogen metabolism.

2.3 S-layer enrichment and proteome analysis

In our previous study we identified a surface layer associated protein (SLAP) domain in the Lactobacillus crispatus pullulanase protein sequence [13], which prompted us to investigate its localization in the surface layer. We hypothesized that the natural variation in the amino-terminal signal peptide for secretion of pullulanase would result in the presence or absence of an active pullulanase in the S-layer. Furthermore, we wanted to confirm that the catabolite repression of pullulanase by glucose, as observed in this study by use of the starch degradation assay (Figure 2), would be reflected by the levels of pullulanase in the S-layer. We utilized galactose as a carbon source since it 1) allows for growth of both pulA− and pulA+ strains and 2) is a condition under which we observed cell-associated starch degradation activity (figure 2). We compared cells of the pulA+ RL10 strain grown on galactose, glucose and glycogen in order to verify expression of pullulanase in an S-layer enriched fraction under these conditions.

The S-layers form the outermost structure of the cell envelope and represent up to 15% of the total protein content [27]. As the S-layer protein subunits (SLPs) and S-layer associated proteins (SLAPs) are non-covalently linked to the cell wall, we released the S-layer fraction from the cells and let it disintegrated into monomeric polypeptides by exposure of the cells to high concentrations of the denaturing agent LiCl. Mass-spectrometric analysis of this isolated S-layer fraction led to the identification of peptides of at least 28 SLAP-domain containing proteins, including a number of abundant SlpA−domain containing proteins (Table 1). The most prominent distinction in the amount of all proteins identified in RL09 and RL10 S-layers is a 90-fold difference for pullulanase (also known as alpha-dextrin endo-1,6-alpha-glucosidase). The minor background of pullulanase peptides in RL09 may result from lysed cells releasing intracellular pullulanase. S-layer fractions of the pulA+ strain RL10 grown on glucose showed a 24-fold lower pullulanase presence compared to fractions from glycogen grown cells. Two uncharacterized surface layer proteins (A0A7T1TNA6 and A0A125P6L8)
are strongly upregulated in the presence of glucose by a factor of over 100; all other proteins show minor differences in levels with ratios between 0.2 and 2.7. The overall fraction of S-layer proteins appears to have been reduced in the presence of glucose (0.2 versus 0.4-0.5 in the other three fractions).
<table>
<thead>
<tr>
<th>ACC NO</th>
<th>DESCRIPTION</th>
<th>PEPT</th>
<th>RL09 GAL</th>
<th>RL10 GAL</th>
<th>RL10 GLU</th>
<th>RL10 GLY</th>
<th>(^2)logR(_1)</th>
<th>(^2)logR(_2)</th>
</tr>
</thead>
</table>
2.4 Metagenome analysis

Our previous comparative genomics analysis of a collection of *Lactobacillus crispatus* strains isolated from the outpatient clinic in Amsterdam revealed variations in the N-terminal signal peptide region of the pullulanase gene that correlated with the ability of a strain to grow on glycogen [13]. In this strain collection, we found seven pullulanase-deficient strains sharing four different types of disruptions of the N-terminal sequence (Table 2). Although not all published *L. crispatus* genomes show the presence of a *pulA* gene [15], the N-terminal disruptions of the *pulA* sequence observed in our collection were not found in other *L. crispatus* genomes reported thus far.

<table>
<thead>
<tr>
<th>Strain</th>
<th><em>pulA</em> variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>RL03, RL10, RL11, RL13, RL14, RL15, RL16, RL20, RL21, RL22, RL23, RL24, RL25, RL27, RL28, RL30, RL33</td>
<td>Wild type</td>
</tr>
<tr>
<td>RL02 and RL09</td>
<td>Two nucleotide deletion</td>
</tr>
<tr>
<td>RL06 and RL07</td>
<td>N-terminal insertion of mobile element (variant 1)</td>
</tr>
<tr>
<td>RL19 and RL26</td>
<td>N-terminal insertion of mobile element (variant 2)</td>
</tr>
<tr>
<td>RL05</td>
<td>Large deletion</td>
</tr>
</tbody>
</table>

Table 2: Overview of pullulanase variants in our strain collection of human vaginal *Lactobacillus crispatus* isolates [13].

