EMbaRC

European Consortium of Microbial Resource Centres

Grant agreement number: **228310**

Seventh Framework Programme

Capacities

Research Infrastructures

Combination of Collaborative Project and Coordination and Support Actions

Deliverable D.15.26 (formerly D.JRA2.3.2)

Title: Publication on comparative genomic analysis of prokaryote targeted species

Due date of deliverable: M30

Actual date of submission: M44

Start date of the project: 1st February 2009 **Duration:** 44 months

Organisation name of the lead beneficiary: IP

Version of this document: V1

Dissemination level: PU

PU	**Public**	
PP	Restricted to **other programme participants** (including the Commission)	
RE	Restricted to a **group defined by the Consortium** (including the Commission)	

EMbaRC is financially supported by the Seventh Framework Programme (2007-2013) of the European Communities, Research Infrastructures action

Document properties	
Project	EMbaRC
Workpackage	WP JRA2
Deliverable	
Title	D.15.26 (formerly D.JRA2.3.2)
Version number	V1
Authors	Christiane Bouchier, IP
Abstract	In order to improve species identification by founding new genes [Partners involved in JRA2.3.2 decided to carry on a whole genome analysis for *Lactobacillus* species based on the Next-Generation Sequencing technology Illumina. 20 strainsof Lactobacillus selected by the partners have been sequenced and reads have been assembled. *De novo* contigs have been scaffolded into a draft genome. Totally, ten draft genome sequences have been published and sequences are available in public database].
Validation process	Document prepared by IP and submitted to the Coordinator for agreement.

Revision table			
Date	***Version***	***Revised by***	***Main changes***
Sept. 2013	0.1	Inra	Complements in the text and in the list of publications

Table of contents

Abbreviations

NGS	Next Generation Sequencing
BLAST	Basic Local Alignment Search Tool

1 Background and Objectives

The genus *Lactobacillus* represents the largest group of optionally anaerobic, catalase-negative, non-spore-forming, gram-positive, rod-shaped organisms which produce lactic acid as a major end product of metabolism (1) and currently contains more than 150 recognized species or subspecies. Species of the genus *Lactobacillus* have been isolated from a wide range of habitats such as the oral cavity, the vagina and the gastrointestinal tracts of humans and animals, food, vegetation and sewage are of great commercial importance due to their use in the production of a range of fermented dairy, meat and vegetable products. They are also the most widely used probiotics meant for promoting a healthy lifestyle. Some species within the Lactobacillus genus are hard to differenciate by current approaches.

The preliminary step of the work package JRA2 was to build a tree using the 16S rRNA gene sequences of the strains available in order to identify areas of closest species hard to separate by using the 16S rRNA gene sequences. Then, the subtask JRA2.1.2 (Development of new molecular markers for prokaryotes) assayed some gene markers for a better characterisation of *Lactobacillus* species). The third step was to perform whole genome sequencing, in order to detect new gene markers able to offer better resolution of some clusters in the phylogenetic tree.

Here is the list of some clusters “hard to separate” selected by the partners for further investigation:

- *Lactobacillus plantarum* group
- *Lactobacillus acidophilus* group and *L. helveticus*
- *Lactobacillus delbrueckii* group

Type strains of new Lactobacillus species, which were deposited years ago but never deeply characterized were also included in this task, in order to provide more complete data about these new species and to facilitate also their identification : *Lactobacillus hominis* CRBIP 24.179^{T} isolated from an human intestine (clinical isolate), *Lactobacillus pasteurii* CRBIP 24.76^{T} and the closely related *Lactobacillus gigeriorum* isolated from chicken crop.

Partners selected type strains and strains of *Lactobacillus* species according to their biotope.

Whole genome sequencing was performed for 20 strains using the NGS technology Illumina.

2 Methods

2.1 Lactobacillus strains

Lactobacillus acidophilus group:

Selection of five strains of *Lactobacillus acidophilus*: CIP 76.13^T (type strain) isolated from human, two strains CIRM-BIA 442, CIRM-BIA 455 isolated from dairy products and strains DSM 9126, DSM 20242 of unknown origin but of technological industrial interest.

