Characterization of the first vaginal *Lactobacillus crispatus* genomes isolated in Brazil

Marcelle Oliveira de Almeida[1], Rodrigo Carvalho[1], Flavia Figueira Aburjaile[1], Fabio Malcher Miranda[1], Janaína Canário Cerqueira[1], Bertram Brenig[2], Preetam Ghosh[3], Rommel Ramos[4], Rodrigo Bentes Kato[5], Siomar de Castro Soares[6], Artur Silva[4], Vasco Azevedo[1] and Marcus Vinicius Canário Viana[1]

[1] Department of Genetics, Ecology, and Evolution, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[2] Institute of Veterinary Medicine, University of Göttingen, Göttingen, Germany
[3] Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA
[4] Department of Genetics, Federal University of Pará, Belém, Pará, Brazil
[5] Post-graduation Program in Bioinformatics, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[6] Department of Immunology, Microbiology, and Parasitology, Federal University of Triângulo Mineiro, Uberaba, Minas Gerais, Brazil

ABSTRACT

Background: *Lactobacillus crispatus* is the dominant species in the vaginal microbiota associated with health and considered a homeostasis biomarker. Interestingly, some strains are even used as probiotics. However, the genetic mechanisms of *L. crispatus* involved in the control of the vaginal microbiome and protection against bacterial vaginosis (BV) are not entirely known. To further investigate these mechanisms, we sequenced and characterized the first four *L. crispatus* genomes from vaginal samples from Brazilian women and used genome-wide association study (GWAS) and comparative analyses to identify genetic mechanisms involved in healthy or BV conditions and selective pressures acting in the vaginal microbiome.

Methods: The four genomes were sequenced, assembled using ten different strategies and automatically annotated. The functional characterization was performed by bioinformatics tools comparing with known probiotic strains. Moreover, it was selected one representative strain (*L. crispatus* CRI4) for in vitro detection of phages by electron microscopy. Evolutionary analysis, including phylogeny, GWAS and positive selection were performed using 46 public genomes strains representing health and BV conditions.

Results: Genes involved in probiotic effects such as lactic acid production, hydrogen peroxide, bacteriocins, and adhesin were identified. Three hemolysins and putrescine production were predicted, although these features are also present in other probiotic strains. The four genomes presented no plasmids, but 14 known families insertion sequences and several prophages were detected. However, none of the mobile genetic elements contained antimicrobial resistance genes. The genomes harbor a CRISPR-Cas subtype II-A system that is probably inactivated due to fragmentation of the genes *csn2* and *cas9*. No genomic feature was associated with a

health condition, perhaps due to its multifactorial characteristic. Five genes were identified as under positive selection, but the selective pressure remains to be discovered. In conclusion, the Brazilian strains investigated in this study present potential protective properties, although in vitro and in vivo studies are required to confirm their efficacy and safety to be considered for human use.

INTRODUCTION

The vaginal microbiota of reproductive-age women is classified into at least five types called community state types (CST). Four of them are dominated by *Lactobacillus crispatus* (CST-I), *L. gasseri* (CST-II), *L. iners* (CST-III), or *L. jensenii* (CST-V). The CST-IV is characterized as having a significantly lower number of lactobacilli and an increased number and diversity of strict and facultative anaerobes (*Ravel et al., 2011*). The lactobacilli use the glycogen supplied by the host as a carbon source and create a protective environment against infections or colonization by pathogens and non-indigenous microbes by the production of L- and/or D-lactic acid, bacteriocins, hydrogen peroxide, competition for tissue adhesion, enhancement of the protective mucus layer integrity and modulation of the innate immune system response (*Ravel et al., 2011*; *Smith & Ravel, 2017*).

CST-IV and III increase the risk of bacterial vaginosis (BV) due to *L. iners* being less effective in controlling the vaginal microbiota. BV is a condition characterized by a microbiota similar to CST-IV, vaginal pH > 4.5 and production of amino acid compounds, sometimes associated with clinical symptoms including discharge, fishy odor, and presence of clue cells. These conditions damage the host defenses and favor the development of opportunistic microorganisms that behave like pathogens (*Smith & Ravel, 2017*; *Barrientos-Durán et al., 2020*). BV can be asymptomatic or associated with gynecological and obstetric complications, besides increasing the risk of sexually transmitted infections (*Barrientos-Durán et al., 2020*).

CST-I, dominated by *L. crispatus*, is most associated with vaginal health. This species is considered a biomarker of a healthy microbiota, and some strains are used as a probiotic to treat BV. Its effect as a probiotic is not entirely clear, but it is believed to involve competitive exclusion strategies (*Almeida et al., 2019*). It was shown to outcompete *Gardinerella vaginallis* for tissue adhesion in vitro (*Ojala et al., 2014*) and, as other lactobacilli, inhibit the growth of pathogens in vivo by the production of lactic acid, but not hydrogen peroxide (*Tachedjian, O'Hanlon & Ravel, 2018*). The decrease in *L. crispatus* is associated with BV, but the causes are not well understood. The differences in persistence and protection among strains is probably influenced by genetic differences that, if described, could be applied in the screening of more efficient probiotic strains (*Almeida et al., 2019*).

Comparative genomic analyses have been performed to identify genomic features of *L. crispatus* associated with female urogenital tract using samples from North America, Europe and Asia (*Ojala et al., 2014*; *Abdelmaksoud et al., 2016*; *France, Mendes-Soares & Forney, 2016*; *Van der Veer et al., 2019*; *Pan, Hidalgo-Cantabrana & Barrangou, 2020*; *Petit & Read, 2020*; *Zhang et al., 2020*). Analyses that searched for genomic features associated with lactobacilli-dominated (healthy) or BV microbiota were performed using samples isolated from the USA (*Abdelmaksoud et al., 2016*) and Netherlands (*Van der Veer et al., 2019*). Those studies point out mechanisms related to persistence during BV, such as phase variation, rather than protection against this infirmity (*Van der Veer et al., 2019*). In Latin America, although *L. crispatus* has been previously studied in Brazil as vaginal isolates (*Branco et al., 2010*) or as part of the vaginal microbiota (*Martinez et al., 2008*; *Marconi et al., 2020*), no complete genome has been sequenced, analyzed, and deposited in GenBank.

In this study, we characterize the first *L. crispatus* genomes isolated in Brazil from healthy vaginal microbiomes and used genome-wide association study (GWAS) and positive selection analyses to identify genetic mechanisms involved healthy or BV conditions and selective pressures acting in the vaginal microbiome.

MATERIALS AND METHODS

Genome sequencing, assembly, and annotation

In a previous study, vaginal fluid samples were collected from individuals diagnosed as healthy or with BV, with the approval of the ethics committee in research (COEP) of the Federal University of Minas Gerais (protocols ETIC 062/03 and 212/07) (*Branco et al., 2010*). *L. crispatus* strains were identified by cellular morphology (Gram-positive bacilli or coccobacilli), biochemistry test (catalase-negative) and 16S-23S rRNA restriction profiling (*Branco et al., 2010*). The genomes of strains CRI4, CRI8, CRI10 and CRI17 isolated from four healthy patients were sequenced using HiSeq 2500 (Illumina, San Diego, CA, USA) with paired-end libraries of 2×150 bp. The sequencing reads quality was examined using FastQC v0.11.8 (*Andrews, 2015*). The sequencing reads were mapped to the genomes of 46 *L. crispatus* vaginal isolates to filter out contaminants (Table 1) using bowtie v2 (*Langmead & Salzberg, 2012*), and the mapped reads were extracted using Samtools v1.7-2 (*Li et al., 2009*).

Ten different assemblies were generated for each genome using SPAdes v3.14.0 (*Bankevich et al., 2012*) (assembly 1), Unicycler v0.4.5 (*Wick et al., 2017*) with SPAdes v3.14.0 (2), ABySS v2.0 (*Simpson et al., 2009*) (3), MaSuRCA v3.4.0 (*Zimin et al., 2013*) (4) and an *in house* pipeline (https://github.com/engbiopct/assembly-hiseq) that generates six other assemblies (5–10). In the *in-house* pipeline, six strategies combine different software. The best k-mer values were identified using KmerStream v1.1 (*Melsted & Halldórsson, 2014*). Adapters were removed using AdapterRemoval v2.3.1 (*Schubert, Lindgreen & Orlando, 2016*). The genome assemblers were Edena v3.131028 (*Hernandez et al., 2008*) and SPAdes v3.13.0 (*Bankevich et al., 2012*). The six assembly strategies were: Edena (5), KmerStream and SPAdes (6), KmerStream and SPAdes, using Edena assembly as trusted contigs (7), AdapterRemoval, KmerStream and SPAdes (8), reads

Table 1 List of *Lactobacillus crispatus* genomes.

