Comparative genomics of human Lactobacillus crispatus isolates reveals genes for glycosylation and glycogen degradation

Van Der Veer, Charlotte; Hertzberger, Rosanne Y.; Bruisten, Sylvia M.; Tytgat, Hanne L.P.; Swanenburg, Jorne; De Kat Angelino-Bart, Alie; Schuren, Frank; Molenaar, Douwe; Reid, Gregor; De Vries, Henry; Kort, Remco

published in
Microbiome
2019

DOI (link to publisher)
10.1101/441972
10.1186/s40168-019-0667-9

document version
Early version, also known as pre-print

citation for published version (APA)
Van Der Veer, C., Hertzberger, R. Y., Bruisten, S. M., Tytgat, H. L. P., Swanenburg, J., De Kat Angelino-Bart, A., Schuren, F., Molenaar, D., Reid, G., De Vries, H., & Kort, R. (2019). Comparative genomics of human Lactobacillus crispatus isolates reveals genes for glycosylation and glycogen degradation: Implications for in vivo dominance of the vaginal microbiota. *Microbiome*, *7*(1), 1-14. [49]. https://doi.org/10.1101/441972, https://doi.org/10.1186/s40168-019-0667-9

Comparative genomics of human *Lactobacillus crispatus* isolates reveals genes for glycosylation and glycogen degradation: Implications for *in vivo* dominance of the vaginal microbiota.

Charlotte van der Veer[1], Rosanne Y. Hertzberger[2], Sylvia M. Bruisten[1,7], Hanne L.P. Tytgat[3], Jorne Swanenburg[2,4], Alie de Kat Angelino-Bart[4], Frank Schuren[4], Douwe Molenaar[2], Gregor Reid[5,6], Henry de Vries[1,7], and Remco Kort[2,4 *]

Affiliations:

[1]Public Health Service, GGD, Department of Infectious diseases, Amsterdam, the Netherlands

[2]Department of Molecular Cell Biology, Faculty of Earth and Life Sciences, VU University, Amsterdam, the Netherlands

[3]Institute of Microbiology, ETH Zürich, Zurich, Switzerland

[4]Netherlands Organization for Applied Scientific Research (TNO), Microbiology and Systems Biology, Zeist, the Netherlands

[5]Canadian R&D Centre for Human Microbiome and Probiotics, Lawson Health Research Institute

[6]Departments of Microbiology and Immunology, and Surgery, Western University, London, Ontario, Canada.

[7] Amsterdam Public Health research institute, Amsterdam UMC, the Netherlands

*Corresponding author at Netherlands Organization for Applied Scientific Research (TNO), Microbiology and Systems Biology, Utrechtseweg 48, 3704 HE, Zeist, the Netherlands

E-mail: remco.kort@tno.nl; r.kort@vu.nl

ABSTRACT

Background: A vaginal microbiota dominated by lactobacilli (particularly *Lactobacillus crispatus*) is associated with vaginal health, whereas a vaginal microbiota not dominated by lactobacilli is considered dysbiotic. Here we investigated whether *L. crispatus* strains isolated from the vaginal tract of women with *Lactobacillus*-dominated vaginal microbiota (LVM) are pheno- or genotypically distinct from *L. crispatus* strains isolated from vaginal samples with dysbiotic vaginal microbiota (DVM).

Results: We studied 33 *L. crispatus* strains (n=16 from LVM; n=17 from DVM). Comparison of these two groups of strains showed that, although strain differences existed, both groups were heterofermentative, produced similar amounts of organic acids, inhibited *Neisseria gonorrhoeae* growth and did not produce biofilms. Comparative genomics analyses of 28 strains (n=12 LVM; n=16 DVM) revealed a novel, 3-fragmented glycosyltransferase gene that was more prevalent among strains isolated from DVM. Most *L. crispatus* strains showed growth on glycogen-supplemented growth media. Strains that showed less efficient (n=6) or no (n=1) growth on glycogen all carried N-terminal deletions (respectively, 29 and 37 amino acid-deletions) in a putative pullulanase type I gene.

Discussion: *L. crispatus* strains isolated from LVM were not phenotypically distinct from *L. crispatus* strains isolated from DVM, however, the finding that the latter were more likely to carry a 3-fragmented glycosyltransferase gene may indicate a role for cell surface glycoconjugates, which may shape vaginal microbiota-host interactions. Furthermore, the observation that variation in the pullulanase type I gene associated with growth on glycogen discourages previous claims that *L. crispatus* cannot directly utilize glycogen.

INTRODUCTION

The vaginal mucosa hosts a community of commensal, symbiotic and sometimes pathogenic micro-organisms. Increasing evidence has shown that the bacteria within this community, referred to here as the vaginal microbiota (VM), play an important role in protecting the vaginal tract from pathogenic infection, which can have far reaching effects on a woman's sexual and reproductive health [1, 2]. Several VM compositions have been described, including VM dominated by: 1) *Lactobacillus iners;* 2) *L. crispatus;* 3) *L. gasseri;* 4) *L. jensenii* and; 5) VM that are not dominated by a single bacterial species but rather consist of diverse anaerobic bacteria, including *Gardnerella vaginalis* and members of Lachnospiraceae and Leptotrichiaceaeprevotella [3-5]. Particularly VM that are dominated by *L. crispatus* are associated with vaginal health, whereas a VM consisting of diverse anaerobes – commonly referred to as vaginal dysbiosis - have been shown to increase a woman's odds for developing bacterial vaginosis (BV), acquiring STI's, including HIV, and having an adverse pregnancy outcome [1, 2, 4, 6].

The application of human vaginal *L. crispatus* isolates as therapeutic agents to treat dysbiosis may have much potential [7, 8], but currently there are still many gaps in our knowledge concerning the importance of specific physiological properties of *L. crispatus* for a sustained domination on the mucosal surface of the vagina. Comparative genomics approaches offer a powerful tool to identify novel important physiological properties of bacterial strains. The genomes of nine human *L. crispatus* isolates have previously been studied, also in the context of vaginal dysbiosis [9, 10]. Comparative genomics of these strains showed that about 60% of orthologous groups (genes derived from the same ancestral gene) were conserved among all strains; i.e. comprising a 'core' genome [10]. The accessory genome was defined as genes shared by at least two strains, while unique genes are specific to a single strain. Currently it is unclear whether traits pertaining to *in vivo* dominance are shared by all strains (core genome), or only by a subset of strains (accessory genome). For example, both women with and without vaginal dysbiosis can be colonized with *L. crispatus* (see e.g.[11]) and we do not yet fully understand why in some women *L. crispatus* dominates and in others not.

The following bacterial traits may be of importance for *L. crispatus* to successfully dominate the vaginal mucosa: 1) the formation of an extracellular matrix (biofilm) on the vaginal mucosal surface; 2) the production of antimicrobials such as lactic acid, bacteriocins and H_2O_2 that inhibit the growth and/or adhesion of urogenital pathogens; 3) efficient utilization of available nutrients – particularly glycogen, as this is the main carbon source in the vaginal lumen; and; 4) the modulation of host-immunogenic responses. Considering these points, firstly, Ojala *et al.* [10] observed genomic islands encoding enzymes involved in exopolysacharide (EPS) biosynthesis in the accessory genome of *L. crispatus* and postulated that strain differences in this trait could contribute to differences in biofilm formation, adhesion and competitive exclusion of pathogens. Secondly, experiments have shown that *L. crispatus* effectively inhibits urogenital pathogens through lactic acid production, but these studies included only strains originating from healthy women [12-16]. Abdelmaksoud *et al.* [9] compared *L. crispatus* strains isolated from *Lactobacillus*-dominated VM (LVM) with strains isolated from dysbiotic VM (DVM) and indeed observed decreased lactic acid production in one of the strains isolated from DVM, providing an explanation for its low abundance. However, no significant conclusion could be made as their study included only eight strains. Thirdly, there is a general consensus that vaginal lactobacilli (including *L. crispatus*) ferment glycogen thus producing lactic acid, but no actual evidence exists that *L. crispatus* produces the enzymes to directly degrade glycogen [10, 17]. Lastly, *L. crispatus*-dominated VM are associated with an anti-inflammatory vaginal cytokine profile [18, 19] and immune evasion is likely a crucial (but poorly studied) factor that allows *L. crispatus* to dominate the vaginal niche. A proposed underlying mechanism is that *L. crispatus* produces immunomodulatory molecules [20], but *L. crispatus* may also accomplish immune modulation by alternating its cell surface glycosylation, as has been suggested for gut commensals [21]. Taken together, there is a clear need to study the properties of more human (clinical) *L. crispatus* isolates to fully appreciate the diversity within this species.

Here we investigated whether *L. crispatus* strains isolated from the vaginal tract of women with LVM are pheno- or genotypically distinct from *L. crispatus* strains isolated from vaginal samples with DVM, with the aim to identify bacterial traits pertaining to a successful domination of lactobacilli of the vaginal mucosa.