We hypothesize that the isolation of these *pulA* variants is not a result of selective enrichment during the isolation procedure, but that *Lactobacillus crispatus* strains with defective pullulanase genes are common in the vaginal microbiota. To test this hypothesis we assessed the prevalence and frequency of disruptive *pulA* gene sequence variants such as those observed in our strain collection (Table 2). We analysed vaginal metagenome datasets from 1507 vaginal microbiota samples [28]. To our knowledge, this is the largest publicly available vaginal microbiome data collection which includes samples from several different studies of North American, African, and Chinese women. In total, we analysed 28.6 billion sequencing reads that made up 3.9 tera base pairs (Tbp) of metagenome sequencing data (on average, 19 million reads and 2.6 Gbp data per sample).

We first determined the prevalence of small but disruptive variations across the *pulA* gene sequence, similar to those observed in *pulA*-strains RL02 and RL09 (Table 2). Out of 1507 samples we aligned to the *L. crispatus* RL10 *pulA* gene sequence, 270 resulted in alignments with a minimum of 10 times full length coverage. Variant calling on these 270 samples resulted in a total of 3202 unique variants (Appendix Figure A3). After low-frequency (allele frequency lower than 0.05) and low prevalence (observed in less than 3 samples) variants were filtered out, 54 variants remained, of which 19, 28 and 6 were low (e.g.
synonymous mutations), moderate (e.g., mutations that cause an amino acid change), and high impact (stop gain or frameshift mutations), respectively (Figure 4).

Fig. 4. Small variants in the *Lactobacillus crispatus* pullulanase gene in the metagenome of vaginal microbial communities. The list and frequency of SNPs and short indels (small variants) identified in the 270 vaginal microbiome samples that showed at least ten times full length coverage for the *pulA* gene sequence. The variants were classified in terms of their impact on the protein sequence.

In 15 out of 270 samples (5.5%), at least one of the high impact mutations was observed at a high allele frequency (>0.7), where we expect a pullulanase deficient phenotype (Supplementary file pulA_small_variants.xlsx). Three of these high-impact mutations were near the N-terminus of the *pulA* sequence at positions 156, 207, and 225, impacting 11 samples in total.

Subsequently, we analysed the same set of 270 samples for larger structural variants compared to the pulA* RL10 pulA gene sequence. In 56 out of 270 samples (20%), we identified large structural variants which had a high frequency (>0.7) (Supplemental File pulA_structural_variants.xlsx). The disruptive structural variation in 17 out of 56 of these samples was found to be due to the insertion of transposase sequences into the pullulanase gene sequence, as in the case of the *pulA* locus from strains RL06, RL07, RL19, and RL26 in our own strain collection. Taken together, we identified high-frequency mutations in 62 of 270 *L. crispatus* pulA strains indicative of a dysfunctional pullulanase of the dominant strain and the inability of *L. crispatus* to contribute to the acidification of the vaginal environment with glycogen as a carbon and energy source.
3. Discussion

In this study we focus on genetic variation and disruption of the Lactobacillus crispatus pullulanase gene. Pullulanase is thought to play a central role in vaginal metabolomics. It catalyzes the first step of a metabolic pathway turning the most abundant vaginal carbohydrate (glycogen) into the most abundant metabolite (lactate) in one of the most abundant vaginal species Lactobacillus crispatus. In a previous study we published 33 genomes of vaginal Lactobacillus crispatus isolates that displayed remarkable variation in the gene locus. Strains with a disrupted pullulanase gene, through small indels or structural variants, were incapable of growth on glycogen.

The wealth of publically available metagenome datasets, repositories and gene catalogs allows for a more hypothesis driven research approach concentrating on specific genes that are associated with in vivo functionality. In the largest available metagenome database a substantial number of samples had similar genetic variation as were present in our previously reported whole genome sequences of Lactobacillus crispatus isolates. These variations consisted of both small deletions as well as relatively large structural variations disrupting the functionality of this gene.

Disruptions in the secretome of catalytically active enzymes are a well-known phenomenon in the microbial world. Bacteria may take advantage of cells in close proximity that cleave ‘public goods’ such as proteins or glycans into smaller peptides or maltodextrins. For instance, a protease-positive Lactococcus lactis population gradually lost protease activity in the entire population in a spatial-dependent manner. In this case the “cheating” proteolytic negative strains can utilize peptides supplied by the proteolytic positive strains without having the cost of protease production [29]. These disruptions may provide a fitness advantage to isolates, especially when, as we show in our mass spectrometric analysis, the enzyme forms a substantial fraction of the outer surface layer.