Strain	Condition	Microbiome	Metadata	Country	GenBank	Reference
RL02	BV	DVM	BV-positive	Netherlands	NKLR01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL07	BV	DVM	BV-positive	Netherlands	NKLN01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL13	BV	DVM	BV-positive	Netherlands	NKLI01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL14	BV	DVM	BV-positive	Netherlands	NKLH01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL15	BV	DVM	BV-positive	Netherlands	NKLG01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL17	BV	DVM	BV-positive	Netherlands	NKLE01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL19	BV	DVM	BV-positive	Netherlands	NKLD01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL20	BV	DVM	BV-positive	Netherlands	NKLC01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL21	BV	DVM	BV-positive	Netherlands	NKLB01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL23	BV	DVM	BV-positive	Netherlands	NKLA01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL24	BV	DVM	BV-positive	Netherlands	NKKZ01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL25	BV	DVM	BV-positive	Netherlands	NKKY01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL28	BV	DVM	BV-positive	Netherlands	NKKV01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL30	BV	DVM	BV-positive	Netherlands	NKKT01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL31	BV	DVM	BV-positive	Netherlands	NKKS01	*Dols et al. (2016), Van der Veer et al. (2019)*
RL33	BV	DVM	BV-positive	Netherlands	NKKQ01	*Dols et al. (2016), Van der Veer et al. (2019)*
VMC1	BV	DVM	History of BV. <50% lactobacilli and >50% of BV-associated taxa	USA	LJCZ01	*Abdelmaksoud et al. (2016)*
VMC2	BV	DVM	History of BV. <50% lactobacilli and >50% of BV-associated taxa	USA	LJDA01	*Abdelmaksoud et al. (2016)*
VMC3	BV	DVM	History of BV. <50% lactobacilli and >50% of BV-associated taxa	USA	LJGP01	*Abdelmaksoud et al. (2016)*
VMC4	BV	DVM	History of BV. ~86% *L. crispatus* and ~12% BV-associated taxa	USA	LJGQ01	*Abdelmaksoud et al. (2016)*
CRI4	Healthy	–	Healthy	Brazil	JABERN01	This study
CRI8	Healthy	–	Healthy	Brazil	JABERO01	This study
CRI10	Healthy	–	Healthy	Brazil	JABERP01	This study

Table 1 (continued)

Strain	Condition	Microbiome	Metadata	Country	GenBank	Reference
CRI17	Healthy	–	Healthy	Brazil	JABERQ01	This study
2029	Healthy	–	Healthy, probiotic strain	Russia	AVFH2	*Abramov et al. (2014)*
125-2-CHN	Healthy	–	Healthy	China	ACPV01	*Ojala et al. (2014)*, www.beiresources.org
AB70	Healthy	–	Healthy	South Korea	CP026503, CP026504	*Chang et al. (2019)*
CIP 104459	Healthy	–	Healthy	France	VOMA01	*Clabaut et al. (2020)*
CTV-05	Healthy	–	Healthy, probiotic strain	–	ADML01	*Hemmerling et al. (2010)*, *Ojala et al. (2014)*
JV-V01	Healthy	–	Normal human vaginal flora	–	ACKR01	*Witkin et al. (2013)*, *Ojala et al. (2014)*, www.beiresources.org
MV-1A-US	Healthy	–	Healthy	USA	ACOG02	*Witkin et al. (2013)*, *Ojala et al. (2014)*, www.beiresources.org
MV-3A-US	Healthy	–	Healthy	USA	ACQC01	*Witkin et al. (2013)*, *Ojala et al. (2014)*, www.beiresources.org
RL03	Healthy	LVM	BV-negative	Netherlands	NKLQ01	*Dols et al. (2016)*, *Van der Veer et al. (2019)*
RL05	Healthy	LVM	BV-negative	Netherlands	NKLP01	*Dols et al. (2016)*, *Van der Veer et al. (2019)*
RL06	Healthy	LVM	BV-negative	Netherlands	NKLO01	*Dols et al. (2016)*, *Van der Veer et al. (2019)*
RL08	Healthy	LVM	BV-negative	Netherlands	NKLM01	*Dols et al. (2016)*, *Van der Veer et al. (2019)*
RL09	Healthy	LVM	BV-negative	Netherlands	NKLL01	*Dols et al. (2016)*, *Van der Veer et al. (2019)*
RL10	Healthy	LVM	BV-negative	Netherlands	NKLK01	*Dols et al. (2016)*, *Van der Veer et al. (2019)*
RL11	Healthy	LVM	BV-negative	Netherlands	NKLJ01	*Dols et al. (2016)*, *Van der Veer et al. (2019)*
RL16	Healthy	LVM	BV-negative	Netherlands	NKLF01	*Dols et al. (2016)*, *Van der Veer et al. (2019)*
RL26	Healthy	LVM	BV-negative	Netherlands	NKKX01	*Dols et al. (2016)*, *Van der Veer et al. (2019)*
RL27	Healthy	LVM	BV-negative	Netherlands	NKKW01	*Dols et al. (2016)*, *Van der Veer et al. (2019)*
RL29	Healthy	LVM	BV-negative	Netherlands	NKKU01	*Dols et al. (2016)*, *Van der Veer et al. (2019)*
RL32	Healthy	LVM	BV-negative	Netherlands	NKKR01	*Dols et al. (2016)*, *Van der Veer et al. (2019)*
SJ-3C-US	Healthy	–	Healthy	Iran	ADDT01	*Eslami et al. (2016)*
V4	Healthy	–	Healthy	France	SRLG01	*Clabaut et al. (2019)*

(Continued)

Table 1 (continued)

Strain	Condition	Microbiome	Metadata	Country	GenBank	Reference
VMC5	Healthy	LVM	No history of BV. >90 % *L. crispatus* and <10 % BV-associated bacterial taxa	USA	LJOK01	*Abdelmaksoud et al. (2016)*
VMC6	Healthy	LVM	No history of BV. Dominated by *L. crispatus* and *L. jensenii*	USA	LJOL01	*Abdelmaksoud et al. (2016)*
VMC7	Healthy	LVM	No history of BV. >90 % *L. crispatus* and <10 % BV-associated bacterial taxa	USA	LJOM01	*Abdelmaksoud et al. (2016)*
VMC8	Healthy	LVM	No history of BV. >90 % *L. crispatus* and <10 % BV-associated bacterial taxa	USA	LJON01	*Abdelmaksoud et al. (2016)*

Note:

LVM, Lactobacilli-dominated vaginal microbiota; DVM, Dysbiotic vaginal microbiota.

processed by AdapterRemoval and the raw reads as input, KmerStream, and SPAdes, using Edena assembly as trusted contigs (9), and AdapterRemoval, KmerStream, and SPAdes, using Edena assembly as trusted contigs (10).

The best assembly for each genome was determined using QUAST v5.0.2 (*Gurevich et al., 2013*). Then, the genome's paired-read sequencing data was used for scaffolding with SSPACE v3.0 (*Boetzer et al., 2011*) and contig extension and gap filling with GapFiller v1.1.1 (*Boetzer & Pirovano, 2012*). Finally, more gaps were closed using contigs from the other nine assemblies and the chromosome of *L. crispatus* strain AB70 (CP026503.1) (*Chang et al., 2019*) as a reference, using GFinisher (*Guizelini et al., 2016*). The presence of plasmids was investigated using PlasmidFinder 2.1 (*Carattoli et al., 2014*). The four genomes were identified from the species *L. crispatus* using the Type (Strain) Genome Server (*Meier-Kolthoff & Göker, 2019*). The assemblies completeness was evaluated using BUSCO v4.0.6 (*Seppey, Manni & Zdobnov, 2019*), based on the presence of 402 single-copy orthologous genes shared within Lactobacillales. The genomes were annotated using Prokka v1.11 (*Seemann, 2014*). The GenBank accession numbers of the genomes from strains CRI4, CRI8, CRI10 and CRI17, isolated from healthy patients, are JABERN01, JABERO01, JABERP01 and JABERQ01, respectively (Table 2). Type Strain Genome Server (*Meier-Kolthoff & Göker, 2019*) was used to confirm the taxonomic classification of the 50 samples as *L. crispatus* strains.