RESULTS

Lactobacillus crispatus strain selection and whole genome sequencing

For this study, 40 nurse-collected vaginal swabs were obtained from the Sexually Transmitted Infections clinic in Amsterdam, the Netherlands, from June to August 2012, as described previously by Dols *et al.* [4]. In total, 33 *L. crispatus* strains were isolated from these samples (n=16 from LVM samples; n=17 *L. crispatus* strains from DVM samples). Following whole genome sequencing, four contigs (n=3 strains from LVM; n=1 strains from DVM) were discarded as they had less than 50% coverage with other assemblies or with the reference genome (ST1), suggesting that these isolates belonged to a different *Lactobacillus* species. One contig (from a strain isolated from LVM) aligned to the reference genome, but its genome size was above the expected range, suggestive of contamination with a second strain and was therefore also discarded. The remaining 28 isolates (n=12 LVM and n=16 DVM) were assembled and used for comparative genomics. These genomes have been deposited at DDBJ/ENA/GenBank under the accession numbers NKKQ00000000-NKLR00000000. The versions described in this paper are versions NKKQ01000000-NKLR01000000 (Table 1).

Lactobacillus crispatus pan genome

The 28 *L. crispatus* genomes had an average length of 2.31 Mbp (range 2.16 – 2.56 MB) (Table 1), which was slightly larger than the reference genome (ST1; 2.04Mbp). The GC content of the genomes was on average 36.8%, similar to other lactobacilli [10]. An average of 2099 genes were annotated per strain (Table 1; Figure 1). This set of 28 *L. crispatus* genomes comprised 4261 different gene families. The core genome consisted of 1429 genes (which corresponds to ~68% of a given genome) and the accessory genome averaged at 618 genes (~30%) per strain. Each strain had on average 54 unique genes (~2.0%). The number of accessory and unique genes did not significantly differ between strains isolated from LVM or from DVM, with respectively an average of 621 (range: 481-855) and 55 (range: 5-243) genes for LVM strains and 615 (range: 488-837) and 53 (range: 1-250) genes for DVM strains. The distribution of cluster of ortholog groups (COG) also did not differ between strains from *Lactobacillus*-dominated and DVM. The gene accumulation model [22] describes the expansion of the pan-genome as function of the number of genomes and estimated that this species has access to a larger gene pool than described here; the model estimated the *L. crispatus* pan genome to include 4384 genes.

A fragmented glycosyltransferase gene was abundant among strains isolated from DVM

In a comparative genomics analysis we aimed to identify genes that were specific to strains isolated from either LVM or DVM. We observed that three transposases, one of which was further classified as an IS30 family transposase, were more abundant among strains isolated from DVM than among strains from LVM. IS30 transposases are associated with genomic instability and have previously been found to flank genomic deletions in commercial *L. rhamnosus* GG probiotic strains [23]. Most notably, we observed that strains from DVM were more likely to carry three gene fragments of a single glycosyltransferase (GT) than strains isolated from LVM. GTs are enzymes that are involved in the transfer of a sugar moiety to a substrate and are thus essential in synthesis of glycoconjugates like exopolysaccharides, glycoproteins and glycosylated teichoic acids [24, 25].

The three differentially abundant GT gene fragments all align to different regions of a family 2 A-fold GT of the ST1 *L. crispatus* strain (CGA_000165885.1) and are flanked by other genes potentially encoding GTs (Figure 2). Fragment 1 aligns with 472 bp of the original unfragmented GT, while fragment 2 overlaps with the last 3 bp of fragment 1 and fragment 3 overlaps 7 bp with fragment 2. Given that all these fragments align to the non-fragmented GT gene in in *L. crispatus* ST1, we hypothesize that the three fragments belong to the same GT. The *L. crispatus* genomes however contained a combination of one or more of the three GT fragments, while the surrounding genes were conserved among the strains. The first fragment of 510 bp contains the true GT fold domain and is thus responsible for the catalytic activity of the GT. The second and third fragment are considerably shorter, respectively 228 and 328 bp, and do not harbor any significant relation to a known GT-fold (Figure 3). Four different combinations of

GT fragments were observed in the studied genomes, namely a variant with: (1) no fragments, (2) all three fragments, (3) fragment 1 and 3, and (4) fragment 1 and 2 (Figure 2; Table 2).

Strains isolated from LVM were not phenotypically distinct from strains isolated from DVM

Phenotypic studies on the *L. crispatus* strains did not reveal any biofilm formation – as assessed by crystal violet assays, except for one strain (RL19) which produced a weak biofilm. In line with this, very low levels of autoaggregation (on average 5%) were observed and this also did not differ between the two groups of strains. Strain specific carbohydrate fermentation profiles were observed, as assessed by a commercial API CH50 test, but the distribution of these profiles did not relate to whether the strains were isolated from LVM or from DVM. Strains isolated from LVM produced similar amounts of organic acids compared with strains isolated from DVM when grown on chemically defined medium mimicking vaginal fluids [26]. The strains mainly produced lactic acid. Other acids such as succinate acid, butyric acid, glutamic acid, phenylalanine, isoleucine and tyrosine were also produced, but four-fold lower compared to lactic acid. Very small acidic molecules, such as acetic and propionic acid, were out of the detection range and could thus not be measured. We also assessed antimicrobial activity against a common urogenital pathogen *Neisseria gonorrhoeae.* Inhibition was similar for strains isolated from LVM and from DVM: *N. gonorrhoeae* growth was inhibited (i.e. lower OD_{600nm} in stationary phase compared to the control), in a dose-dependent way, by on average 27.9 ± 15.8% for undiluted *L. crispatus* supernatants compared to the *N. gonorrhoeae* control. Undiluted neutralized *L. crispatus* supernatants inhibited *N. gonorrhoeae* growth by on average 15.7 ± 16.3% (Supplementary information).

Strain-specific glycogen growth among both LVM and DVM isolates

Of the 28 strains for which full genomes were available, we tested 25 strains (n=12 LVM and n=13 DVM) for growth on glycogen. We compared growth on glucose-free NYCIII medium supplemented with glycogen as carbon source to growth on NYCIII medium supplemented with glucose (positive control) and NYCIII medium supplemented with water (negative control). All except one strain (RL05) showed growth on glycogen; however six strains showed substantially less efficient growth on glycogen. One strain showed a longer lag time (RL19; on average 4.5 hours, compared to an average of 1.5 hours for other strains) and five strains (RL02, RL06, RL07, RL09 and RL26) showed a lower OD after 36 hours of growth compared to other strains (Figure 4). Growth on glycogen did not correlate to whether the strain was isolated from LVM or DVM.

Growth on glycogen corresponded with variation in a putative pullulanase type I gene

We followed-up on the glycogen growth experiments with a gene-trait analysis as glycogen is considered to be a key, although disputed, nutrient (directly) available to *L. crispatus*. We searched the *L. crispatus* genomes for the presence/absence of enzymes that can potentially be involved in glycogen metabolism. We thus searched for orthologs of the: 1) glycogen debranching enzyme (encoded by *glgX*) in *Escherichia coli* [27, 28]; 2) *Streptococcus agalactiae* pullulanase [29]; 3) SusB of *Bacteroides thetaiotaomicron* [30]; and 4) the amylase (encoded by *amyE*) of *Bacillus subtilis* [31]. This search revealed a gene that was similar to the *glgX* gene; this gene was annotated as a pullulanase type I gene. In other species this pullulanase is bound to the outer S-layer of the cell wall, suggesting that this enzyme utilizes extracellular glycogen [32]. All except two strains (RL31, RL32) carried a copy of this gene. The genes are conserved except for variation in the N-terminal sequence that encodes a putative signal peptide that may be involved in subcellular localization of the enzyme. All strains with less efficient growth on glycogen had a 29 amino acid deletion in the N-terminal sequence (strains: RL02, RL06, RL07, RL09, RL19 and RL26) and the strain that showed no growth (RL05) had an 8 amino acid deletion in the same region as the other strains in addition to 37 amino acid deletion further downstream (Table 3).

DISCUSSION

Key findings of this paper

Here we report the full genomes of 28 *L. crispatus* clinical isolates; the largest contribution of *L. crispatus* clinical isolates to date. These strains were isolated from women with LVM and from women with DVM. A comparative genomics analysis revealed that a glycosyltransferase gene was more frequently found in the genomes of strains isolated from DVM as compared with strains isolated from LVM, suggesting a fitness advantage for carrying this gene in *L. crispatus* under dysbiotic conditions and a role of surface glycoconjugates in microbiota-host interactions. Comparative experiments pertaining to biofilm formation, antimicrobial activity and nutrient utilization showed that these two groups of strains did not phenotypically differ from each other. Of particular novelty value, we found that these clinical *L. crispatus* isolates were capable of growth on glycogen and that variation in a pullulanase type I gene correlates to the level of this activity.