In the metagenome data of samples we did not find evidence of extensive variety of pullulanase sequences within a single vaginal microbiome: most samples only showed presence of one pullulanase variant, which we assume coincides to the dominant presence of one strain. We hypothesize that the metabolic advantage that a cheating strain in complex ecological samples may experience is not so much a matter of intraspecies competition, but rather due to interspecies crossfeeding of metabolites. The presence of other glycogen degrading bacteria or human amylase may create an environment in which a Lactobacillus crispatus strain may obtain energetic advantages from shutting down the expression and secretion of S-layer associated enzymes, while benefiting from the breakdown products of extracellular glycogen hydrolysis by other species.

A recent study did not find such evolutionary pressure to mutate the pullulanase: even after cultivation of up to 1,000 generations by sequential sub-culturing in synthetic vaginal fluid, no genetic variations in the pulA locus of strains of Lactobacillus crispatus were detected. [30]. We hypothesize that this may be due to the available carbohydrates in this simulated vaginal fluid, which contains both glucose and glycogen and is further supplemented with amylase which degrades the glycogen into smaller maltodextrins. Our study implies that L. crispatus grown in such an environment will first grow on glucose, maltose and maltotriose and will not express pullulanase until these carbohydrates are depleted. This will diminish the advantage of cheating or any other evolutionary pressure on the production of alpha-glucosidase.

Genetic disruption affected a minority of vaginal samples in the metagenomic database or isolates in our strain collection. In most environments L. crispatus maintains an intact pullulanase gene with no sign of a genetic alteration disrupting its activity. However, our study shows that carbon catabolite repression may further reduce activity of Lactobacillus crispatus pullulanase activity in vivo. Bacteria generally regulate metabolic pathways
optimizing their growth strategy by prioritizing carbon sources based on nutritional quality. The presence of glucose, maltose and maltotriose in the vagina has been well-documented, especially in Lactobacillus-dominated environments [31,32].

Given the prevalence of these carbon sources in metabolomics studies, as well as human amylase we expect that Lactobacillus crispatus in the vaginal environment preferentially relies on glycogen cleavage by other bacteria or host amylase and only expresses pullulanase when other carbon sources are depleted.

4. Materials and Methods

4.1 Standard cultivation conditions

Lactobacillus crispatus pullulanase deficient strain (L. crispatus RL09) and a pullulanase sufficient strain (L. crispatus RL10) were routinely cultivated on NYCIII medium with HEPES (2.4 g/L), proteose peptone #3 (Becton Dickinson, 211693) (15 g/L), yeast extract (Oxoid, LP0021) (3.8 g/L), NaCl (5 g/L), glucose monohydrate (Santa Cruz Biotechnology, 14431-43-7) (5.5 g/L g), heat inactivated horse serum (10%) and for agar plates 1.5% (w/v) agarose. Plates, precultures and subcultures were grown under anaerobic conditions (gas mixture of 10% CO₂ and 90% N₂ at 37°C.

4.2 PCR and sequencing of signal peptide region L. crispatus pullulanase

Previously, Illumina-whole genome sequencing showed a two nucleotide deletion in the N-terminal pulA signal peptide sequence of Lactobacillus crispatus RL09. In order to confirm this frame shift mutation we performed genetic analysis of the area with the deletion by PCR amplification and Sanger sequencing. Genomic DNA was isolated from a 24 hour liquid culture using the GenElute™ Bacterial Genomic DNA Kit (Merck). A 300 basepair sequence at the start site of the pullulanase was amplified with forward primer GCAATGAAACGCGATACGTTT (annealing 42 basepairs upstream the start site) and reverse primer TGTTGACGCTGCTTTGCTT (annealing 257 basepairs downstream the start site) using a High Fidelity Phusion® DNA polymerase (ThermoFisher). The size of the PCR product was verified with ethidium bromide on gel to be the predicted size of 299 bp. The PCR product was purified using the GeneJET PCR Purification Kit (Thermo Fisher) and sequenced by Sanger Dideoxy sequencing (Eurofins).