Criteria of public genomes selection

We included in the analysis genomes available in public databases using the following criteria: (i) vaginal isolate, (ii) the metadata explicitly informs the health condition of the individual, and/or the microbiome classification. A total of 46 samples were classified as belonging to "healthy" (26) or "BV" (20) condition groups and were used in previous

Table 2 Statistics of genome sequencing, assembly, and annotation of the *Lactobacillus crispatus* strains CRI4, CRI8, CRI10, and CRI17.

	CRI4	CRI8	CRI10	CRI17
SRA accession	SRR13201099	SRR13201098	SRR13201097	SRR13201096
Replicon accession	JABERN01	JABERO01	JABERP01	JABERQ01
Completeness (%)	99.5	99.2	99.0	99.0
Size (bp)	2,376,268	2,330,310	2,418,420	2,384,332
Contig	100	65	65	79
N50 (bp)	44.691	69.313	75.243	58.032
L50	19	10	11	14
CDS	2.438	2.329	2.478	2.393
Plasmid	None	None	None	None

studies (*Hemmerling et al., 2010*; *Ojala et al., 2014*; *Abramov et al., 2014*; *Abdelmaksoud et al., 2016*; *Eslami et al., 2016*; *Dols et al., 2016*; *Chang et al., 2019*; *Clabaut et al., 2019*, *2020*; *Van der Veer et al., 2019*). Table 1 shows the genomes list, including the microbiome classification, when available, and the terms in metadata used to classify the sample.

Probiotic features

Bacteriocins and linear azol(in)e-containing peptides (LAPs) were predicted using BAGEL4 (*Van Heel et al., 2018*). The enzymes involved in the production of L- and D-lactate and hydrogen peroxide were identified using KEGG Mapper/BlastKOALA, under the pyruvate metabolism pathway. Adhesins were predicted using eggNOG-mapper v2 (*Huerta-Cepas et al., 2017*). Protein IDs were identified using BLASTp (*Camacho et al., 2009*) using the GenBank non-redundant (nr) database, selecting hits with 100% identity and coverage. Pathways involved in the biosynthesis of antimicrobial drugs with clinical importance (*Chokesajjawatee et al., 2020*) were predicted using KEGG Mapper.

Safety assessment of the strains for probiotic applications

Detection of plasmids, insertion sequences, prophages, and CRISPR-Cas elements was performed using PlasmidFinder (*Carattoli et al., 2014*), ISEScan (*Xie & Tang, 2017*), PHASTER (*Arndt et al., 2016*), and CRISPRCasFinder (https://crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index), respectively. The domains of the identified Cas proteins were predicted using InterProScan (*Jones et al., 2014*) and NCBI's Conserved Domain Database (*Lu et al., 2020*). Multiple alignments of sequences of interest were performed using the Clustal Omega web service (*Larkin et al., 2007*). Local alignment of sequences of interest across genomes was performed using BLASTn (*Camacho et al., 2009*) implemented in PATRIC (*Davis et al., 2020*).

Virulence factor genes were detected using the databases VFDB (*Liu et al., 2019*) and Ecoli_VF (https://github.com/phac-nml/ecoli_vf), while antimicrobial resistance genes were detected using the databases ARG-ANNOT (*Gupta et al., 2014*), CARD (*Alcock et al., 2020*), MEGARes (*Doster et al., 2019*), NCBI AMRFinderPlus (*Feldgarden et al., 2019*) and

ResFinder (*Bortolaia et al., 2020*). The screening using these databases was performed using ABRicate (https://github.com/tseemann/abricate) with default parameters.

The presence of the toxins hemolysins, enzymes involved in the synthesis of biogenic amines and other undesirable genes, as listed by *Chokesajjawatee et al. (2020)*, were manually screened using KEGG Mapper (*Kanehisa & Sato, 2020*) implemented in BlastKOALA v2.2 (*Kanehisa, Sato & Morishima, 2016*).

In vitro detection of phages

One representative strain (*L. crispatus* CRI4) was subjected to the induction of lysogenic bacteriophages according to previous studies (*Kiliç et al., 1996*; *Raya & H'bert, 2009*). A total of 200 µL of the culture was initially inoculated in 5 mL of Man–Rogosa–Sharpe medium. The optical density (OD) was measured until reaching $DO_{600} = 0.2$. Then, 0.4 µg/mL of Mitomycin C (Sigma, St. Louis, MO, USA) was added in the culture and incubated at 37 °C overnight. Afterward, the supernatant was collected and filtered in sterile 0.22 µm membrane. A 10 µL sample of the filtered lysate was applied to a 200 mesh grid at the UFMG electron microscopy center (CM-UFMG) and visualization was performed in Tecnai G2-12 Transmission Electron Microscope, SpiritBiotwin FEI, 120 kV.

Comparative analyses using public genomes

We performed comparative analysis including the for strains from Brazil and 46 public genomes (Table 1) to reconstruct their phylogeny and to identify adaptations that could be related to the colonization of the vaginal niche by testing for association between gene presence/absence and features of interest, and detection of positive selection in protein-coding genes.

For phylogenomic analysis, *L. helveticus* DSM20075 genome (GenBank accession ACLM01) was used as an outgroup, the conserved genes across all 51 genomes were estimated by Roary v3.6.0, and their nucleotide sequences were aligned using MAFFT (*Katoh et al., 2005*) implemented in Roary. The alignment was used as input for IQ-Tree v1.6.12 (*Nguyen et al., 2015*) for phylogenetic inference using the Maximum Likelihood. The confidence values were estimated using 1,000 rounds of bootstrapping. The tree was edited using iTOL (*Letunic & Bork, 2019*).

We tested the associations of gene presence/absence with a health condition and geographical locations suggested by the phylogenetic tree. A GWAS based on gene presence/absence was performed using Scoary v1.6.16 (*Brynildsrud et al., 2016*). Scoary estimates association by pairwise comparisons on a phylogeny (*Maddison, 2000*) to correct population structure and permutation. The input was a gene presence-absence matrix from the 50 *L. crispatus* genomes estimated using Roary and a phylogenetic tree generated by IQ-Tree v1.6.12, utilizing the core gene alignment calculated by Roary, and a matrix containing the presence-absence of the features across the samples.

A genome-scale positive selection analysis was performed using POTION v1.1.2 (*Hongo et al., 2015*). To generate the input, FastOrtho (https://github.com/PATRIC3/FastOrtho) was used to identify ortholog groups across the 50 *L. crispatus* genomes. The file containing the orthologous group's information and multifasta files containing nucleotide sequences of

protein-coding genes were used as input for POTION v1.1.2. The genome-scale positive selection analysis used *site* tests with the models M1 and M2, and M7 and M8 (*Yang & Nielsen, 2002*). The POTION configuration file is available as Data S1. The function of the identified proteins was annotated using eggNOG-mapper (*Huerta-Cepas et al., 2017*), the subcellular localization using SufG+ v1.2.1 (*Barinov et al., 2009*), and the GenBank protein ID using BLASTp (*Agarwala et al., 2016*).

RESULTS

Genome sequencing and taxonomy

The four sequenced genomes were assembled as drafts, and no plasmid was found. Table 2 shows the statistics of genome assembly and annotation. The genomes completeness ranged from 99% to 99.5%, while the reference strain AB70, available as a complete genome sequenced using PacBio RS II platform (*Chang et al., 2019*), had its completeness estimated as 99%. The four sequenced genomes and the 46 public genomes were classified as *L. crispatus* by TYGS, with dDDH > 70% and G+C content divergence of less than 1% to the strain JCM 1185^T (Data S2).

Probiotic features

Features associated with probiotic effects in the four strains from Brazil are shown in Table 3. We identified genes involved in the biosynthesis of D-lactate (1), L-lactate (3), hydrogen peroxide (1), bacteriocins (9), LAPs (1) and adhesins (10, five classes) across the four genomes. No pathway for biosynthesis of antimicrobial drugs of clinical importance was found.

Safety assessment of the strains for probiotic applications

Features associated with safety in the four strains from Brazil are shown in Table 4. About mobile elements, no plasmid was predicted, as previously stated. A total of 131 to 184 IS from 14 known families were identified in the four genomes, and one new family was predicted across CRI8 (2 copies), CRI10 (2) and CRI17 (1) (Data S3). The new IS has a size of 2,088 bp and harbors two genes coding a site-specific integrase-resolvase (IS607-like family, GenBank protein ID WP_005728427.1) and a transposase (IS605 family, AZR15009.1) (Data S4). The multiple alignments of the five copies show the IS is fragmented in the contig 44 of strain CRI10, and the difference among the complete sequences is a SNP (T to G) in position 1,350 (Data S5). A BLASTn against the four strains from Brazil and other 123 public *L crispatus* genomes database identified hits with ≥99% identity and ≥98% coverage with 42 genomes from strains isolated from the female human urogenital tract, 39 of them public genomes (Data S3).