Vaginal dysbiotic conditions may pressurize *Lactobacillus crispatus* to vary its glycome

Several studies have shown that vaginal dysbiosis is associated with an increased pro-inflammatory response, including an increase in pro-inflammatory chemokines and cytokines, but also elevated numbers of activated CD4+ T cells [3, 19], although no clinical signs of inflammation are present and vaginal dysbiosis is seen as a condition rather than as a disease [33]. Nonetheless, it indicates that the vaginal niche in a dysbiotic state is indeed under some immune pressure and that immune evasion could be a key (but poorly studied) trait for probiotic bacterial survival and dominance on the vaginal mucosa.

Our comparative genomics analysis revealed a glycosyltransferase gene (GT) gene that was more common in strains isolated from DVM compared with strains isolated from LVM. The identified GT consists of three fragments, which all align to a single GT in the reference *L. crispatus* genome (ST1). Sequence analyses showed that the first and longest fragment exhibits close homology to a known GT-A fold and most probably harbors the active site of the GT (Figure 3). The latter two fragments do not harbor any structural motifs resembling known GTs and most probably do not harbor any catalytic GT activity. We hypothesize that these two fragments play a role in steering the specific activity of the GT (e.g. towards donor or substrate specificity). This might point towards *L. crispatus* harnessing its genetic potential to change its surface glycome. Such a process is termed phase variation and allows bacteria to rapidly adapt and diversify their surface glycans, resulting in an evolutionary advantage in the arms race between the immune system and invading bacteria. Modulation of the surface glycome by phase variation of the GT coding sequence is a common immune evasion strategy, which has been extensively studied in pathogenic bacteria like *Campylobacter jejuni* [25], but could be utilized by commensals as well [21]. We hypothesize that *L. crispatus* in DVM exploits this genetic variation to allow for (a higher) variation in cell wall glycoconjugates providing a mechanism for *L. crispatus* to persist at low levels in DVM and remain stealth from the immune system (Figure 5). Of note, evidence for expression of all of the 3 GT-fragments comes from a recent transcriptomics study that studied the effect of metronidazole treatment on the VM of women with (recurring) BV [11]. Personal communication with Dr. Zhi-Luo Deng revealed that high levels of expression for the three putative GT peptides were present in the vaginal samples of two women who were responsive to treatment (i.e. their VM was fully restored to a *L. crispatus*-dominated VM following treatment). This finding is in line with our hypothesis that the presence of the fragmented GT gene has a selective advantage for *L. crispatus* under dysbiotic conditions. Further functional experiments are needed to test this hypothesized host-microbe interaction and to coin if and how the variation of glycoconjugates is affected by this GT. Additionally, the immunological response of the host must be further studied in reference to these hypothesized microbial adaptations. The bacterial surface glycome and related variability events are currently overlooked features in probiotic strain selection, while they might be crucial to a strain's survival and *in vivo* dominance [21].

No distinct phenotypes pertaining to dominance *in vivo* were observed

It has previously been postulated, relying merely on genomics data, that the accessory genome of *L. crispatus* could lead to strain differences relating to biofilm formation, adhesion and competitive exclusion of pathogens [9, 10]; all of which could influence whether a strain dominates the vaginal mucosa or not. Our comparative experimental work, however, showed that *L. crispatus* - irrespective of whether the strain was isolated from a woman with LVM or with DVM – all formed little to no biofilm, demonstrated effective lactic acid production and effective antimicrobial activity against *N. gonorrhoeae.* The previous genomic analyses also suggested that *L. crispatus* is herterofermentative [10]. Indeed, we observed that *L. crispatus* ferments a broad range of carbohydrates, as assessed by a commercial API test, but these profiles did not differ between strains isolated from LVM or from DVM.

First evidence showing that *Lactobacillus crispatus* grows on glycogen

The vaginal environment of healthy reproductive-age women is distinct from other mammals in that it has low microbial diversity, a high abundance of lactobacilli and high levels of lactic acid and luminal glycogen [34]. It has been postulated that proliferation of vaginal lactobacilli is supported by estrogen-driven glycogen production [35], however the 'fly in the ointment' - as finely formulated by Nunn *et al.* [17] - is that evidence for direct utilization of glycogen by vaginal lactobacilli is absent. Moreover, previous reports have stated that the core genome of *L. crispatus* does not contain the necessary enzymes to break down glycogen [10, 36]. It has even been suggested that *L. crispatus* relies on amylase secretion by the host or other microbes for glycogen breakdown [17, 37], as *L. crispatus* does contain all the appropriate enzymes to consume glycogen breakdown products such as glucose and maltose [36].

Here we provide the first evidence suggesting that *L. crispatus* human isolates are capable of growing on extracellular glycogen and we identified variation in a gene which correlated with this activity. The identified gene putatively encodes a pullulanase type I enzyme belonging to the glycoside hydrolase family 13 [38]. Its closest ortholog is an extracellular cell-attached pullulanase found in *L. acidophilus* [32]. The *L. crispatus* pullulanase gene described here carries three conserved domains, comprising an N-terminal carbohydrate-binding module family 41, a catalytic module belonging to the pullulanase super family and a C-terminal bacterial surface layer protein (SLAP) [39] (Figure 6). We observed that all except two of the strains in our study carry a copy of this gene. These two strains (RL31 and RL32), were no longer cultivable after their initial isolation. The six strains that showed less efficient or no growth on glycogen all showed variation in the N-terminal part of the pullulanase gene. All of these deletions are upstream of the carbohydrate-binding module in a sequence encoding a putative signal peptide. Furthermore, the presence of a SLAP-domain suggests that this enzyme is assigned to the outermost S-layer of the cell wall and is hence expected to be capable of degrading extracellular glycogen [32]. Further functional experiments are needed to fully characterize this pullulanase enzyme and to assess whether it degrades intra- or extracellular glycogen. Importantly, this pullulanase is likely part of a larger cluster of glycoproteins involved in glycogen metabolism in *L. crispatus,* which should be considered in future research.

Of note, we analyzed just one *L. crispatus* strain per vaginal sample, while it is plausible that multiple strain types co-exist in the vagina. So strain variability in growth on glycogen (and other carbohydrates) might actually benefit the *L. crispatus* population as a whole and explain the variation in growth on glycogen that we observed, especially considering that glycogen availability may fluctuate along with oscillating estrogen levels during the menstrual cycle. When developing probiotics, it could thus be beneficial to select for *L. crispatus* strains that ferment different carbohydrates (in addition to glycogen) [8] and also to supplement the probiotic with a prebiotic [40, 41].

Conclusion

Here we report whole-genome sequences of 28 *L. crispatus* human isolates. Our comparative study led to a total of three novel insights: 1) gene fragments encoding for a glycosyltransferase were disproportionally higher abundant among strains isolated from DVM, suggesting a role for cell surface glycoconjugates that shape vaginal microbiota-host interactions; 2) *L. crispatus* strains isolated from LVM do not differ from those isolated from DVM regarding the phenotypic traits studied here, including biofilm formation, pathogen inhibitory activity and carbohydrate utilization; and 3) *L. crispatus* is able to grow on glycogen and this correlates with the presence of a full-length pullulanase type I gene.

METHODS

L. crispatus strain selection

For this study, nurse-collected vaginal swabs were obtained from the Sexually Transmitted Infections clinic in Amsterdam, the Netherlands, from June to August 2012, as described previously by Dols *et al.* [4]. These vaginal samples came from women with LVM (Nugent score 0-3) and from women with DVM (Nugent score 7- 10). LVM and DVM vaginal swabs were plated on Trypton Soy Agar supplemented with 5% sheep serum, 0.25% lactic acid and pH set to 5.5 with acetic acid and incubated under microaerobic atmosphere (using an Anoxomat; Mart Microbiology B.V., the Netherlands) at 37°C for 48-72 hours. Candidate *Lactobacillus* spp. strains were selected based on colony morphology (white, small, smooth, circular, opaque colonies) and single colonies were subjected to 16S rRNA sequencing. One *L. crispatus* isolate per vaginal sample was taken forward for whole genome sequencing. A DNA library was prepared for these isolates using the Nextera XT DNA Library preparation kit and the genome was sequenced using the Illumina Miseq generate FASTQ workflow.

Genome assembly and quality control

All analyses were run on a virtual machine running Ubuntu version 16.02. Contigs were assembled using the Spades assembly pipeline [42]. Contigs were discarded if they had less than 50% coverage with other assemblies or with the reference genome (N50 and NG50 values deviated more than 3 standard deviations from the mean as determined using QUAST [43]. The genomes were assembled with Spades 3.5.0 using default settings. The Spades pipeline integrates read-error correction, iterative k-mer (nucleotide sequences of length k) based short read assembling and mismatches correction. The quality of the assemblies was determined with Quast (History 2013) using default settings and the *Lactobacillus crispatus* ST1 strain as reference genome (Genbank FN692037).