4.3 Serial propagation on various carbohydrates

Growth on various carbohydrate sources was tested with NYCIII without glucose supplemented with D-lactose monohydrate (Sigma-Aldrich), D-raffinose pentahydrate (Alfa Aesar), D-cellobiose (Sigma-Aldrich), D-melibiose (Sigma-Aldrich), D-maltose monohydrate (Fluka), D-trehalose dihydrate (Sigma-Aldrich), D-galactose (Sigma-Aldrich), lactulose (Sigma-Aldrich), glucose monohydrate (Merck), bovine glycogen (Sigma-Aldrich) and oyster glycogen (Alfa Aesar). Growth of L. crispatus RL09 and L. crispatus RL10 on these carbohydrate sources was examined by serial propagation on glucose-free NYCIII medium supplemented with 5% (w/v) of the different carbohydrates. The serial propagation was executed using two passages of a twentyfold sub-culturing in a 96-wells plate. Between each passage, the cells were anaerobically grown for 24h in jars (Oxoid). To monitor growth, optical density was measured at 600 nm in a sterile 96-well flat bottom plates in a spectrophotometer (Spectramax Plus 384; Molecular devices) after ten times dilution in sterile phosphate-buffered saline solution (PBS). Acidification of the medium by L. crispatus was assessed by pH measurements (HI 2210 pH meter, Hanna instrument) after every passage after 24h of anaerobic growth. Statistical testing of optical densities and pH values were done with t-tests: two samples assuming unequal variations (α = 0.05, two-sided).
4.4 The starch degradation assay

Glycogen is a polymer of α-1,4-linked and α-1,6-linked glucose residues. Starch degradation, measured in a basic starch-iodine assay, can be used as a proxy since the amylopectin moiety of starch has a similar branched structure with α-1,4-linked and α-1,6-linked glucose residues. To test α-glucosidase activity strains were grown for 24 hours. Cells were spun down and resuspended in amylase buffer (100 mM Na-Acetate with 5 mM CaCl₂, pH 5.3). 50 µL supernatant or resuspended pellet was added to 150 µL of 10 g/L starch solution (Sigma, S9765) in amylase buffer. Chloramphenicol (Sigma, C0378-25G) from a stock solution in ethanol was added in a final concentration of 10 ng/mL to prevent growth. 50 µL of resuspended pellet of culture spent medium was mixed with 150 µL of this starch working solution and incubated at 37°C for 24 hours. To visualize residual starch, 10 µL was added to 290 µL of an iodine stock solution consisting of 0.2 g I₂ and 2 g KI in 100 mL 50 mM HCl. Absorption was measured at 600 nm in a 96 well plate reader. Experiments were carried out with a minimum of 2 biological replicates for RL09 on maltose and maltotriose; all other conditions had between 3 and 9 replicates carried out on separate days by a total of 4 independent experiments.

4.5 Enrichment for surface layer (associated) proteins

The enrichment for SLP’s and SLAP’s was modified from a standard LiCl S-layer extraction protocol for *L. acidophilus* (Goh et al., 2009; Lortal et al., 1992). Bacterial cells of *L. crispatus* were grown to stationary phase (24 h), centrifuged at 2236 g for 10 min (4°C), and washed twice with 25 ml cold PBS (Gibco), pH 7.4. Cells were agitated for 15 min at 4°C following the addition of 5 M LiCl (Fisher Scientific). Supernatants, containing SLP’s and SLAP’s, were harvested via centrifugation at 8994 g for 10 min (4°C) and transferred to a 6000–8000 kDa Spectra/Por molecular porous membrane (Spectrum Laboratories) and dialysed against cold distilled water for 24 h, changing the water every 2 h for the first 8 h. The dialysed precipitate was harvested via centrifugation at 20,000 g for 30 min (4°C).

4.6 Sample Preparation for Proteomics Analysis

Protein pellets were resuspended in 200 µL of digestion buffer (5 mM DTT, 50 mM NH₄HCO₃, pH 8.0) and incubated at 55°C for one hour to reduce disulfide bridges. Then samples were treated with 15 mM iodoacetamide (IAA) to alkylate free cysteines in the dark for 45 min. Subsequently samples were digested by trypsin (1:100 w/w protease:protein ratio) at 37°C for 18 hours. The mixture of peptides was desalted using C18 solid phase extraction columns (Omix, Agilent) following the manufacturer’s protocol. The final peptide mixture was dried in a vacuum centrifuge and suspended in 0.1% formic acid in water (ULC-MS grade, Biosolve) for MS analysis.