Five or six prophages were predicted across the four genomes, classified as questionable (score 70–90) or incomplete (score < 70), with most of them close to contig ends. In strains CRI8 and CRI17, some of the predicted prophages harbored the new IS (Fig. 1; Data S6). We identified CRISPR-Cas systems subtype II-A in the four strains and subtype II-C only in strain CRI10 (Table 4; Data S7). A closer examination of the subtype II-A system reveals that all four genomes contain CRISPR loci and Cas proteins-encoding genes *cns2*, *cas2*, *cas1*,

Table 3 Genomic features related to probiotic effects in *Lactobacillus crispatus* strains CRI4, CRI8, CRI10 and CR17.

Feature	Product (Gene)	KEGG ID	Protein ID	CRI4	CRI8	CRI10	CRI17
Lactic acid synthesis							
	D-lactate			1	1	1	1
	D-lactate dehydrogenase [EC:1.1.1.28] (*ldhA*)	K03778	WP_005720611	1	1	1	1
	L-lactate			3	3	3	3
	L-lactate dehydrogenase [EC:1.1.1.27] (*ldh*)	K00016	WP_005721100	1	1	1	1
		K00016	WP_005720302	1	WP_170080485	1	WP_005721074
	L-2-hydroxyisocaproate dehydrogenase (*hicDH*)	K00016	WP_005727148	WP_005719855	1	1	1
Bacteriocin							
	Class: 210.2; SakT_alpha			2	2	3	2
	ggmotif; ComC; Bacteriocin_IIc;	–	WP_005721006	1	1	1	1
	ComC; L_biotic_typeA; Bacteriocin_IIc; 20.2; bacteriocin_LS2chaina		WP_005721005	1	1	1	1
	ComC; L_biotic_typeA; Bacteriocin_IIc;	–	WP_005720990	–	–	1	–
	Class: 70.3; Helveticin-J			1	2	2	2
	70.3; Helveticin-J	–	WP_005729773	–	1	1	1
		–	WP_005720754	1	1	1	1
	Class: 64.3; Enterolysin_A			1	2	2	2
	64.3; Enterolysin_A	–	WP_005728076	–	1	1	1
		–	WP_005719715	1	1	1	1
	Class: 163.2; Penocin_A			–	–	2	1
	bacteriocinII; Bacteriocin_II; ComC; Bacteriocin_IIc; 163.2; Penocin_A	–	WP_005723822	–	–	1	1
	Bacteriocin_IIc;	–	WP_005727428	–	–	1	–
	Class: 6.3; Bacteriocin_helveticin_J			1	1	1	1
	6.3; Bacteriocin_helveticin_J	–	WP_005728268	WP_005718134	1	1	1
	Class: LAPs			-	-	1	-
	Putative nitr oreductase MJ1384	–	WP_005721909	–	–	1	–
Hydrogen peroxide synthesis							
	Pyruvate oxidase [EC:1.2.3.3] (*poxL*)	K00158	WP_005723618	1	1	1	1
Adhesin							
	Putative adhesin		WP_005728236	1	1	1	1
			WP_005729490	–	1	1	1
Antimicrobial production	–	–	–	–	–	–	–

**Table 4 Genomic features related to safety in *Lactobacillus crispatus* strains CRI4, CRI8, CRI10, and CR17.**

Feature	Product (Gene)	KEGG ID	BLASTp hit	Strain			
				CRI4	CRI8	CRI10	CRI17
Mobile elements							
Plasmid				–	–	–	–
Insertion sequences				131	171	178	184
Prophages				2 questionable, 4 incomplete	5 incomplete	1 questionable, 5 incomplete	1 questionable, 4 incomplete
CRISPR-Cas system				CAS-TypeIIA	CAS-TypeIIA	CAS-TypeIIA, CAS-TypeIIC	CAS-TypeIIA
Bacterial toxins							
	Hemolysin A (*tlyA*)	K06442	WP_005723149	–	1	1	1
	Putative hemolysin (*tlyC*)	K03699	WP_005727867	1	1	1	1
	Hemolysin-III related (*hlyIII*)	K11068	WP_005720215	1	1	1	1
Bile salt deconjugation							
	Choloylglycine hydrolase/ bile salt hydrolase (*cbh*)	K01442	WP_005718943	1	1	1	1
Biogenic amine formation							
	Ornithine decarboxylase (*odcI*)	K01581	WP_005727730	1	1	1	1
Antimicrobial resistance	–			–	–	–	–

and *cas9*. However, a transposase is inserted in *csn2* in each strain, resulting in two gene fragments. The gene *cas9* is also fragmented in all strains, with a transposase inserted between fragments in strains CRI4 and CRI8. CRI17 has an extra copy of *cas9*. The subtype II-C system is represented in CRI10 by a single *cas9* gene (Fig. 2). All the *cas9* CDSs lack one or more domains.

No virulence or antimicrobial resistance gene was predicted using ABRicate. With the search for virulence and undesirable genes using KEGG Mapper/BlastKOALA, we identified two hemolysins (*tlyC* and *hlyIII*), one bile salt hydrolase (*cbh*), and one enzyme involved in the biosynthesis of putrescine (Ornithine decarboxylase, *odcI*). Manual screening of Prokka annotation showed a third hemolysin, Hemolysin A. The search for toxin biosynthesis identified an "Ornithine decarboxylase" (*odcI*) (WP_005727730.1) involved in the synthesis of the biogenic amine putrescine, in all for strains.

In vitro detection of phages

Several rod-shaped particles of varying length and thickness were observed by electron microscopy, some of them similar to Myoviridae bacteriophages. This pattern was repeated throughout the slide (Fig. 3).
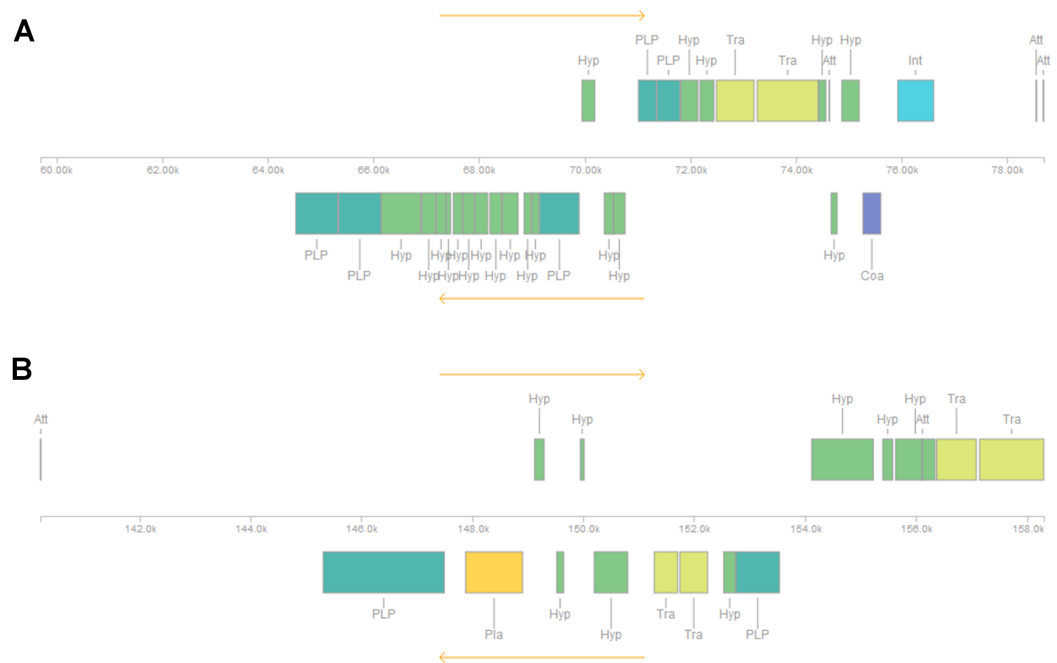
Figure 1 Insertion sequence from IS607-like family located in two prophages (regions) in *Lactobacillus crispatus* strain CRI8. (A) Region 2 and (B) Region 3. The insertion sequences from the IS607-like family are two subsequent transposases, located in region 2 and at the end of region 3. Att, Attachment Site; Coa, Coat protein (purple); Hyp, Hypothetical protein (green); Int, Integrase (blue); Pla, Plate protein (orange); PLP, Phage-like Protein (cyan); Tra, Transposase (olive).