Genome annotation and comparative genome analysis

After assembly, the generated contigs were sorted with Mauve contig mover [44], using the *L. crispatus* ST1 strain as reference genome. Contaminating sequences of human origin and adaptor sequences were identified using BLAST and manually removed. The reordered genomes were annotated using the Prokka automated annotation pipeline [45] using default settings. Additionally, the genomes were uploaded to Genbank and annotated using the NCBI integrated Prokaryotic Genome Annotation Pipeline [46]. The annotated genomes were analyzed using the Sequence element enrichment analysis (SEER), which looks for an association between enriched k-mers and a certain phenotype [47]. Following the developer's instructions, the genomes were split into k-mers using fsm-lite on standard settings and a minimum k-mer frequency of 2 and a maximum frequency of 28. The usage of k-mers enables the software to look for both SNPs as well as gene variation at the same time. After k-mer counting, the resulting file was split into 16 equal parts and g-zipped for parallelization purposes. In order to correct for the clonal population structure of bacteria, the population structure was estimated using Mash with default settings [48]. Using SEER, we looked for k-mers of various lengths that associated with whether the *L. crispatus* strains came from LVM or DVM. The results were filtered for k-mers with a chi-square test of association of <0.01 and a likelihood-ratio test p-value (a statistical test for the goodness of fit for two models) of <0.0001. The resulting list of k-mers was sorted by likelihood-ratio p and the top 50 hits were manually evaluated using BLASTx and BLASTn.

Pan and accessory genome analysis

We used the bacterial pan genome analysis tool developed by Chaudhari *et al.* [49] using default settings. The circular image was created using CGview Comparison Tool [50] by running the build_blast_atlas_all_vs_all.sh script included in the package.

Comparative phenotype experiments

Not all strains were (consistently) cultivable after their initial isolation, so experimental data was collected for a subset of the strains and could differ per experiment. The ratio of cultivable LVM and DVM strains was however similar for each experiment. For a full overview of experimental procedures, we refer to the Supplementary Information. In short, carbohydrate metabolism profiles were assessed using commercial API CH50 carbohydrate fermentation tests (bioMérieux, Inc., Marcy l'Etoile, France) according to the manufacturer's protocol. To assess organic acid production, strains were grown on medium that mimicked vaginal secretions [26]. Total metabolite extracts from spent medium were assessed as previously described by Collins *et al.* [41]. Biofilm formation was assessed using the crystal violet assay as described by Santos *et al.* [51] and auto-aggregation as described by Younes *et al.* [52]. Antimicrobial activity against *Neisseria gonorrhoeae* was assessed by challenging *N. gonorrhoeae* (WHO-L strain) with varying (neutralized with NaOH to pH 7.0) dilutions of *L. crispatus* supernatants. Inhibitory effect was assessed as percentile difference in OD_{600nm} in a conditional stationary phase as compared to the control.

Glycogen degradation assay

Starter cultures were grown in regular NYCIII glucose medium for 72 hours. For this assay, 1.1x carbohydrate deprived NYCIII medium was supplemented with water (negative control), 5% glucose (positive control) or 5% glycogen (Sigma-Aldrich, Saint Louis, US) and subsequently inoculated with 10% (v/v) bacterial culture (OD~0.5; 10^9 CFU/ml). Growth on glycogen was compared to growth on NYCII without supplemented carbon source and to NYCIII with glucose. Growth curves were followed in a BioScreen (Labsystems, Helsinki, Finland). At least two independent experiments per strain were performed in triplicate.

LIST OF ABBREVIATIONS

VM: vaginal microbiota

LVM: *Lactobacillus*-dominated vaginal microbiota

DVM: dysbiotic vaginal microbiota

COG: cluster ortholog genes

GT: glycosyltransferase

TSB: Trypton Soya Broth

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The research proposed in this study was evaluated by the ethics review board of the Academic Medical Center (AMC), University of Amsterdam, The Netherlands. According to the review board no additional ethical approval was required for this study, as the vaginal samples used here were collected as part of routine procedure for cervical examinations at the STI clinic in Amsterdam (document reference number W12_086 # 12.17.0104). Clients of the STI clinic were notified that remainders of their samples could be used for scientific research, after anonymisation of client clinical data and samples. If the clients objected, their data and samples were discarded. This procedure has been approved by the AMC ethics review board (reference number W15_159 # 15.0193).

CONSENT FOR PUBLICATION

Clients of the STI clinic were notified that remainders of their samples could be used for scientific research, after anonymisation of client clinical data and samples. If the clients objected, their data and samples were discarded. This procedure has been approved by the AMC ethics review board (reference number W15_159 # 15.0193).

AVAILABILITY OF DATA AND MATERIAL

The 28 *Lactobacillus crispatus* sequenced genomes described in this paper have been deposited at DDBJ/ENA/GenBank under the accessions NKKQ00000000-NKLR00000000.

COMPETING INTERESTS

The authors declare no conflict of interest.

FUNDING

This research was funded by Public Health Service Amsterdam (GGD), the VU University of Amsterdam (VU) and the Netherlands Organization for Applied Scientific Research (TNO). HT holds a Marie Sklodowska-Curie fellowship of the European Union's Horizon 2020 research and innovation program under agreement No 703577 (Glycoli) to support her work at ETH Zurich.

AUTHORS' CONTRIBUTIONS

RK, SB, HdV and FS conceptualized the study. CV and JS performed the experimental work, supervised by AdKA, SB and RK. JS performed the bio-informatic analyses, supervised by DW and RK. RH did the initial glycogen finding and provided further expertise. HT provided expertise for the glycosyltransferase finding and GR for the potential of probiotic applications. CV drafted the manuscript. All authors contributed to and approved the final manuscript.

ACKNOWLEDGMENTS

We thank Dr. Titia Heijman of the Sexually Transmitted Infections clinic in Amsterdam, the Netherlands, for organizing the collection of the clinical vaginal samples. We thank Liesbeth Hoekman (TNO) for isolation and initial characterization of *Lactobacillus crispatus* strains. We thank Mark Sumarah and Justin Renaud for facilitating the metabolomics analysis. We also thank Dr. Zhi-Luo Deng for mining his transcriptomics data for the GT gene fragments and pullulanase gene.