Lactobacillus delbrueckii group:

Selection of three type strains: *L. delbrueckii* subsp. *delbrueckii* CIP 57.8^T from sour grain mash, *L. delbrueckii* subsp. *lactis* CIP 101028^T isolated from Emmental swiss and *L. delbrueckii* subsp.*indicus* CIP 57.8^T from dairy fermented product. Two additional strains included in a subcluster in the *L. delbrueckii* phylogenetic group have been sequenced: *L. equicursoris* CIP 110162^T isolated from racehorse in Japan and *Lactobacillus sp.* CRBIP 24.137, a clinical isolate from a human urine sample.

Lactobacillus plantarum group:

Selection of two strains *L. paraplantarum* DSM 10667 & *L. plantarum* subsp.*argentoratensis* DSM 16365.

Lactobacillus helveticus strains:

Based on their interest in industrial products, five *Lactobacillus helveticus* strains isolated from dairy products were selected: CIRM-BIA 101 isolated from cheese and the other strains CIRM-BIA 103 and 104 isolated from artisanal starter culture of Comté and Italian cheese respectively and CIRM-BIA 951 and 953 were isolated from milk products from Czech Republic.

Moreover two strains isolated in the early 1960s and deposited in the Institut Pasteur Collection were sequenced : *Lactobacillus pasteurii* CRBIP 24.76^T of unknown origin and *Lactobacillus hominis* CRBIP 24.179^T isolated from an human intestine.

Lactobacillus gigeriorum strain 202^T, renamed as CRBIP 24.85^T isolated in the early 1980s from chicken crop was also included in this whole genome sequencing.

2.2 Sequencing

The whole genome sequencing was done using the HiSeq® 2000 system (Illumina) as next-generation sequencing (NGS) technology except for the strain CIP 57.8^T which was sequenced on the GAIIX system (Illumina) a paired-end of 54 bases. Illumina library preparation and sequencing followed standard protocols developed by the supplier (Illumina, San Diego, CA). Briefly, genomic DNA was sheared by sonication (Bioruptor, Diagenode), and sheared fragments were end-repaired and phosphorylated. Blunt-end fragments were A-tailed, and sequencing adapters were ligated to the fragments. Inserts were sized using AMPure XP Beads (Agencourt) (± 500 bp) and enriched with 10 cycles of PCR before library quantification and validation. Hybridization of the library to the flow cell and bridge amplification was performed to generate clusters, and paired-end reads of 90 or 100 cycles were collected on a HiSeq® 2000 (Illumina). After sequencing was complete, image analysis, base calling, and error estimation were performed using Illumina Analysis Pipeline version 1.7. Technical problems occurred during the sequening phase, in the Illumina technology provoking a significant delay of obtention of the sequencing data.

2.3 Bioinformatics analysis

De novo assembly:

The strategy used for the *de novo* assembly was the following: after quality-filtering the reads were assembled with three different softwares: Abyss1.2.5 (2), VelvetOptimizer 2.2.0 (3) and CLC Genomics 4.5 (CLCbio, Denmark). The resulting contigs were compared using Mauve version 2.3.1 (4) in order to select the best scaffold. Many problems again were encountered in the use of the Mauve software generating additional delay and after several temptatives, a strategic decision was taken between Pasteur and Inra, to use the Agmial platform at Inra Jouy en Josas.

The contigs were thus annotated with the AGMIAL platform (5), an integrated bacterial genome annotation system. Prediction of coding sequences used the self-training gene detection software SHOW based on hidden Markov models (http://genome.jouy.inra.fr/ssb/SHOW/). tRNA and rRNA were detected using tRNAscan-SE (6) and RNAmmer (7) software, respectively.

Mapping analysis:

After quality-filtering the reads were mapped on the genome reference sequence using CLC Assembly Cell v 4.5 and SNPs and indels were detected by CLC Genomics Workbench v4.5.1 (CLCbio, Denmark).