4.7 LC-MS/MS Analysis

Mass spectrometric analysis of 200 ng peptides was carried out on a timsTOF Pro (Bruker, Bremen, Germany) equipped with an Ultimate 3000 nanoRSLC UHPLC system (Thermo Scientific, Germeringen, Germany). Samples were injected onto a C18 column (75 μm, 250 mm, 1.6 μm particle size, Aurora, Ionopticks, Fitzroy, Australia) kept at 50°C. Peptides were loaded at 400 nl/min for 1 minute in 3% solvent B and separated by a multi-step gradient: 5% solvent B at 2 min 17%, solvent B at 24 min, 25% solvent B at 29 min, 34% solvent B at 32 min, 99% solvent B at 33 min held until 40 minute, returning to 3% solvent B at 40.1 min and held until 58 min to recondition the column (Solvent A: 0.1% formic acid in water, Solvent B: 0.1% formic acid in acetonitrile). MS analysis of eluting peptides was performed by a time-of-flight mass spectrometer. The precursor scan ranged from 100 to 1700 m/z and a tims range of 0.6–1.6 V.s/cm² in PASEF mode. A total of 10 PASEF MS/MS scans following collision
induced dissociation (collision energy from 20–59 eV) were collected with a total cycle time of 1.16 s.

Raw MS/MS data were processed using Maxquant software (version 1.6.14.0) [33] searching a pan proteome database of *L. crispatus* (Uniprot, downloaded 1-11-2021). To estimate false spectrum assignment rate a reverse version of the same database was also searched. The settings were as follows: Enzyme Trypsin/P allowing for a maximum of two missed cleavages, variable modifications: Oxidation (M), fixed modifications: Carbamidomethyl (C). Settings were default for timsDDA, match between runs was enabled with a matching time window of 0.2 minutes and a matching ion mobility window of 0.05 indices. For label free quantification, both iBAQ and LFQ were enabled [33].

4.8 Analysis of the pulA gene region

The contig of *Lactobacillus crispatus* RL10 was analyzed with SnapGene version 5.3.2. (GSL Biotech LLC). The promoter sequences (-35 and -10 elements) were predicted by use of BPROM (Prediction of Bacterial Promoters) and FindTerm (Finding Terminators in bacterial genomes) [34]. The cabolite responsive element (*cre* site) has been identified by matching the palindromic nucleotide motif of *Lactobacillus plantarum* (TGWAANCGNTNWCA) [35].

4.9 Analysis of publicly available vaginal metagenomes

The raw shotgun metagenome sequencing data of 1507 publicly available vaginal microbiome samples were downloaded from the VIRGO database [28]. In addition to the 211 samples which were sequenced for the construction of VIRGO, this large data collection contains samples from other studies of North American [36], African [5], and Chinese women [37]. The Illumina reads were processed using FastQC to determine data statistics such as data yield per sample. The samples were taxonomically characterised using Kraken 2 [38] and Bracken [39]. The human read counts in the species-level results from Bracken were removed and the data was normalised to obtain relative abundances of non-human taxa such as bacteria, archaia, viruses, and fungi. Taxa that had less than 0.5% relative abundance and which were observed in less than 5% of the samples were filtered out. The downstream taxonomic analyses and visualisation were performed using R v3.6.3 (R Core Team, 2020) with R packages vegan, phyloseq [40], ComplexHeatmap [41], and ggplot2 v3.3.3.

4.10 Pullulanase sequence variant detection

The raw reads were aligned to the RL10 *pulA* gene sequence using bowtie2 [42] with default parameters except for "--very-sensitive-local --no-unal". Alignment statistics for each sample such as coverage were calculated using samtools v1.13 [43]. Single-nucleotide polymorphisms (SNPs) and short indels in samples with a minimum 10X coverage were determined using FreeBayes v1.2 [44] in a multi-sample variant calling approach using default parameters except for "--min-alternate-count 2 --min-coverage 10 --min-alternate-fraction 0.01 --use-duplicate-reads --min-mapping-quality 0 --ploidy 1 --pooled-continuous". The variants were annotated in terms of their impact on the RL10 *pulA* protein sequence using SnpEff [45]: high (e.g. mutations that result in frameshifts or stop gain), moderate (e.g. mutations that result in an amino acid change), and low (synonymous mutations). The variant allele frequency heatmap for variants which had at least 0.05 alternate allele frequency (i.e., supported by at least 5% of the reads at that position) and were observed in at least 3 samples was plotted using R packages VariantAnnotation [46], and ggplot2. SNPs and short indels with a minimum frequency of 0.7 were classified as high-frequency short variants.