REFERENCES

1. DiGiulio DB, Callahan BJ, McMurdie PJ, Costello EK, Lyell DJ, Robaczewska A, Sun CL, Goltsman DS, Wong RJ, Shaw G *et al*: **Temporal and spatial variation of the human microbiota during pregnancy**. *Proc Natl Acad Sci U S A* 2015, **112**(35):11060-11065.
2. Tamarelle J, Thiebaut ACM, de Barbeyrac B, Bebear C, Ravel J, Delarocque-Astagneau E: **The vaginal microbiota and its association with Human Papillomavirus, *Chlamydia trachomatis*, *Neisseria gonorrhea* and *Mycoplasma genitalium* infections: a systematic review and meta-analysis**. *Clin Microbiol Infect* 2018.
3. Borgdorff H, van der Veer C, van Houdt R, Alberts CJ, de Vries HJ, Bruisten SM, Snijder MB, Prins M, Geerlings SE, Schim van der Loeff MF *et al*: **The association between ethnicity and vaginal microbiota composition in Amsterdam, the Netherlands**. *PLoS One* 2017, **12**(7):e0181135.
4. Dols JA, Molenaar D, van der Helm JJ, Caspers MP, de Kat Angelino-Bart A, Schuren FH, Speksnijder AG, Westerhoff HV, Richardus JH, Boon ME *et al*: **Molecular assessment of bacterial vaginosis by *Lactobacillus* abundance and species diversity**. *BMC Infect Dis* 2016, **16**:180.
5. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO *et al*: **Vaginal microbiome of reproductive-age women**. *Proc Natl Acad Sci U S A* 2011, **108 Suppl 1**:4680-4687.
6. van der Veer C, Bruisten SM, van der Helm JJ, de Vries HJ, van Houdt R: **The Cervicovaginal Microbiota in Women Notified for *Chlamydia trachomatis* Infection: A Case-Control Study at the Sexually Transmitted Infection Outpatient Clinic in Amsterdam, The Netherlands**. *Clin Infect Dis* 2017, **64**(1):24-31.
7. Kort R: **Personalized therapy with probiotics from the host by TripleA**. *Trends Biotechnol* 2014, **32**(6):291-293.
8. Kort R, van der Veer C: **A new probiotic composition for the prevention of bacterial vaginosis.** *European Patent 17181005* 2017.
9. Abdelmaksoud AA, Koparde VN, Sheth NU, Serrano MG, Glascock AL, Fettweis JM, Strauss JF, 3rd, Buck GA, Jefferson KK: **Comparison of *Lactobacillus crispatus* isolates from *Lactobacillus*-dominated vaginal microbiomes with isolates from microbiomes containing bacterial vaginosis-associated bacteria**. *Microbiology* 2016, **162**(3):466-475.
10. Ojala T, Kankainen M, Castro J, Cerca N, Edelman S, Westerlund-Wikstrom B, Paulin L, Holm L, Auvinen P: **Comparative genomics of *Lactobacillus crispatus* suggests novel mechanisms for the competitive exclusion of *Gardnerella vaginalis***. *BMC Genomics* 2014, **15**:1070.
11. Deng ZL, Gottschick C, Bhuju S, Masur C, Abels C, Wagner-Dobler I: **Metatranscriptome Analysis of the Vaginal Microbiota Reveals Potential Mechanisms for Protection against Metronidazole in Bacterial Vaginosis**. *mSphere* 2018, **3**(3).
12. Atassi F, Brassart D, Grob P, Graf F, Servin AL: ***Lactobacillus* strains isolated from the vaginal microbiota of healthy women inhibit *Prevotella bivia* and *Gardnerella vaginalis* in coculture and cell culture**. *FEMS Immunol Med Microbiol* 2006, **48**(3):424-432.
13. Foschi C, Salvo M, Cevenini R, Parolin C, Vitali B, Marangoni A: **Vaginal lactobacilli reduce *Neisseria gonorrhoeae* viability through multiple strategies: An *in vitro* study**. *Front Cell Infect Microbiol* 2017, **7**:502.
14. Gong Z, Luna Y, Yu P, Fan H: **Lactobacilli inactivate *Chlamydia trachomatis* through lactic acid but not H2O2**. *PLoS One* 2014, **9**(9):e107758.
15. Graver MA, Wade JJ: **The role of acidification in the inhibition of *Neisseria gonorrhoeae* by vaginal lactobacilli during anaerobic growth**. *Ann Clin Microbiol Antimicrob* 2011, **10**:8.
16. Nardini P, Nahui Palomino RA, Parolin C, Laghi L, Foschi C, Cevenini R, Vitali B, Marangoni A: ***Lactobacillus crispatus* inhibits the infectivity of *Chlamydia trachomatis* elementary bodies, *in vitro* study**. *Sci Rep* 2016, **6**:29024.
17. Nunn KL, Forney LJ: **Unraveling the Dynamics of the Human Vaginal Microbiome**. *Yale J Biol Med* 2016, **89**(3):331-337.
18. Borgdorff H, Gautam R, Armstrong SD, Xia D, Ndayisaba GF, van Teijlingen NH, Geijtenbeek TB, Wastling JM, van de Wijgert JH: **Cervicovaginal microbiome dysbiosis is associated with proteome changes related to alterations of the cervicovaginal mucosal barrier**. *Mucosal Immunol* 2016, **9**(3):621-633.

19. Gosmann C, Anahtar MN, Handley SA, Farcasanu M, Abu-Ali G, Bowman BA, Padavattan N, Desai C, Droit L, Moodley A *et al*: ***Lactobacillus*-deficient dervicovaginal bacterial communities are associated with Increased HIV Acquisition in young South African women**. *Immunity* 2017, **46**(1):29-37.

20. Witkin SS, Mendes-Soares H, Linhares IM, Jayaram A, Ledger WJ, Forney LJ: **Influence of vaginal bacteria and D- and L-lactic acid isomers on vaginal extracellular matrix metalloproteinase inducer: implications for protection against upper genital tract infections**. *MBio* 2013, **4**(4).

21. Tytgat HLP, de Vos WM: **Sugar Coating the Envelope: Glycoconjugates for Microbe-Host Crosstalk**. *Trends Microbiol* 2016, **24**(11):853-861.

22. Tettelin H, Riley D, Cattuto C, Medini D: **Comparative genomics: the bacterial pan-genome**. *Curr Opin Microbiol* 2008, **11**(5):472-477.

23. Sybesma W, Molenaar D, van IW, Venema K, Kort R: **Genome instability in *Lactobacillus rhamnosus* GG**. *Appl Environ Microbiol* 2013, **79**(7):2233-2239.

24. Lairson LL, Henrissat B, Davies GJ, Withers SG: **Glycosyltransferases: structures, functions, and mechanisms**. *Annu Rev Biochem* 2008, **77**:521-555.

25. Tytgat HL, Lebeer S: **The sweet tooth of bacteria: common themes in bacterial glycoconjugates**. *Microbiol Mol Biol Rev* 2014, **78**(3):372-417.

26. Geshnizgani AM, Onderdonk AB: **Defined medium simulating genital tract secretions for growth of vaginal microflora**. *J Clin Microbiol* 1992, **30**(5):1323-1326.

27. Strydom L, Jewell J, Meier MA, George GM, Pfister B, Zeeman S, Kossmann J, Lloyd JR: **Analysis of genes involved in glycogen degradation in *Escherichia coli***. *FEMS Microbiol Lett* 2017, **364**(3).

28. Dauvillee D, Kinderf IS, Li Z, Kosar-Hashemi B, Samuel MS, Rampling L, Ball S, Morell MK: **Role of the *Escherichia coli glgX* gene in glycogen metabolism**. *J Bacteriol* 2005, **187**(4):1465-1473.

29. Santi I, Pezzicoli A, Bosello M, Berti F, Mariani M, Telford JL, Grandi G, Soriani M: **Functional characterization of a newly identified group B *Streptococcus* pullulanase eliciting antibodies able to prevent alpha-glucans degradation**. *PLoS One* 2008, **3**(11):e3787.

30. Kitamura M, Okuyama M, Tanzawa F, Mori H, Kitago Y, Watanabe N, Kimura A, Tanaka I, Yao M: **Structural and functional analysis of a glycoside hydrolase family 97 enzyme from *Bacteroides thetaiotaomicron***. *J Biol Chem* 2008, **283**(52):36328-36337.

31. Yamazaki H, Ohmura K, Nakayama A, Takeichi Y, Otozai K, Yamasaki M, Tamura G, Yamane K: **Alpha-amylase genes (amyR2 and amyE+) from an alpha-amylase-hyperproducing *Bacillus subtilis* strain: molecular cloning and nucleotide sequences**. *J Bacteriol* 1983, **156**(1):327-337.

32. Moller MS, Goh YJ, Rasmussen KB, Cypryk W, Celebioglu HU, Klaenhammer TR, Svensson B, Abou Hachem M: **An extracellular cell-attached pullulanase confers branched alpha-glucan utilization in human gut *Lactobacillus acidophilus***. *Appl Environ Microbiol* 2017, **83**(12).

33. Reid G: **Is bacterial vaginosis a disease?** *Appl Microbiol Biotechnol* 2018, **102**(2):553-558.

34. Petrova MI, van den Broek M, Balzarini J, Vanderleyden J, Lebeer S: **Vaginal microbiota and its role in HIV transmission and infection**. *FEMS Microbiol Rev* 2013, **37**(5):762-792.

35. Mirmonsef P, Hotton AL, Gilbert D, Burgad D, Landay A, Weber KM, Cohen M, Ravel J, Spear GT: **Free glycogen in vaginal fluids is associated with *Lactobacillus* colonization and low vaginal pH**. *PLoS One* 2014, **9**(7):e102467.

36. France MT, Mendes-Soares H, Forney LJ: **Genomic comparisons of *Lactobacillus crispatus* and *Lactobacillus iners* reveal potential ecological drivers of community composition in the vagina**. *Appl Environ Microbiol* 2016, **82**(24):7063-7073.

37. Spear GT, French AL, Gilbert D, Zariffard MR, Mirmonsef P, Sullivan TH, Spear WW, Landay A, Micci S, Lee BH *et al*: **Human alpha-amylase present in lower-genital-tract mucosal fluid processes glycogen to support vaginal colonization by *Lactobacillus***. *J Infect Dis* 2014, **210**(7):1019-1028.

38. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B: **The carbohydrate-active enzymes database (CAZy) in 2013**. *Nucleic Acids Res* 2014, **42**(Database issue):D490-495.

39. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI *et al*: **CDD: NCBI's conserved domain database**. *Nucleic Acids Res* 2015, **43**(Database issue):D222-226.

40. Gibson GR, Hutkins R, Sanders ME, Prescott SL, Reimer RA, Salminen SJ, Scott K, Stanton C, Swanson KS, Cani PD *et al*: **Expert consensus document: The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of prebiotics**. *Nat Rev Gastroenterol Hepatol* 2017, **14**(8):491-502.

41. Collins SL, McMillan A, Seney S, van der Veer C, Kort R, Sumarah MW, Reid G: **Promising prebiotic candidate established by evaluation of lactitol, lactulose, raffinose, and oligofructose for maintenance of a *Lactobacillus*-dominated vaginal microbiota**. *Appl Environ Microbiol* 2018, **84**(5).

42. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al*: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing**. *J Comput Biol* 2012, **19**(5):455-477.