3 Results

3.1 De novo assembly strategy of lactobacilli genomes:

Here is the summary of results obtained for all sequenced strains:

	Strain	**HiSeq 2000 Illumina - run**	***de novo* assembly program selected**	**Number of Scaffolds with size above 1000 bases**	**Genome size in Mb**	**Length of the longest contig in Kb**
L. d.delbrueckii	CIP 57.8^{T}	GAII - PE -54 bases	Abyss K25	267	1,8	48
L. d.lactis	CIP 101028^{T}	PE -100 bases	Abyss 40	173	1,9	90
L. d.indicus	CIP 108704^{T}	PE -100 bases	Velvet	1342	3,3	24
L. equicursoris	CIP 110162^{T}	PE -100 bases	Abyss K60	116	2,2	123
L. sp strain 66C	CRBIP 24.137	PE -100 bases	Abyss K60	62	2,4	254
L. hominis	CRBIP 24.179^{T}	PE -100 bases	Velvet	28	1,9	365
L. pasteurii	CRBIP 24.76^{T}	PE -90 bases	Abyss K40	29	1,9	313
L. gigeriorum	CRBIP 24.85^{T}	PE -90 bases	Velvet	60	1,9	205
L. acidophilus	DSM 20242	PE -90 bases	Abyss K60	20	2	313
L. acidophilus	DSM 9126	PE -90 bases	Velvet	26	2	552
L. acidophilus	CIRMBIA 442	PE -90 bases	Abyss 40	19	1,9	673
L. acidophilus	CIRMBIA 445	PE -90 bases	Velvet	22	2	877
L. acidophilus	CIP 76.13^{T}	PE -90 bases	Velvet	34	1,9	671
L. helveticus	CIRMBIA 101$^{(T)}$	PE -100 bases	Abyss 40	212	2,1	60
L. helveticus	CIRMBIA 103	PE -100 bases	Velvet	191	2	64
L. helveticus	CIRMBIA 104	PE -100 bases	Velvet	171	2,1	56
L. helveticus	CIRMBIA 951	PE -100 bases	Velvet	159	1,9	84
L. helveticus	CIRMBIA 953	PE -100 bases	Velvet	151	2,1	89
L. paraplantarum	DSM10667	PE -100 bases	Velvet	152	3,4	233
L. p.argentoratensis	DSM16365	PE -100 bases	Abyss K60	150	3,5	213

PE = paired end run

Currently, excepted for *L. plantarum*, the draft genome sequence of all these strains were deposited into EMBL/GenBank databank and publication are in press or on going for most of them (list below)

List of publications done by the EMbaRC consortium relatives to the JRA2.3.2 :

1- Cousin S., Gullat-Okalla ML., Motreff L., Gouyette C., Bouchier C., Clermont D., and Bizet C., **2012**. *Lactobacillus gigeriorum* sp. Nov., isolated from chicken crop. International Journal of systematic and Evolutionary Microbiology, 62, 330-334

2- **Cousin S**, Ma L, Creno S, Clermont D, Loux V, Bizet C, Bouchier C, **2012**. Draft genome sequence of ***Lactobacillus gigeriorum*** CRBIP 24.85^{T} isolated from a chicken crop. J Bacteriol. ;194(21):5973.

3- Cousin S,, Motreff L., Gullat-Okalla ML, Gouyette C., Spröer C., Schumann P., Begaud E., Bouchier C., Clermont D., and C. Bizet, **2013**. ***Lactobacillus pasteurii*** sp. nov. and ***Lactobacillus hominis*** sp.nov., two novel species. International Journal of systematic and Evolutionary microbiology, 63, 53-59

4- Cousin S, Clermont D, Creno S, Ma L, Loux V, Bizet C, Bouchier C, **2013.** Draft genome sequence of ***Lactobacillus pasteurii*** CRBIP 24.76^{T} Genome Announc. July/August 2013 1:e00660-13. PMID23969061