Full-size 🖼 DOI: 10.7717/peerj.11079/fig-1

Comparative analyses using public genomes

The strains did not cluster according to a health condition. However, strains from Brazil were the closest to strains from the Netherlands, even in different clusters (Fig. 4). The GWAS based on gene presence/absence implemented in Scoary did not identify an association with a health condition ($p > 0.05$). Due to the clustering of strains from Brazil and the Netherlands, we tested the association with these two geographical locations and the result was also negative. A total of 8 protein-coding genes were identified as under positive selection ($q < 0.05$). After manual curation for false positives caused by alignment artifacts, five genes were obtained: a surface exposed "S-layer protein precursor," three phage related proteins, and a hypothetical protein (Table 5; Data S8).

DISCUSSION

In the first four *L. crispatus* sequenced genomes from Brazil, we identified genes involved in the protective properties of vaginal lactobacilli (*Smith & Ravel, 2017*) such as the production of lactic acid, bacteriocins, hydrogen peroxide and adhesion (Table 3).

For safety reasons, a probiotic strain should not present features associated with virulence and antimicrobial resistance, as well as mobile elements that could transfer those features to other microorganisms from the host microbiome (*Pariza et al., 2015*; *Chokesajjawatee et al., 2020*). No antimicrobial resistance gene was found, and only three hemolysins were predicted as virulence factors, which have also been previously identified

Figure 2 CRISPR-Cas systems subtype II-A predicted in four *Lactobacillus crispatus* strains from Brazil. (A) Strain CRI4, (B) CRI8, (C) CRI10 and (D) CR17. The genes *csn2* and *cas9* are fragmented in the four genomes. Colors of the protein coding sequences green—*csn2*, *cas2* and *cas1*; red—*cas9* fragments; pink—transposase. Full-size ⊡ DOI: 10.7717/peerj.11079/fig-2

in other *Lactobacillus* probiotic strains (*Zafar & Saier, 2020*) (Table 4). The hemolysin-III is widespread across *Lactobacillus* species, including strains considered as safe and commercially available (*Chokesajjawatee et al., 2020*). This result implies these genes should not be a safety concern. However, the hemolytic activity of those proteins from *Lactobacillus* must be verified in further studies. We predicted the synthesis of the toxic biogenic amine putrescine by the ornithine decarboxylase pathway in all four strains (Table 4). This is one of the pathways in which the decarboxylation of amino acids and organic acids generates a proton motive force that can regulate intracellular pH (acidic stress response) and generate ATP (*Romano et al., 2014*; *Del Rio et al., 2018*). The production of biogenic amines is an important safety issue when screening probiotic strains as they may cause intoxication when consumed in high concentrations (*García-Villar, Hernández-Cassou & Saurina, 2009*; *Linares et al., 2011*).

Concerning mobile elements, as expected for probiotic strains, no plasmids were found. Despite of numerous ISs from 14 known families (Table 4; Data S3) and one from a new family-specific of vaginal isolates (Datas S3, S4, and S5), none of them were associated with antibiotic resistance genes.

Four to six incomplete prophages across the genomes were revealed (Table 4; Data S6). Some incomplete prophages could be the result of an assembly artifact, as the genomes were assembled as drafts and most prophages were located close to a contig end. The possible role of the newly described IS, and the associated prophages in the genome evolution (*Siguier, Gourbeyre & Chandler, 2014*) or adaptation to the urogenital niche
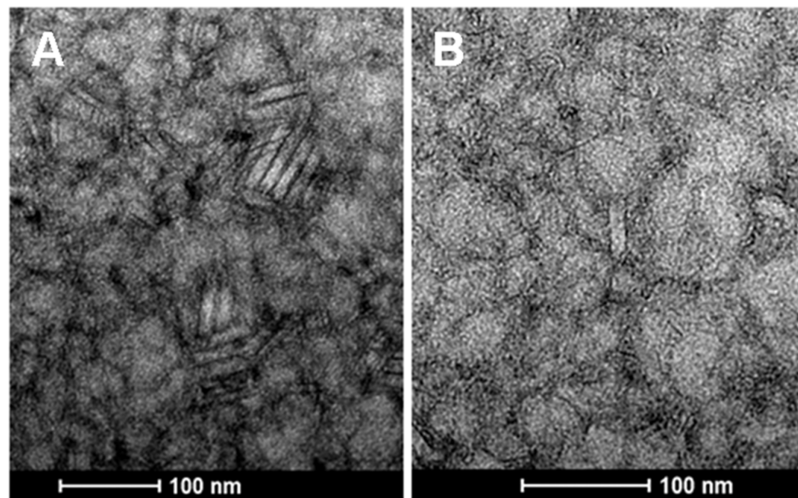
Figure 3 Transmission electron microscopy of the strain CRI4 filtrate. (A) Presence of stick-shaped structures. (B) Structure similar to a capsid with contractile tail (Myoviridae).

Full-size 🖼 DOI: 10.7717/peerj.11079/fig-3

should be investigated. Moreover, the presence of complete phages was confirmed by electron microscopy in CRI4 lysate after Mitomycin C induction (Fig. 3). Therefore, our results suggest the possibility of lysogeny of this strain.

The presence of a functional CRISPR-Cas system could prevent the infection by bacteriophages (*Crawley et al., 2018*) that could influence the vaginal microbiota (*Macklaim et al., 2013*). We detected the subtype II-A in the four genomes and II-C in CRI10 (Table 4; Data S7). However, all *csn2* and *cas9* genes across the genomes are fragmented (Fig. 2), lacking one or more domains. Also, the subtype II-C in CRI10 is probably a prediction artifact, as a single subtype with two sets of *cas* genes can be detected as multiple subtypes (*Crawley et al., 2018*) and only the subtypes I-B, I-E, and II-A were reported for this species (*Petit & Read, 2020*). The result suggests that the CRISPR-Cas systems are not functional due to gene fragmentation which could compromise the prevention of phage infections.

In the phylogenomic tree, health-related condition does not show clustering (Fig. 4), a result also found in other studies (*Abdelmaksoud et al., 2016*; *Pan, Hidalgo-Cantabrana & Barrangou, 2020*). However, Brazil-Netherlands clusters were formed. The GWAS performed using Scoary did not find gene presence/absence associated with the health condition or the Brazil-Netherlands clusters. This result is different from a previous study, which has shown an association of transposases and a glycosyltransferase being more abundant in BV strains (*Van der Veer et al., 2019*). The differences obtained between association studies could be due to the multifactorial characteristic of BV (*Barrientos-Durán et al., 2020*; *Marconi et al., 2020*).