The same sample set used in small variant detection which had a minimum coverage of 10 times along the RL10 *pulA* sequence was used for detecting larger and structural variations such as insertions, deletions, inversions, translocations, and duplications in read mappings.
To this end, the indels and Structural Variants tool in CLC Genomics v.12.0.3 was used with default parameters (P-value threshold=0.0001) except for “minimum number of reads=10”. Structural variants with a minimum frequency of 0.7 were classified as high-frequency large variants.

4.11 Identification of pulA genes with transposase insertions

To create a database of *L. crispatus* transposase gene sequences, we first downloaded 147 annotated *L. crispatus* genomes from NCBI (assembly level: scaffold, chromosome, and complete). We then created a FASTA file of 19,551 redundant transposase gene sequences by combining the coding sequences which had a ‘transposase’ annotation in the 147 *L. crispatus* genomes. This database was clustered at 95% sequence identity using CD-HIT using default parameters [47] to obtain a non-redundant transposase database of 508 gene sequences.

To identify pulA-rich samples that had transposase insertions, we mapped the reads which were mapped to the *L. crispatus* RL10 pulA sequence to the above-mentioned transposase database using bowtie2 as described above. The list of samples where more than 20 transposase reads were found was compared to the list of samples with high-frequency structural variants to determine the impact of transposase insertions in causing disruptive structural variations.

**Supplementary Materials:** The following are available online at www.mdpi.com/xxxx/s1,

Supplementary file data Figure 1

Supplementary file data Figure 2 and Figure A2

Supplementary file pulA_small_variants.xlsx

Supplementary file pulA_structural_variants.xlsx

**Author Contributions:**

Conceptualization, RYH, RK; Methodology, RYH, GK, RK; Software, AM; Experiments: RYH, GK, IvV, DM, RK; data curation: AM, RYH, Writing—original draft preparation, RYH, AM, GK, RK; Writing – review and editing RYH, RK; Supervision, RYH, DM, RK; All authors have read and agreed to the published version of the manuscript.”

**Funding:** This research received no external funding

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Data Availability Statement:** All data presented here are included in the manuscript, appendix or supplemental file. The mass-spectrometry data have been deposited in Proteome Exchange.

**Acknowledgments:** We would like to thank the MSc-students Omar Beganovic, Deborah Jekel, Yumiko van Diest, Leon Steenbergen, and Ritesh Panchoe for their contributions to the research presented in this publication.

**Conflicts of Interest:**
RYH and RK are unpaid board members of the *crispatus* foundation: [www.crispatus.org](http://www.crispatus.org). The other authors declare no conflict of interest.

**Appendix A**

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

**References**


18. Nunn Kenetta L.; Clair Geremy C.; Adkins Joshua N.; Engbrecht Kristin; Fillmore Thomas; Forney Larry J.; Young Vincent B. Amylases in the Human Vagina. mSphere 2020, 5, e00943-20, doi:10.1128/mSphere.00943-20.


Appendix A

Figure A1. Resequencing the *Lactobacillus crispatus* pulA signal peptide region of a pulA* strains RL10 and RL09.
Figure A2 Starch degrading activity in spent medium of *Lactobacillus crispatus* pullulanase sufficient and deficient strains grown on NYC-, glucose, maltose, maltotriose, glycogen (only *pulA*+) and galactose. Cells were cultured in NYCIII medium supplemented with these carbohydrates and after centrifugation and resuspension of the pellet incubated in a starch solution (7.5 g/L). Experiments were carried out with a minimum of (two) biological replicates; averages and standard deviations are shown. Asterisks indicate *p*-values calculated by Mann Whitney of * p<0.05, ** p<0.005, *** p<0.0005.

Figure A3 Variant calling on the *pulA* gene positive 270 samples in the vaginal metagenome data set.