5- Cousin S, Creno S, Ma L, Clermont D, Loux V, Bizet C, Bouchier C, **2013**. Draft genome sequence of ***Lactobacillus hominis*** Strain CRBIP 24.179^{T}, isolated from human intestine. Genome Announc. July/August 2013 1:e00662-13. PMID 23969062

6- Cousin S, Loux V, Ma L, Creno S, Clermont D, Bizet C, Bouchier C, 2013. **Draft genome sequences of *Lactobacillus equicursoris*** CIP 110162^{T} and ***Lactobacillus sp*.** Strain CRBIP 24.137, isolated from thoroughbred racehorse feces and human urine, respectively. Genome Announc. July/August 2013 1:e00663-13.

. **7-** Falentin H**,** Cousin S, Clermont D, Creno S, Ma L, Chuat V , Loux V, Rüdiger P, Bizet C, Bouchier C. 2013. **Draft genome sequences of five strains of *Lactobacillus acidophilus***, Strain CIP 76.13^{T}, isolated from humans, strains CIRM-BIA 442 and CIRM-BIA 445, isolated from dairy products, and strains DSM 20242 and DSM 9126 of unknown origin. Genome Announc. July/August 2013 1:e00658-13.PMID23969059

The publication of the draft genome sequence of three *Lactobacillus delbrueckii* species is in progress as well as the genome sequence of the five strains of *Lactobacillus helveticus* whose draft genomes were deposited in the EMBL under the following references :

'LHCIRMBIA101' > PRJEB1536.
'LHCIRMBIA103' > PRJEB1537.
'LHCIRMBIA104' > PRJEB1540.
'LHCIRMBIA951' > PRJEB1541.
LHCIRMBIA953' > PRJEB1542.

All this work increases significantly the data available regarding whole genomes in lactobacilli. Indeed before this task, in the group *L. acidophilus* (acidophilus) only one strain was completely sequenced (NCFM) whereas 4 *L. gasseri* (whole sequence for ATCC 33323 and

scaffold for JV-V03, MV-22 and 202-4), 2 *L johnsonii*. (FI9785, NCC 533), and only one complete sequence of *L. debrueckii* subsp. *bulgaricus* ATCC 11842T. were accessible in Genbank.

Regarding *L. helveticus* before the work described here two complete sequence were already made (strain DPC 4571 and Strain H10) as well as scaffold for strain DSM20075 and MTCC 5463. The whole genome sequence of five more strains including the type strain of *L. helveticus* were provided in this work.