43. Gurevich A, Saveliev V, Vyahhi N, Tesler G: **QUAST: quality assessment tool for genome assemblies**. *Bioinformatics* 2013, **29**(8):1072-1075.

44. Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT: **Reordering contigs of draft genomes using the Mauve aligner**. *Bioinformatics* 2009, **25**(16):2071-2073.

45. Seemann T: **Prokka: rapid prokaryotic genome annotation**. *Bioinformatics* 2014, **30**(14):2068-2069.

46. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J: **NCBI prokaryotic genome annotation pipeline**. *Nucleic Acids Res* 2016, **44**(14):6614-6624.

47. Lees JA, Vehkala M, Valimaki N, Harris SR, Chewapreecha C, Croucher NJ, Marttinen P, Davies MR, Steer AC, Tong SY *et al*: **Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes**. *Nat Commun* 2016, **7**:12797.

48. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM: **Mash: fast genome and metagenome distance estimation using MinHash**. *Genome Biol* 2016, **17**(1):132.

49. Chaudhari NM, Gupta VK, Dutta C: **BPGA- an ultra-fast pan-genome analysis pipeline**. *Sci Rep* 2016, **6**:24373.

50. Grant JR, Arantes AS, Stothard P: **Comparing thousands of circular genomes using the CGView Comparison Tool**. *BMC Genomics* 2012, **13**:202.

51. Santos CM, Pires MC, Leao TL, Hernandez ZP, Rodriguez ML, Martins AK, Miranda LS, Martins FS, Nicoli JR: **Selection of *Lactobacillus* strains as potential probiotics for vaginitis treatment**. *Microbiology* 2016, **162**(7):1195-1207.

52. Younes JA, van der Mei HC, van den Heuvel E, Busscher HJ, Reid G: **Adhesion forces and coaggregation between vaginal staphylococci and lactobacilli**. *PLoS One* 2012, **7**(5):e36917.

Table 1. Overview and properties of 28 *L. crispatus* strains isolated from vaginal swabs with *Lactobacillus*-dominated vaginal microbiota or dysbiotic vaginal microbiota.

Strain information		**Clinical information vaginal sample**				**Pan-genome overview**				
Accession no.	ID	Group	Nugent score	VM Cluster [4]	Urogenital infection	Genome size (Mb)	GC content	No. of core genes	No. of accessory genes	No. of unique genes
NKLQ00000000	RL03	LVM	0	II	None	2.52	36.86	1429	846	12
NKLP00000000	RL05	LVM	0	II	None	2.53	36.39	1429	553	243
NKLO00000000	RL06	LVM	0	II	None	2.16	36.92	1429	481	11
NKLM00000000	RL08	LVM	0	I	None	2.25	36.82	1429	606	43
NKLL00000000	RL09	LVM	0	II	None	2.25	36.83	1429	559	21
NKLK00000000	RL10	LVM	0	I	None	2.15	36.91	1429	612	31
NKLJ00000000	RL11	LVM	0	II	None	2.17	36.90	1429	482	5
NKLF00000000	RL16	LVM	3	II	None	2.56	36.49	1429	855	27
NKKX00000000	RL26	LVM	3	II	None	2.21	36.90	1429	525	103
NKKW00000000	RL27	LVM	3	I	None	2.51	36.84	1429	815	78
NKKU00000000	RL29	LVM	2	II	None	2.20	36.88	1429	501	44
NKKR00000000	RL32	LVM	1	II	CA	2.34	36.97	1429	644	63
NKLR00000000	RL02	DVM	9	III	None	2.22	36.88	1429	528	13
NKLN00000000	RL07	DVM	10	IV	None	2.16	36.94	1429	498	6
NKLI00000000	RL13	DVM	9	V	None	2.19	36.89	1429	488	28
NKLH00000000	RL14	DVM	9	V	None	2.56	36.76	1429	837	63
NKLG00000000	RL15	DVM	8	V	CT	2.27	36.79	1429	593	74
NKLE00000000	RL17	DVM	8	III	None	2.31	37.08	1429	605	250
NKLD00000000	RL19	DVM	8	V	None	2.41	36.93	1429	527	117
NKLC00000000	RL20	DVM	10	III	Candida	2.49	36.47	1429	660	41
NKLB00000000	RL21	DVM	9	V	None	2.49	36.79	1429	807	72

NKLA00000000	RL23	DVM	10	III	None	2.30	36.84	1429	621	1
NKKZ00000000	RL24	DVM	9	III	None	2.37	36.72	1429	682	9
NKKY00000000	RL25	DVM	9	V	None	2.32	36.84	1429	618	16
NKKV00000000	RL28	DVM	10	IV	None	2.17	36.88	1429	489	63
NKKT00000000	RL30	DVM	10	IV	None	2.27	36.76	1429	603	20
NKKS00000000	RL31	DVM	10	IV	CA	2.31	36.93	1429	652	48
NKKQ00000000	RL33	DVM	8	I†	TV	2.37	36.73	1429	631	31

VM: vaginal microbiota; LVM: *Lactobacillus*-dominated VM; DVM: dysbiotic VM; CT: *Chlamydia trachomatis;* CA: Condylomata accuminata TV: *Trichomonas vaginalis;* VM clusters: I-*L. iners;* II-*L. crispatus;* III-*G. vaginalis*-Sneathia; IV-Sneathia-Lachnospiraceae; V-Sneathia

† This sample clustered together with *L. iners*-dominated samples, but contained many reads belonging to BV-associated bacteria.

Table 2. Comparison of distribution of glycosyltransferase (GT) gene fragments in *Lactobacillus crispatus* genomes isolated from vaginal samples with *Lactobacillus*-dominated or dysbiotic vaginal microbiota.

	LVM	**DVM**	**p-value***
	N = 12 (%)	**N = 16 (%)**	
No GT fragments	6 (50.0)	3 (18.8)	0.114
1st and 2nd GT fragments	3 (25.0)	3 (18.8)	1.000
1st and 3rd GT fragment	1 (8.3)	0 (0.0)	0.429
All 3 GT fragments	2 (16.6)	10 (62.5)	**0.023**

LVM: *Lactobacillus*-dominated VM; DVM: dysbiotic VM

* Fisher's Exact test.

Table 3. Overview of *Lactobacillus crispatus* strain specific growth on glycogen and corresponding translated amino acid sequence at the N-terminal of a pullulanase type I gene.

Strain ID	Group	Growth on glycogen	Pullulanase Type I amino acid sequence (N-terminal)
RL3	LVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL5	LVM	-	M________NKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAP_____________________________________PQNVPTVLAA
RL6	LVM	+/-	M_____________________________SLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL8	LVM	NA	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL9	LVM	+/-	M_____________________________SLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL10	LVM	NA	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL11	LVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL16	LVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL22†	LVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL26	LVM	+/-	M_____________________________SLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL27	LVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL29	LVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL32	LVM	NC	---
RL2	DVM	+/-	M_____________________________SLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL7	DVM	+/-	M_____________________________SLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL13	DVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL14	DVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL15	DVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL17	DVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL19	DVM	EL	M_____________________________SLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL20	DVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL21	DVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL23	DVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA

RL24	DVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL25	DVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL28	DVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL30	DVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA
RL31	DVM	NC	---
RL33	DVM	+	MILWRNLFMNKKSGHNIKFKSIFVCTSAIMSLWLGANLTTTQVHAAEDNAAPKSSEVVGQTNSSKDNAATATVQNQSNAKAKQRQQGVAPQNVPTVLAA

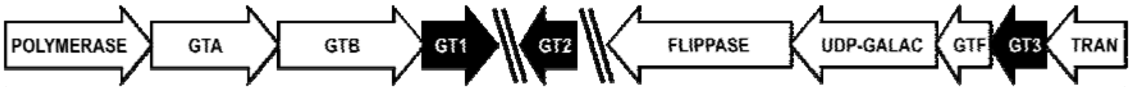
LVM: *Lactobacillus*-dominated vaginal microbiota; DVM: dysbiotic vaginal microbiota; NA: not available; NC: non-cultivable; EL: extended lag time.

† The genome of RL22 was not deposited in GenBank as the sequencing depth was too low and the N50 and NG50 values gave an inconclusive image of the assembly's quality.

FIGURES

Figure 1. Whole genome alignments of the coding sequences from the *Lactobacillus crispatus* clinical isolates described in this study. The outermost ring represents COG annotated genes on the forward strand (color coded according to the respective COG). The positions of the genes discussed in this article are indicated. The third ring represents COG annotated genes on the reverse strand (color coded according to the respective COG). The next twelve rings each represent one genome of the LVM strains, followed by a separator ring and 16 rings each representing a genome of the DVM strains. The height of the bar and the saturation of the color in these rings indicate a BLAST hit of either >90% identity (darker colored) or >70% identity (lightly colored). Hits below 70% identity score are not shown and appear as white bars in the plots. The two inner most rings represent the GC content of that area and the GC-skew respectively. The presence or absence of the gene variants discussed in this article is indicated in each genome by black and white dots. A black dot indicates that a wild-type gene (as compared to the STI reference genome) is present in that genome, a white dot indicates that no copy of that gene (fragment) was present or that it carried a deletion (for the type 1 pullulanase). Abbreviations: COG: cluster ortholog genes; LVM: *Lactobacillus*-dominated vaginal microbiota; DVM: dysbiotic vaginal microbiota; WT: wild type.