The positive selection analysis identified five genes (Table 5; Data S8). The S-layer precursor has the surface layer A protein (SLAP) domain described in *slpB* from *L. acidophilus* ATCC 4356 (*Boot, Kolen & Pouwels, 1995*). S-layer proteins form symmetric, porous, lattice-like layers that cover the cell surface with poorly known
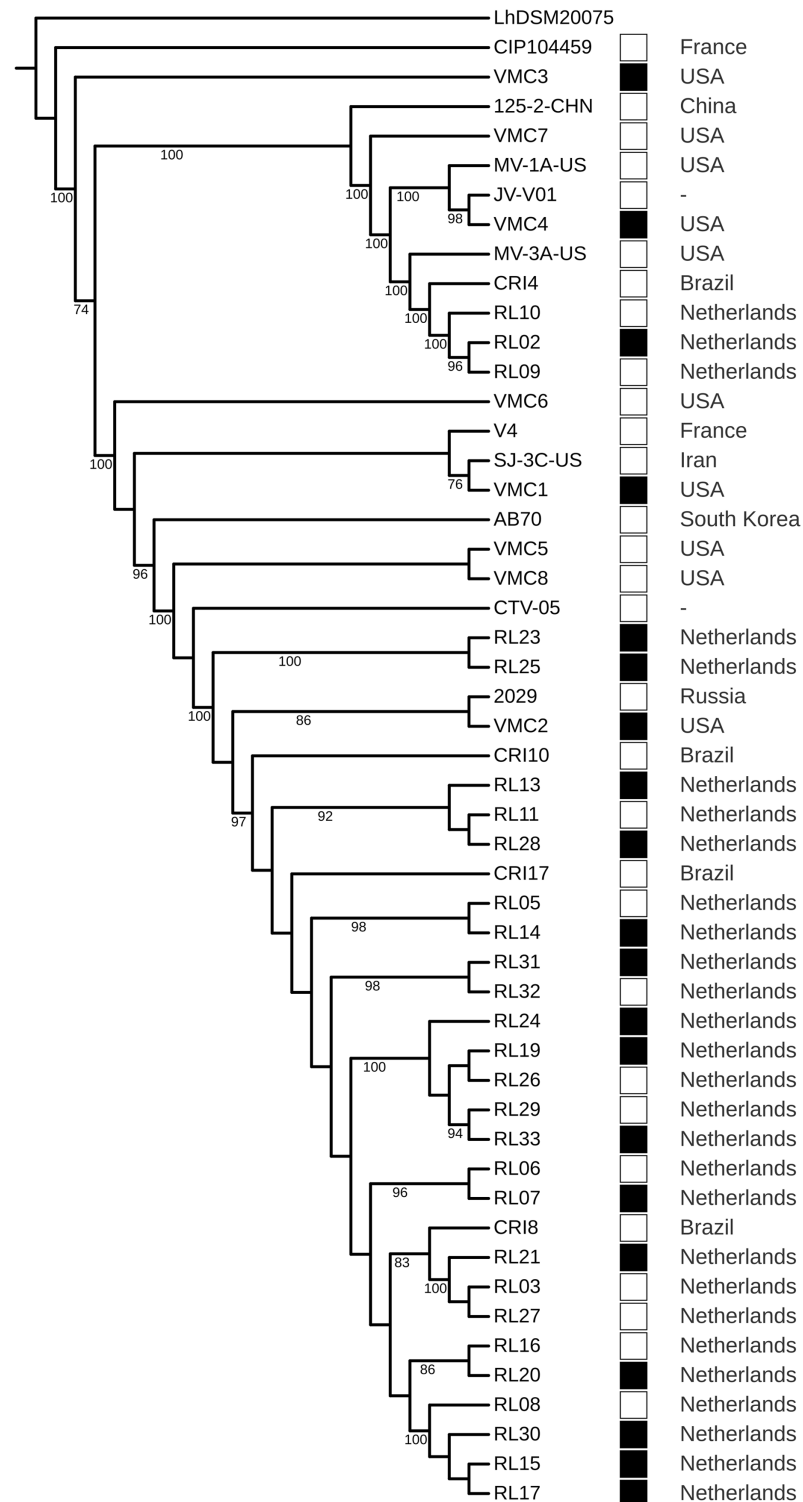
Figure 4 Phylogenomic tree of *Lactobacillus crispatus* strains from vaginal isolates. The tree was built using the nucleotide sequences of 118 core genes predicted by Roary and aligned by MAFFT, 1,000 rounds of bootstrapping, and Maximum Likelihood phylogenetic inference implemented in IQ-TREE. Filled squares—bacterial vaginosis samples, Empty squares—samples from healthy individuals.

Full-size DOI: 10.7717/peerj.11079/fig-4

Table 5 Genes under positive selection in *Lactobacillus crispatus* genomes from healthy and bacterial vaginosis samples.

Product (Gene)	COG	Location	Sequences/genomes	PS sites	Positions	Protein ID
Integrase core domain protein	L	CYTOPLASMIC	53/50	1	160	EKB63650
S-layer protein precursor	S	SECRETED	42/50	1	130	EKB61518
Hypothetical protein	–	CYTOPLASMIC	15/50	11	31, 39, 49, 48, 52, 58, 94, 95, 103, 107, 126	WP_126708926
ORF6N domain protein	K	CYTOPLASMIC	10/50	2	240, 256	WP_133463822
Tyrosine recombinase XerS (*xerS*)	L	CYTOPLASMIC	4/50	1	231	WP_133475995

Note:
PS, positively selected sites.

functions that can involve mediation of bacterial adherence to host cells, extracellular matrix proteins, or protective or enzymatic functions (*Hynönen & Palva, 2013*). Surface exposed proteins are located in the interface with the environment and can be under positive selection due to interaction with several factors such as antimicrobial compounds, viruses, hosts, and other bacteria (*Petersen et al., 2007*). The three phage related proteins have domains for DNA biding, integration, and recombination. The selective pressures acting in those proteins have yet to be identified.

CONCLUSIONS

The first *L. crispatus* genomes from vaginal isolates from Brazil presented several genes associated with probiotic characteristics. Although mobile genetic elements were detected, they do not present antimicrobial resistance genes that could be transmitted to other bacteria. For safety issues, the functionality of the hemolysin related genes must be further experimentally confirmed. No genomic feature was associated with healthy and BV conditions, and the positive selection was predicted in an S-layer protein and phage related genes but have yet to be investigated.

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests

Rommel Ramos and Vasco Azevedo are Academic Editors for PeerJ.

Author Contributions

- Marcelle Oliveira de Almeida conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Rodrigo Carvalho conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Flavia Figueira Aburjaile conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Fabio Malcher Miranda performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Janaína Canário Cerqueira performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Bertram Brenig performed the experiments, authored or reviewed drafts of the paper, genome sequencing, and approved the final draft.
- Preetam Ghosh analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Rommel Ramos conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Rodrigo Bentes Kato performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Siomar de Castro Soares conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Artur Silva conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Vasco Azevedo conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Marcus Vinicius Canário Viana conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

The genome sequences described here are available at GenBank: JABERN01, JABERO01, JABERP01 and JABERQ01.

Data Availability

The following information was supplied regarding data availability:

The genome sequencing reads are available at NCBI SRA: SRR13201099, SRR13201098, SRR13201097 and SRR13201096.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.11079#supplemental-information.

REFERENCES

Abdelmaksoud AA, Koparde VN, Sheth NU, Serrano MG, Glascock AL, Fettweis JM, Strauss JF, Buck GA, Jefferson KK. 2016. Comparison of *Lactobacillus crispatus* isolates from Lactobacillus-dominated vaginal microbiomes with isolates from microbiomes containing bacterial vaginosis-associated bacteria. *Microbiology* **162(3)**:466–475 DOI 10.1099/mic.0.000238.

Abramov V, Khlebnikov V, Kosarev I, Bairamova G, Vasilenko R, Suzina N, Machulin A, Sakulin V, Kulikova N, Vasilenko N, Karlyshev A, Uversky V, Chikindas ML, Melnikov V. 2014. Probiotic properties of *Lactobacillus crispatus* 2,029: homeostatic interaction with cervicovaginal epithelial cells and antagonistic activity to genitourinary pathogens. *Probiotics and Antimicrobial Proteins* **6(3–4)**:165–176 DOI 10.1007/s12602-014-9164-4.

Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister JR, Bryant SH, Canese K, Charowhas C, Clark K, Dicuccio M, Dondoshansky I, Federhen S, Feolo M, Funk K, Geer LY, Gorelenkov V, Hoeppner M, Holmes B, Johnson M, Khotomlianski V, Kimchi A, Kimelman M, Kitts P, Klimke W, Krasnov S, Kuznetsov A, Landrum MJ, Landsman D, Lee JM, Lipman DJ, Lu Z, Madden TL, Madej T, Marchler-Bauer A, Karsch-Mizrachi I, Murphy T, Orris R, Ostell J, O'sullivan C, Panchenko A, Phan L, Preuss D, Pruitt KD, Rodarmer K, Rubinstein W, Sayers E, Schneider V, Schuler GD, Sherry ST, Sirotkin K, Siyan K, Slotta D, Soboleva A, Soussov V, Starchenko G, Tatusova TA, Todorov K, Trawick BW, Vakatov D, Wang Y, Ward M, Wilbur WJ, Yaschenko E, Zbicz K. 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **44(D1)**:D7–D19 DOI 10.1093/nar/gkv1290.

Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen ALV, Cheng AA, Liu S, Min SY, Miroshnichenko A, Tran HK, Werfalli RE, Nasir JA, Oloni M, Speicher DJ, Florescu A, Singh B, Faltyn M, Hernandez-Koutoucheva A, Sharma AN, Bordeleau E, Pawlowski AC, Zubyk HL, Dooley D, Griffiths E, Maguire F, Winsor GL, Beiko RG, Brinkman FSL, Hsiao WWL, Domselaar GV, McArthur AG. 2020. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research* **48(2)**:D517–D525 DOI 10.1093/nar/gkz1136.