3.2 Mapping strategy and comparative genomic :

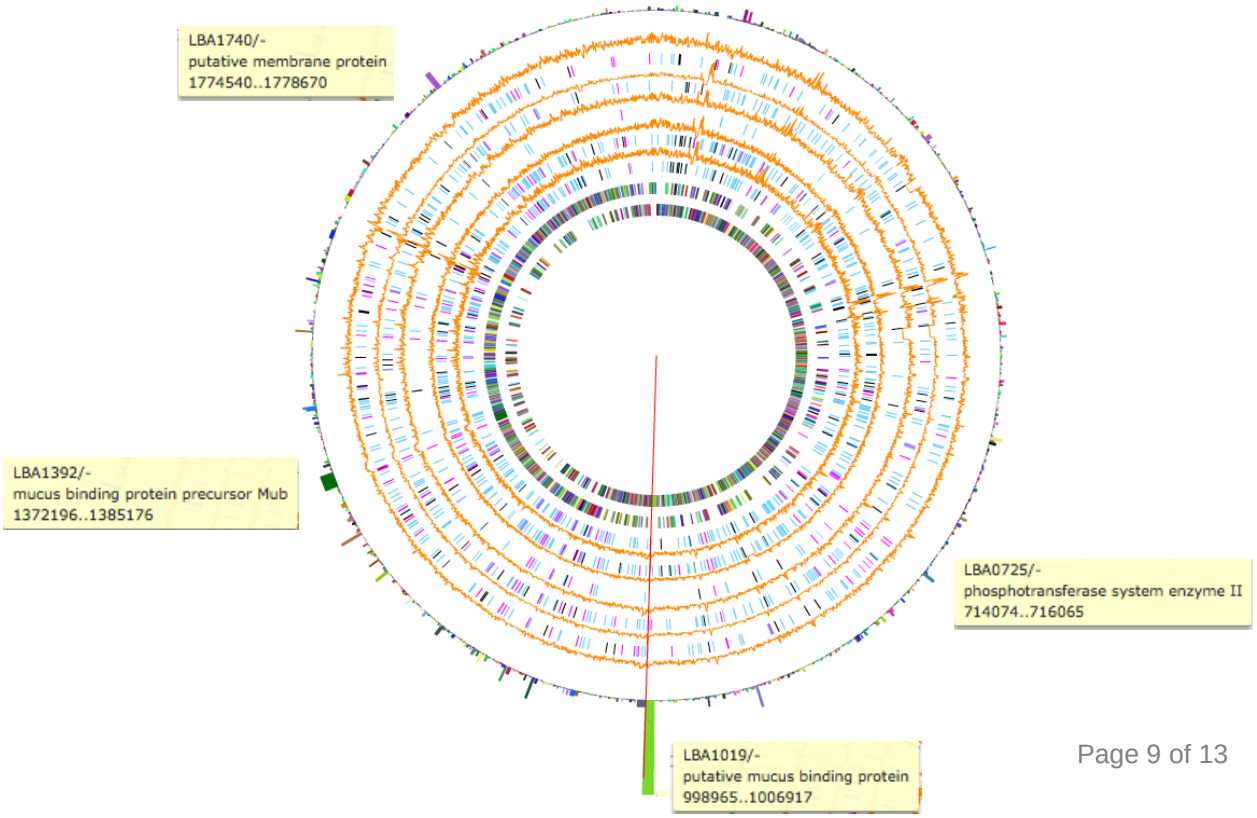
Lactobacillus acidophilus:

Five strains of *L acidophilus* were mapped on the genome reference sequence (*Lactobacillus acidophilus* strain NCFM – GenBank accession CP000033) using CLC Genomics. All SNPs detected by reference genome sequence mapping were visualized by a graphic interface, SynTview (tool developed by Pierre Lechat – Genopole- Institut Pasteur). Data are available on the web page :

http://genopole.pasteur.fr/SynTView/flash/EMBARC/Lactobacillus_acidophilus//SynWebSNP.html

SynTview analysis search for differences between the 5 strains using SNP and indel visualisation.

4 genes presenting a high rate of SNPs between the 5 strains were found and will be further explored as identification tools :



Lactobacillus delbrueckii:

The three type strain genomes of *L. delbrueckii* were sequenced on the HiSeq2000 system and the quality-filtered reads were mapped on the reference genome sequence, *L. delbrueckii* subsp. *bulgaricus* ATCC 11842^T. After a step of detection and filtering of variants, 17 annotated genes which had more than 3 SNPs were selected. The sequence of these 17 genes was extracted from high-throughput sequencing data of *L. delbrueckii delbrueckii*, *L. delbrueckii lactis*, *L. delbrueckii indicus, L. equicursoris* or/and strain 66C in order to search for primer to partially amplify those

genes. Then alignments of genes were performed with BLAST program (v2.2.5) using the 17 genes of *L. delbrueckii bulgaricus* ATCC 11842^T strain. 20 sets of primers were designed to cover these 17 genes. **Primers were used for both PCR amplification and cycle sequencing of purified PCR products in the JRA 2.1.2 to develop a MLST scheme**.