Figure 2. Schematic overview of the organization of the glycosyltransferase fragments in the *Lactobacillus crispatus* genomes. The orientation of the fragments is dependent on the assembly, and can therefore be different than depicted here. Also, the distance between the fragments is undetermined and can be of any length (depicted with diagonal lines). Abbreviations: GT: Glycosyltransferase; GTA, GTB: GT super families; GT1, GT2, GT3: GT fragments 1, 2, 3; UDP-GALAC: UDP-Galactopyranose mutase; GTF: GT family 1; TRAN: transposase; LVM: *Lactobacillus*-dominated vaginal microbiota; DVM: dysbiotic vaginal microbiota.



Strain ID	Group	GT1	GT2	GT3
RL03	LVM	-	-	-
RL05	LVM	-	-	-
RL06	LVM	+	+	-
RL08	LVM	+	+	-
RL09	LVM	-	-	-
RL10	LVM	-	-	-
RL11	LVM	+	+	+
RL16	LVM	+	+	-
RL26	LVM	+	-	+
RL27	LVM	-	-	-
RL29	LVM	+	+	+
RL32	LVM	-	-	-
RL02	DVM	-	-	-
RL7	DVM	+	+	-
RL13	DVM	+	+	+
RL14	DVM	-	-	-
RL15	DVM	+	+	+
RL17	DVM	+	+	-
RL19	DVM	+	+	-
RL20	DVM	+	+	+
RL21	DVM	-	-	-
RL23	DVM	+	+	+
RL24	DVM	+	+	+
RL25	DVM	+	+	+
RL28	DVM	+	+	+
RL30	DVM	+	+	+
RL31	DVM	+	+	+
RL33	DVM	+	+	+

Figure 3. Schematic overview of how the glycosyltransferase fragments align to the *Lactobacillus crispatus* ST1 reference genome. The first fragment comprises the conserved glycosyltransferase family 2 domain with catalytic activity. The shorter second and third fragments most probably do not harbor any catalytic GT activity. We hypothesize that these two fragments play a role in steering the specific activity of the GT (e.g. towards donor or substrate specificity). Abbreviation: GT: glycosyltransferase.

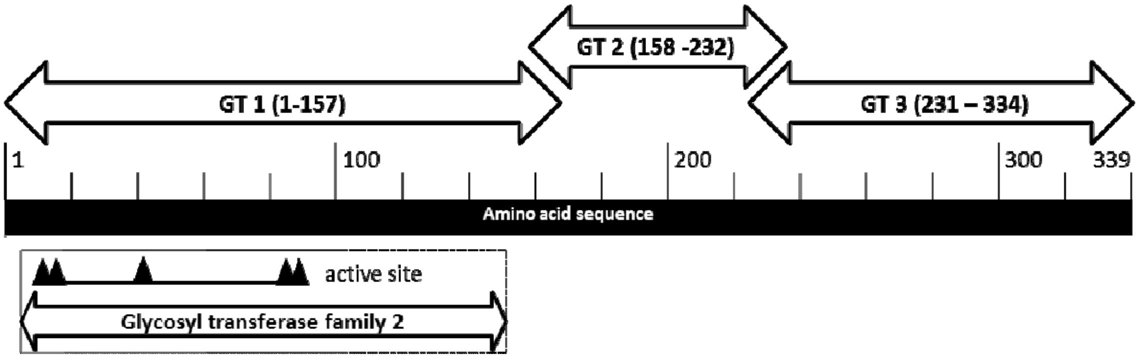


Figure 4. Growth on glycogen for *Lactobacillus crispatus* strains isolated from *Lactobacillus*-dominated and from dysbiotic vaginal microbiota. Strains were grown in minimal medium supplemented with A) 5% glucose and B) 5% glycogen. Strains that showed less efficient or no growth on glycogen carried a mutation in the N-terminal sequence of a putative type I pullulanase gene. RL19 showed a longer lag time compared to other strains; on average 4.5 hours, compared to an average of 1.5 hours for other strains. Abbreviations: LVM: *Lactobacillus*-dominated vaginal microbiota; DVM: dysbiotic vaginal microbiota; WT: wild type.

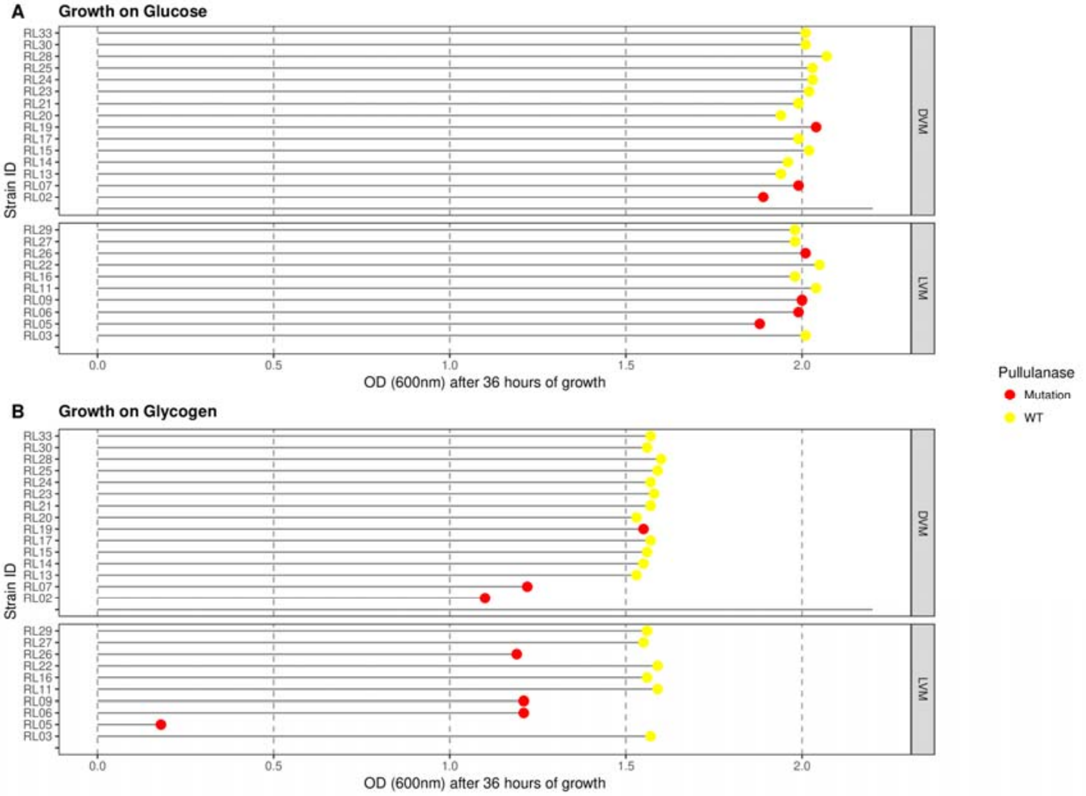


Figure 5. Model for enzymatic activity in glycosylation and glycogen degradation in *Lactobacillus crispatus*. Schematic representation of the vaginal environment with either LVM or DVM. Our comparative genomics analysis revealed a glycosyltransferase gene that was more common in *Lactobacillus crispatus* strains isolated from LVM (red bacteria) and DVM (low abundance of red lactobacilli, diverse bacterial population in multiple colors and forms, thinner mucus layer). We hypothesize that *L. crispatus* in DVM exploits this genetic variation to allow for (a higher) variation in cell wall glycoconjugates providing a mechanism for *L. crispatus* to persist at low levels in DVM and remain stealth from the immune system. Another finding of this work describes the ability of *L. crispatus* strains to utilize glycogen as a food source, which is associated with the presence of a full-length pullulanase gene (red dots on cell wall of *L. crispatus*). Abbreviations: LVM: *Lactobacillus*-dominated vaginal microbiota; DVM: dysbiotic vaginal microbiota, LC: Langerhans cell, CK: cytokines.