Almeida MO, Carmo FLRD, Gala-García A, Kato R, Gomide AC, Drummond RMN, Drumond MM, Agresti PM, Barh D, Brening B, Ghosh P, Silva A, Azevedo V, Viana MVC. 2019. Lactobacillus crispatus protects against bacterial vaginosis. *Genetics and Molecular Research* **18(4)**:GMR18475 DOI 10.4238/gmr18475.

Andrews S. 2015. FASTQC a quality control tool for high throughput sequence data. *Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/*.

Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research* **44(W1)**:W16–W21 DOI 10.1093/nar/gkw387.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19(5)**:455–477 DOI 10.1089/cmb.2012.0021.

Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Ehrlich D, Maguin E, Van de Guchte M. 2009. Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria. *Proteomics* **9(1)**:61–73 DOI 10.1002/pmic.200800195.

Barrientos-Durán A, Fuentes-López A, De Salazar A, Plaza-Díaz J, García F. 2020. Reviewing the composition of vaginal microbiota: inclusion of nutrition and probiotic factors in the maintenance of eubiosis. *Nutrients* **12(2)**:1–30 DOI 10.3390/nu12020419.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27(4)**:578–579 DOI 10.1093/bioinformatics/btq683.

Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biology* **13(6)**:R56 DOI 10.1186/gb-2012-13-6-r56.

Boot HJ, Kolen CPAM, Pouwels PH. 1995. Identification, cloning, and nucleotide sequence of a silent S-layer protein gene of *Lactobacillus acidophilus* ATCC, 4356 which has extensive similarity with the S-layer protein gene of this species. *Journal of Bacteriology* **177(24)**:7222–7230 DOI 10.1128/JB.177.24.7222-7230.1995.

Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL, Rebelo AR, Florensa AF, Fagelhauer L, Chakraborty T, Neumann B, Werner G, Bender JK, Stingl K, Nguyen M, Coppens J, Xavier BB, Malhotra-Kumar S, Westh H, Pinholt M, Anjum MF, Duggett NA, Kempf I, Nykäsenoja S, Olkkola S, Wieczorek K, Amaro A, Clemente L, Mossong J, Losch S, Ragimbeau C, Lund O, Aarestrup FM. 2020. ResFinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy* **75(12)**:3491–3500 DOI 10.1093/jac/dkaa345.

Branco KMGR, Nardi RMD, Moreira JLS, Nunes AC, Farias LM, Nicoli JR, Carvalho MAR. 2010. Identification and in vitro production of *Lactobacillus antagonists* from women with or without bacterial vaginosis. *Brazilian Journal of Medical and Biological Research* **43(4)**:338–344 DOI 10.1590/S0100-879X2010007500013.

Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology* **17(1)**:1–9 DOI 10.1186/s13059-016-1108-8.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST plus: architecture and applications. *BMC Bioinformatics* **10(1)**:1 DOI 10.1186/1471-2105-10-421.

Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014. In silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy* **58(7)**:3895–3903 DOI 10.1128/AAC.02412-14.

Chang D-H, Rhee M-S, Lee S-K, Chung I-H, Jeong H, Kim B-C. 2019. Complete genome sequence of Lactobacillus crispatus AB70, isolated from a vaginal swab from a healthy pregnant korean woman. *Microbiology Resource Announcements* **8(12)**:e01736-18 DOI 10.1128/MRA.01736-18.

Chokesajjawatee N, Santiyanont P, Chantarasakha K, Kocharin K, Thammarongtham C, Lertampaiporn S, Vorapreeda T, Srisuk T, Wongsurawat T, Jenjaroenpun P, Nookaew I, Visessanguan W. 2020. Safety assessment of a nham starter culture *Lactobacillus plantarum* BCC9546 via whole-genome analysis. *Scientific Reports* **10(1)**:1–12 DOI 10.1038/s41598-020-66857-2.

Clabaut M, Boukerb AM, Racine P, Pichon C, Kremser C, Picot J, Karsybayeva M, Redziniak G, Chevalier S, Feuilloley MGJ. 2020. Draft genome sequence of *Lactobacillus crispatus* CIP

104459, isolated from a vaginal swab. *Microbiology Resource Announcements* **9(2)**:18–19 DOI 10.1128/MRA.01373-19.

Clabaut M, Boukerb AM, Racine P-J, Pichon C, Kremser C, Queiroz A, Karsybayeva M, Redziniak G, Chevalier S, Feuilloley MGJ. 2019. Draft genome sequence of *Lactobacillus crispatus* strain V4, isolated from a vaginal swab from a young healthy nonmenopausal woman. *Microbiology Resource Announcements* **8(38)**:1–2 DOI 10.1128/MRA.00856-19.

Crawley AB, Henriksen ED, Stout E, Brandt K, Barrangou R. 2018. Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli. *Scientific Reports* **8(1)**:1–12 DOI 10.1038/s41598-018-29746-3.

Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler RRM, Butler RRM, Chlenski P, Conrad N, Dickerman A, Dietrich EM, Gabbard JL, Gerdes S, Guard A, Kenyon RW, Machi D, Mao C, Murphy-Olson D, Nguyen M, Nordberg EK, Olsen GJ, Olson RD, Overbeek JC, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomas C, VanOeffelen M, Vonstein V, Warren AS, Xia F, Xie D, Yoo H, Stevens R. 2020. The PATRIC bioinformatics resource center: expanding data and analysis capabilities. *Nucleic Acids Research* **48**:D606–D612 DOI 10.1093/nar/gkz943.

Del Rio B, Alvarez-Sieiro P, Redruello B, Martin MC, Fernandez M, Ladero V, Alvarez MA. 2018. *Lactobacillus rossiae* strain isolated from sourdough produces putrescine from arginine article. *Scientific Reports* **8(1)**:1–10 DOI 10.1038/s41598-017-17765-5.

Dols JAM, Molenaar D, Van der Helm JJ, Caspers MPM, De Kat Angelino-Bart A, Schuren FHJ, Speksnijder AGCL, Westerhoff HV, Richardus JH, Boon ME, Reid G, De Vries HJC, Kort R. 2016. Molecular assessment of bacterial vaginosis by *Lactobacillus* abundance and species diversity. *BMC Infectious Diseases* **16(1)**:180 DOI 10.1186/s12879-016-1513-3.

Doster E, Lakin SM, Dean CJ, Wolfe C, Young JG, Boucher C, Belk KE, Noyes NR, Morley PS. 2019. MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Research* **48(D1)**:D561–D569 DOI 10.1093/nar/gkz1010.

Eslami S, Hadjati J, Motevaseli E, Mirzaei R, Farashi Bonab S, Ansaripour B, Khoramizadeh MR. 2016. *Lactobacillus crispatus* strain SJ-3C-US induces human dendritic cells (DCs) maturation and confers an anti-inflammatory phenotype to DCs. *APMIS* **124(8)**:697–710 DOI 10.1111/apm.12556.

Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, Tyson GH, Zhao S, Hsu C-H, McDermott PF, Tadesse DA, Morales C, Simmons M, Tillman G, Wasilenko J, Folster JP, Klimke W. 2019. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrobial Agents and Chemotherapy* **63(11)**:e00483-19 DOI 10.1128/AAC.00483-19.

France MT, Mendes-Soares H, Forney LJ. 2016. Genomic comparisons of *Lactobacillus crispatus* and *Lactobacillus iners* reveal potential ecological drivers of community composition in the vagina. *Applied and Environmental Microbiology* **82(24)**:7063–7073 DOI 10.1128/AEM.02385-16.

García-Villar N, Hernández-Cassou S, Saurina J. 2009. Determination of biogenic amines in wines by pre-column derivatization and high-performance liquid chromatography coupled to mass spectrometry. *Journal of Chromatography A* **1216(36)**:6387–6393 DOI 10.1016/j.chroma.2009.07.029.

Guizelini D, Raittz RT, Cruz LM, Souza EM, Steffens MBR, Pedrosa FO. 2016. GFinisher: a new strategy to refine and finish bacterial genome assemblies. *Scientific Reports* **6(1)**:34963 DOI 10.1038/srep34963.

Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain JM. 2014. ARG-annot, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial Agents and Chemotherapy* **58(1)**:212–220 DOI 10.1128/AAC.01310-13.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29(8)**:1072–1075 DOI 10.1093/bioinformatics/btt086.

Hemmerling A, Harrison W, Schroeder A, Park J, Korn A, Shiboski S, Foster-Rosales A, Cohen CR. 2010. Phase 2a study assessing colonization efficiency, safety, and acceptability of *Lactobacillus crispatus* CTV-05 in women with bacterial vaginosis. *Sexually Transmitted Diseases* **37(12)**:745–750 DOI 10.1097/OLQ.0b013e3181e50026.

Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J. 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research* **18(5)**:802–809 DOI 10.1101/gr.072033.107.

Hongo JA, De Castro GM, Cintra LC, Zerlotini A, Lobo FP. 2015. POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. *BMC Genomics* **16(1)**:567 DOI 10.1186/s12864-015-1765-0.

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution* **34(8)**:2115–2122 DOI 10.1093/molbev/msx148.

Hynönen U, Palva A. 2013. *Lactobacillus* surface layer proteins: structure, function and applications. *Applied Microbiology and Biotechnology* **97(12)**:5225–5243 DOI 10.1007/s00253-013-4962-2.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30(9)**:1236–1240 DOI 10.1093/bioinformatics/btu031.

Kanehisa M, Sato Y. 2020. KEGG mapper for inferring cellular functions from protein sequences. *Protein Science* **29(1)**:28–35 DOI 10.1002/pro.3711.

Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of Molecular Biology* **428(4)**:726–731 DOI 10.1016/j.jmb.2015.11.006.

Katoh K, Kuma KI, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research* **33(2)**:511–518 DOI 10.1093/nar/gki198.

Kiliç AO, Pavlova SI, Ma WG, Tao L. 1996. Analysis of *Lactobacillus* phages and bacteriocins in American dairy products and characterization of a phage isolated from yogurt. *Applied and Environmental Microbiology* **62(6)**:2111–2116 DOI 10.1128/AEM.62.6.2111-2116.1996.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9(4)**:357–359 DOI 10.1038/nmeth.1923.

Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23(21)**:2947–2948 DOI 10.1093/bioinformatics/btm404.

Letunic I, Bork P. 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47(W1)**:W256–W259 DOI 10.1093/nar/gkz239.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25(16)**:2078–2079 DOI 10.1093/bioinformatics/btp352.

Linares DM, MartÍn MC, Ladero V, Alvarez MA, Fernández M. 2011. Biogenic amines in dairy products. *Critical Reviews in Food Science and Nutrition* **51(7)**:691–703 DOI 10.1080/10408398.2011.582813.

Liu B, Zheng D, Jin Q, Chen L, Yang J. 2019. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Research* **47(D1)**:D687–D692 DOI 10.1093/nar/gky1080.

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. 2020. CDD/SPARCLE:the conserved domain database in 2020. *Nucleic Acids Research* **48(D1)**:D265–D268 DOI 10.1093/nar/gkz991.

Macklaim JM, Fernandes AD, Di Bella JM, Hammond JA, Reid G, Gloor GB. 2013. Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis. *Microbiome* **1(1)**:1–11 DOI 10.1186/2049-2618-1-12.

Maddison WP. 2000. Testing character correlation using pairwise comparisons on a phylogeny. *Journal of Theoretical Biology* **202(3)**:195–204 DOI 10.1006/jtbi.1999.1050.

Marconi C, El-Zein M, Ravel J, Ma B, Lima MD, Carvalho NS, Alves RRF, Parada CMGL, Leite SHM, Giraldo PC, Gonçalves AK, Franco EL, Silva MG. 2020. Characterization of the vaginal microbiome in women of reproductive age from 5 regions in Brazil. *Sexually Transmitted Diseases* **47**:562–569 DOI 10.1097/OLQ.0000000000001204.

Martinez RCR, Franceschini SA, Patta MC, Quintana SM, Nunes ÁC, Moreira JLS, Anukam KC, Reid G, De Martinis ECP. 2008. Analysis of vaginal lactobacilli from healthy and infected Brazilian women. *Applied and Environmental Microbiology* **74(14)**:4539–4542 DOI 10.1128/AEM.00284-08.

Meier-Kolthoff JP, Göker M. 2019. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nature Communications* **10(1)**:2182 DOI 10.1038/s41467-019-10210-3.

Melsted P, Halldórsson BV. 2014. KmerStream: streaming algorithms for k-mer abundance estimation. *Bioinformatics* **30(24)**:3541–3547 DOI 10.1093/bioinformatics/btu713.

Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32(1)**:268–274 DOI 10.1093/molbev/msu300.

Ojala T, Kankainen M, Castro J, Cerca N, Edelman S, Westerlund-Wikström B, Paulin L, Holm L, Auvinen P. 2014. Comparative genomics of *Lactobacillus crispatus* suggests novel mechanisms for the competitive exclusion of Gardnerella vaginalis. *BMC Genomics* **15(1)**:1–21 DOI 10.1186/1471-2164-15-1070.

Pan M, Hidalgo-Cantabrana C, Barrangou R. 2020. Host and body site-specific adaptation of *Lactobacillus crispatus* genomes. *NAR Genomics and Bioinformatics* **2(1)**:1–14 DOI 10.1093/nargab/lqaa001.

Pariza MW, Gillies KO, Kraak-Ripple SF, Leyer G, Smith AB. 2015. Determining the safety of microbial cultures for consumption by humans and animals. *Regulatory Toxicology and Pharmacology* **73(1)**:164–171 DOI 10.1016/j.yrtph.2015.07.003.

Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. 2007. Genes under positive selection in Escherichia coli. *Genome Research* **17(9)**:1336–1343 DOI 10.1101/gr.6254707.

Petit RA, Read TD. 2020. Bactopia: a flexible pipeline for complete analysis of bacterial genomes. *BioRxiv* DOI 10.1128/mSystems.00190-20.

Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ. 2011. Vaginal

microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences* **108(Supplement_1)**:4680–4687 DOI 10.1073/pnas.1002611107.

Raya RR, H'bert EM. 2009. Isolation of phage via induction of lysogens. In: Clokie MRJ, Kropinski A, eds. *Bacteriophages Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions*. New York: Humana Press, 23–32.

Romano A, Ladero V, Alvarez MA, Lucas PM. 2014. Putrescine production via the ornithine decarboxylation pathway improves the acid stress survival of *Lactobacillus brevis* and is part of a horizontally transferred acid resistance locus. *International Journal of Food Microbiology* **175**:14–19 DOI 10.1016/j.ijfoodmicro.2014.01.009.

Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes* **9(1)**:88 DOI 10.1186/s13104-016-1900-2.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30(14)**:2068–2069 DOI 10.1093/bioinformatics/btu153.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M, ed. *Gene Prediction*. New York: Humana Press, 227–245.

Siguier P, Gourbeyre E, Chandler M. 2014. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews* **38(5)**:865–891 DOI 10.1111/1574-6976.12067.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Research* **19(6)**:1117–1123 DOI 10.1101/gr.089532.108.

Smith SB, Ravel J. 2017. The vaginal microbiota, host defence and reproductive physiology. *Journal of Physiology* **595(2)**:451–463 DOI 10.1113/JP271694.

Tachedjian G, O'Hanlon DE, Ravel J. 2018. The implausible *in vivo* role of hydrogen peroxide as an antimicrobial factor produced by vaginal microbiota. *Microbiome* **6(1)**:3–7 DOI 10.1186/s40168-018-0418-3.

Van der Veer C, Hertzberger RY, Bruisten SM, Tytgat HLP, 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)**:49 DOI 10.1186/s40168-019-0667-9.

Van Heel AJ, De Jong A, Song C, Viel JH, Kok J, Kuipers OP. 2018. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Research* **46(W1)**:W278–W281 DOI 10.1093/nar/gky383.

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* **13(6)**:e1005595 DOI 10.1371/journal.pcbi.1005595.

Witkin SS, Mendes-soares H, Linhares IM, Jayaram A, Ledger WJ, Forney J. 2013. Influence of vaginal bacteria and D- and L-lactic acid isomers on vaginal extracellular matrix metalloproteinase inducer: implications. *mBio* **4**:1–7 DOI 10.1128/mBio.00460-13.

Xie Z, Tang H. 2017. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* **33(21)**:3340–3347 DOI 10.1093/bioinformatics/btx433.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **19(6)**:908–917 DOI 10.1093/oxfordjournals.molbev.a004148.

Zafar H, Saier MH. 2020. Comparative genomics of the transport proteins of ten *Lactobacillus* strains. *Genes* **11(10)**:1234 DOI 10.3390/genes11101234.

Zhang Q, Zhang L, Ross P, Zhao J, Zhang H, Chen W. 2020. Comparative genomics of *Lactobacillus crispatus* from the gut and vagina reveals genetic diversity and lifestyle adaptation. *Genes* **11(4)**:360 DOI 10.3390/genes11040360.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29(21)**:2669–2677 DOI 10.1093/bioinformatics/btt476.