List of the selected genes:

Gene	Length of gene (bp)	Function
dnaC	1371	**replicative DNA helicase**
guaB/guaA	1554	**inosine-5-monophosphate dehydrogenase / GMP synthase ; IMP dehydrogenase/GMP reductase ; bifunctional gmp synthase/glutamine amidotransferase protein**
mfd	3477	**transcription-repair coupling factor**
uvrB/uvrA	2856	**excinuclease ABC subunit B/A ; excinuclease ATPase subunit; excision endonuclease subunit UvrA ; UvrABC system protein A (UvrA protein)**
ftsW	1203	**cell division protein FtsW ; cell division membrane protein**
comEC	2187	**competence protein ComEC ; metallo-beta-lactamase superfamily hydrolase**
dfp	1200	**phosphopantothenoylcysteine synthetase/decarboxylase; phosphopantothenate-cysteine ligase/phosphopantothenoyl…**
relA	2262	**GTP pyrophosphokinase ; guanosine polyphosphate pyrophosphohydrolase/synthetase ; PpGpp synthetase ; GTP pyrophosphohydrolases/synthetases RelA/SpoT family ; RelA/SpoT protein ; (p)ppGpp synthetase I SpoT/RelA**
addB	3540	**ATP-dependent deoxyribonuclease subunit B ; ATP-dependent nuclease,**

		subunit B ; ATP-dependent helicase/deoxyribonuclease subunit b
prtB	5841	**proteinase precursor subtilisin-like serine protease proteinase b**
glyS/glyQ	2094	**glycyl-tRNA synthetase beta/alpha subunit**
hrcA	1083	**heat-inducible transcription repressor**
smc	3546	**chromosome partition protein SMC chromosome segregation ATPase condensin subunit smc**
polA	2661	**DNA polymerase I**
Ldb1525	657	**tRNA (guanine-N(7)-)-methyltransferase**
mutL/mutS1	1962	**DNA mismatch repair protein MutS/mutl**
thrC	1479	**threonine synthase l-threonine synthase**

Conclusion

The whole-genome sequencing approach was performed on a selection of *Lactobacillus* species hard to separate in the phylogenetic tree based on the 16S rRNA gene. For all the sequenced strains we obtained a draft genome sequence. Currently, half of these draft genomes was now published and available to the scientific community. The other half is about to be published.

The mapping analysis on a genome reference sequence allowed to find some new genes markers for the distinction of the 4 subspecies of *Lactobacillus delbrueckii*. A similar analysis is in progress for the five *L. acidophilus* strains.

The whole-genome sequencing approach provided contiguous segments (contigs) representing the entire genome which could be used for several studies. The results obtained in this subtask contributed to the supply of data to the microbial scientific community an d more particularly will contribute to widening of knowledge about biodiversity in specific phyla.

References

1-Hammes WP, Hertel C. 2009. Genus I. *Lactobacillus*. p 465-511. *In* Vos P, Garrity G, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, Whitman WB (Ed), Bergey's manual of systematic bacteriology, 2nd ed, vol 3. Springer, New York, NY.

2- Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–23.

3- Zerbino, D. R. and E. Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18:821-829.

4- Darling ACE, Mau B, Blattner FR, Perna NT. 2004 Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. **14:**1394–1403.

5-Bryson K, Loux V, Bossy R, Nicolas P, Chaillou S, van de Guchte M, Penaud S, Maguin E, Hoebeke M, Bessières P, Gibrat JF. 2006. AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res. 34:3533–3545.

6- Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25:955–964.

7- Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35:3100–3108.

Significance of this deliverable *D. JRA2.3.2: Comparative genomic analysis of procaryotes targeted species & publications*

The exploration of new approaches in species identification overcome the actual limits in species identification for closely lactic species such as *L. helveticus and L. acidophilus.* The availability in Genbank of the 13 draft genomes of LAB species with economical and technological value will be of great interest for people working on these species which are generally less explored than species of medical interest. For example the draft genomes of *L. delbrueckii* mapped on the reference genome sequence, *L. delbrueckii bulgaricus* ATCC 11842^{T}, allowed to find new molecular markers for the distinction of the 4 subspecies of *Lactobacillus delbrueckii* via a MLST approach which was presented in the JRA2.1.2.

A specific effort has been done to publish the genomes draft obtained here **6 publications already published or in press and 2 submitted**. Submitted Data are now available for genomic comparison that are on going within members of the consortium and within other laboratories either public or private, as the species targeted are of high technological interest.