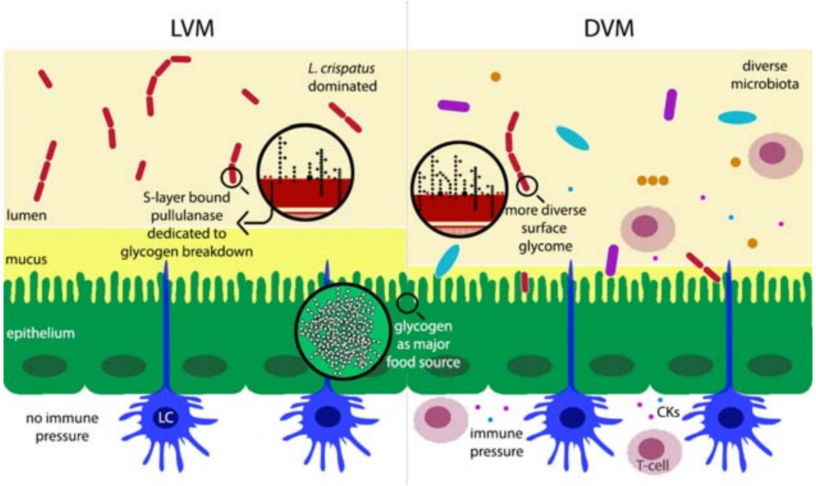
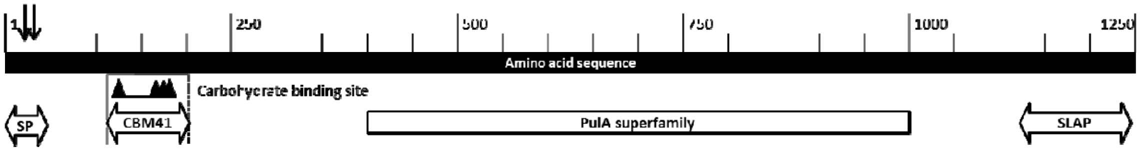
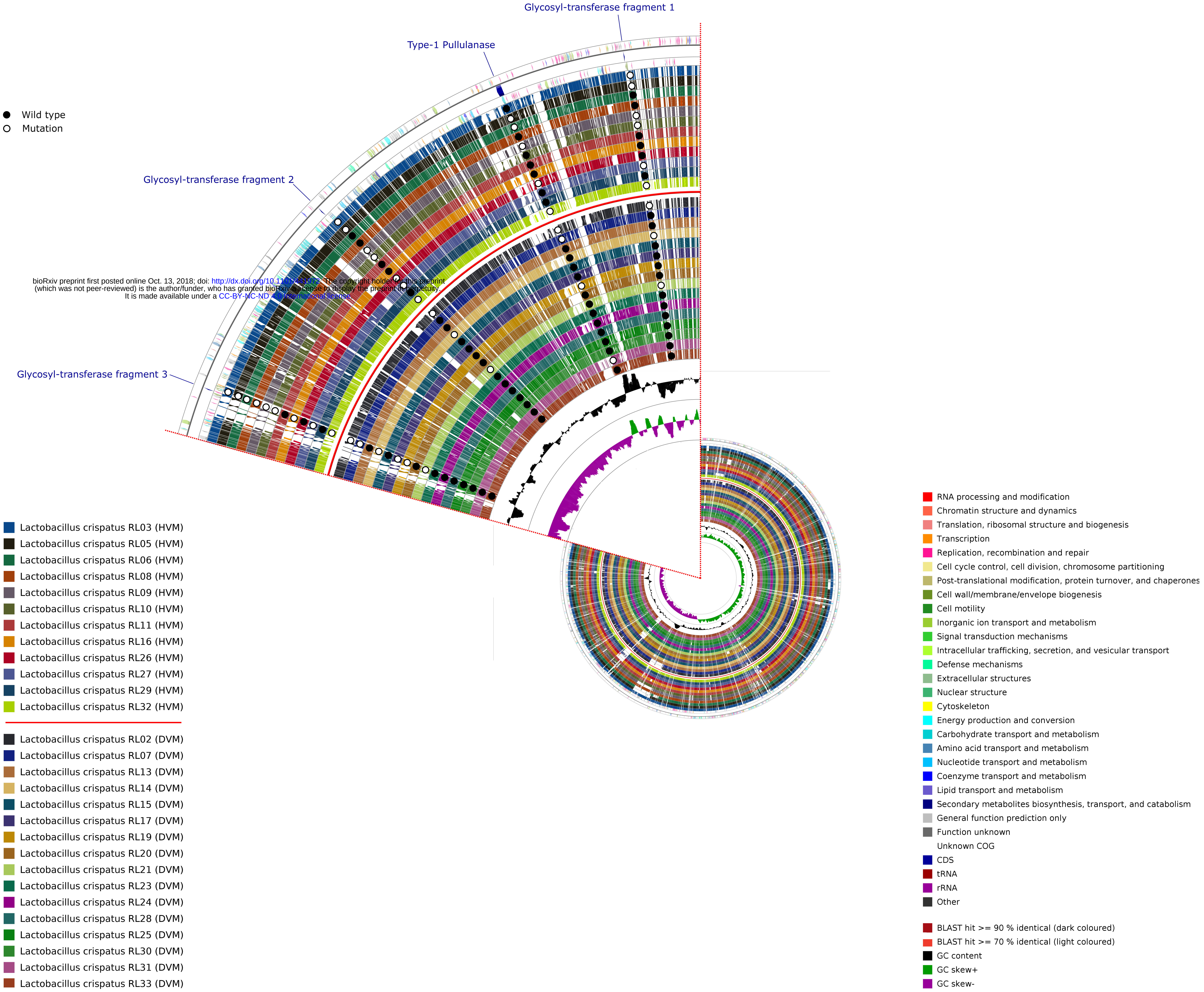

LVM
DVM
L. crispatus dominated
lumen
S-layer bound pullulanase dedicated to glycogen breakdown
mucus
epithelium
glycogen as major food source
no immune pressure
LC
diverse microbiota
more diverse surface glycome
immune pressure
CKs
T-cell

Figure 6. Schematic overview of the organization of the putative pullulanase type I encoding gene in *Lactobacillus crispatus*. The enzyme comprises three conserved domains including an N-terminal carbohydrate-binding module family 41 with specific carbohydrate binding sites, a catalytic module belonging to the pullulanase super family and a C-terminal bacterial surface layer protein (SLAP). The mutations (indicated by arrows) were located in an unconserved area that encodes a putative signal peptide (SP) that may be involved in subcellular localization. Abbreviations: SP: signal peptide; CBM41: carbohydrate-binding module family 41; PulA: pullulanase; SLAP: surface layer protein.



Lactobacillus crispatus clinical isolates coding sequences alignment


Glycosyl-transferase fragment 1
Type-1 Pullulanase
Glycosyl-transferase fragment 2
Glycosyl-transferase fragment 3
● Wild type
○ Mutation
Lactobacillus crispatus RL03 (HVM)
Lactobacillus crispatus RL05 (HVM)
Lactobacillus crispatus RL06 (HVM)
Lactobacillus crispatus RL08 (HVM)
Lactobacillus crispatus RL09 (HVM)
Lactobacillus crispatus RL10 (HVM)
Lactobacillus crispatus RL11 (HVM)
Lactobacillus crispatus RL16 (HVM)
Lactobacillus crispatus RL26 (HVM)
Lactobacillus crispatus RL27 (HVM)
Lactobacillus crispatus RL29 (HVM)
Lactobacillus crispatus RL32 (HVM)
Lactobacillus crispatus RL02 (DVM)
Lactobacillus crispatus RL07 (DVM)
Lactobacillus crispatus RL13 (DVM)
Lactobacillus crispatus RL14 (DVM)
Lactobacillus crispatus RL15 (DVM)
Lactobacillus crispatus RL17 (DVM)
Lactobacillus crispatus RL19 (DVM)
Lactobacillus crispatus RL20 (DVM)
Lactobacillus crispatus RL21 (DVM)
Lactobacillus crispatus RL23 (DVM)
Lactobacillus crispatus RL24 (DVM)
Lactobacillus crispatus RL28 (DVM)
Lactobacillus crispatus RL25 (DVM)
Lactobacillus crispatus RL30 (DVM)
Lactobacillus crispatus RL31 (DVM)
Lactobacillus crispatus RL33 (DVM)
RNA processing and modification
Chromatin structure and dynamics
Translation, ribosomal structure and biogenesis
Transcription
Replication, recombination and repair
Cell cycle control, cell division, chromosome partitioning
Post-translational modification, protein turnover, and chaperones
Cell wall/membrane/envelope biogenesis
Cell motility
Inorganic ion transport and metabolism
Signal transduction mechanisms
Intracellular trafficking, secretion, and vesicular transport
Defense mechanisms
Extracellular structures
Nuclear structure
Cytoskeleton
Energy production and conversion
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Nucleotide transport and metabolism
Coenzyme transport and metabolism
Lipid transport and metabolism
Secondary metabolites biosynthesis, transport, and catabolism
General function prediction only
Function unknown
Unknown COG
CDS
tRNA
rRNA
Other
BLAST hit >= 90 % identical (dark coloured)
BLAST hit >= 70 % identical (light coloured)
GC content
GC skew+
GC skew-