JOURNAL OF BACTERIOLOGY, May 2009, p. 3420–3424 Vol. 191, No. 10
0021-9193/09/$08.00+0 doi:10.1128/JB.01672-08

Large Intergenic Cruciform-Like Supermotifs in the *Lactobacillus plantarum* Genome▽†

Michiel Wels,[1,2,3]* Roger S. Bongers,[1,3] Jos Boekhorst,[2] Douwe Molenaar,[1,3] Mark Sturme,[4] Willem M. de Vos,[1,4] Roland J. Siezen,[1,2,3] and Michiel Kleerebezem[1,3,4]

TI Food and Nutrition, Wageningen, The Netherlands[1]; Radboud University Nijmegen-Medical Centre/NCMLS, Nijmegen, The Netherlands[2]; NIZO Food Research, Ede, The Netherlands[3]; and Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands[4]

Received 28 November 2008/Accepted 5 March 2009

Twenty-four *Lactobacillus plantarum* supermotifs (LPSMs) with lengths from ~800 to 1,000 nucleotides were identified in the *L. plantarum* genome. LPSMs were conserved in other *L. plantarum* strains but not in other species. Secondary structure analysis predicted that LPSMs may fold into cruciform-like structures. Preliminary experiments indicate that the LPSMs are transcribed.

Intergenic regions of bacterial genomes are known to contain many small elements, such as promoters, transcription or translation regulation sequences, or terminators, that play a role in gene expression. Occasionally, much larger functional noncoding regions are encountered that contain larger functional genetic elements. These elements are involved in transcription regulation (e.g., riboswitches and T-box elements [8, 20]) or function as *trans*-acting RNA elements, blocking transcription or translation by binding to their target (e.g., RNAIII in *Staphylococcus aureus* [17, 18]). Another example of well-described noncoding elements is CRISPRs, RNA elements that protect the cell against phage attacks. CRISPRs are known to be distributed in a subset of lactic acid bacteria genomes (3, 10). For these elements, it was shown that they are transcribed to RNA and are processed by the associated enzyme complex named CAS to generate the functional small RNAs that interfere with phage replication or function (4). In addition, there are many examples of repeated elements in bacterial chromosomes with an undefined function (15, 16, 21). The largest repeated elements described to date have been identified exclusively in *Staphylococci* species and have been designated *Staphylococcus aureus* repeat (STAR) elements (5). STAR elements are internally repeated DNA sequences with lengths, in general, of between 300 and 600 nucleotides (nt), which have been found up to 21 times in one genome (5). The function of these STAR elements has not been described in the literature.

In this study, we identify conserved genetic elements in the intergenic regions of *Lactobacillus plantarum*. Initially, all 74 noncoding regions larger than 700 nt were scanned for conserved sequence motifs using MEME software (1) with the settings "anr" (any number of repetitions), a minimum number of occurrences of 20, a width range of between 10 and 100 nt, and a maximum of 20 different motifs. This analysis revealed 20 different motifs with various lengths, from 19 to 100 nt. Remarkably, 19 motifs (all smaller than 50 nt) were found to colocalize in the same intergenic loci, indicating the existence of a large conserved intergenic sequence in *L. plantarum*. Detailed manual inspection of the conserved order of these motifs revealed a >800-nt supermotif, which is termed here *Lactobacillus plantarum* supermotif (LPSM).

A MAST analysis (2) on the total chromosomal sequence revealed 26 intergenic regions that encompassed at least 10 consecutive motifs (<50 nt between two consecutive motifs). A hidden Markov model (HMM) was built from the complete alignment of these 26 regions using Muscle software (7) and used to search the complete genomic sequence of *L. plantarum* WCFS1 (using HMMer [6]). In an iterative procedure, hits below the threshold (E value of 0.1) were used to build a refined HMM until no additional hits were found. In total, 66 LPSM hits were identified. All minus-strand hits exactly colocalized with a plus-strand hit, indicating that there are 33 copies of a conserved bidirectional LPSM within the chromosome of *L. plantarum* WCFS1 (for an alignment of the 33 copies, see file S1 in the supplemental material). The LPSM sequences did not coincide with statistical features on the genome, such as GC%, codon adaptation index, or comparative genome hybridization data (gathered from reference 14) (Fig. 1). However, the LPSMs were found to be statistically overrepresented on the second replichore of the genome (chi-square test; $P < 0.05$). A complete list of LPSM positions, the best score with the HMM (including the direction of the best hit), and their sequence conservation to other LPSM sequences are summarized (see Table S1 in the supplemental material). The genomic context and flanking gene order of all LSPMs was compared with those of other related bacteria. Seventeen out of 33 LPSMs were found in regions that share synteny with other genomes (some examples are displayed in Fig. 2), in particular with genomes of other lactic acid bacteria (13). The corresponding intergenic regions in *L. plantarum* were consistently much longer, indicating that LPSMs are uniquely present in these loci in the *L. plantarum* genome. In addition, no hits above the threshold (0.1) were found by HMM searches against all publicly available genomes found in

* Corresponding author. Mailing address: NIZO Food Research BV, P.O. Box 20, 6710 BA, Ede, The Netherlands. Phone: 31 318 659 511. Fax: 31 318 650 400. E-mail: michiel.wels@nizo.nl.

† Supplemental material for this article may be found at http://jb.asm.org/.

▽ Published ahead of print on 13 March 2009.

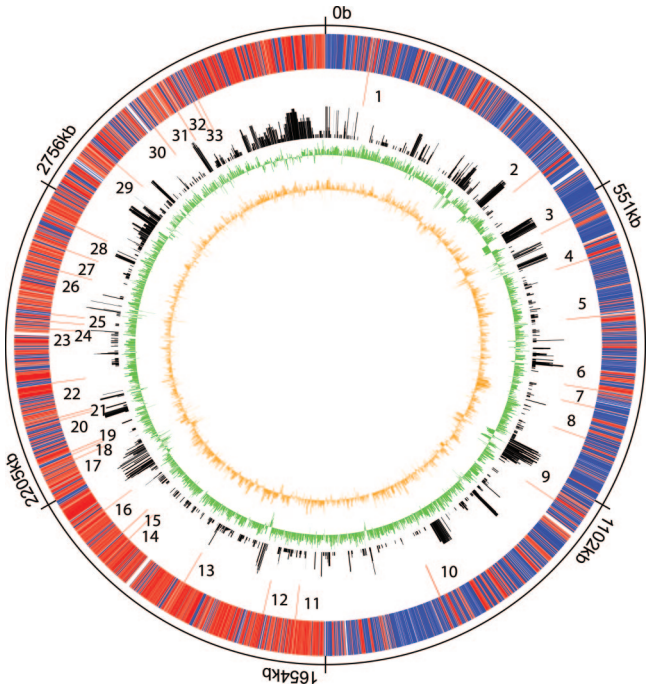

FIG. 1. Chromosome wheel displaying the identified LPSM hits on the chromosome of *L. plantarum*. Outer ring (ring 1), predicted open reading frames (plus and minus strand, dark and light, respectively); ring 2, LPSM occurrences; ring 3, relative conservation among 20 *L. plantarum* strains as determined by array-based chromosome profiling, with the peak height representing the number of strains in which specific chromosomal regions of strain WCFS1 were scored as absent in other strains of *L. plantarum* (14); ring 4, G+C content centered around the median G+C content; ring 5, codon adaptation index. Figure generated using MGV software (11).

the NCBI repository (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/), indicating that the full-length LPSM is unique for *L. plantarum*.

A PCR approach was used to analyze conservation of 15 of the LPSMs among 10 additional *L. plantarum* strains (see file S2 in the supplemental material). PCR amplifications were carried out with an automated thermal cycler (Perkin-Elmer, Shelton, CT) using *Taq* DNA polymerase (Gibco-BRL Life Technologies, Breda, The Netherlands). Primers were designed to specifically anneal to genes that flank 15 of the LPSMs present in the *L. plantarum* WCFS1 genome and were used to amplify the corresponding regions in the chromosomal DNA isolated from 10 other *L. plantarum* strains. Selection of these strains was based on their relative genetic distance to the sequenced strain WCFS1, which was previously determined by gene presence profiling using comparative genome hybridization (14). In 117 cases (out of the total of 150), these PCRs generated detectable amplification products, of which 111 were of a size comparable to the product obtained using *L. plantarum* WCFS1 chromosomal DNA as a template. Of the remaining six products, three appeared to be significantly shorter than the WCFS1 product and three were longer (at least a 500-bp difference). Moreover, none of the LPSMs was found to be unique for WCFS1 only, while all *L. plantarum* strains analyzed contained at least eight LPSMs. Therefore, it can be concluded that LPSMs are generally found on the chromosomes of different *L. plantarum* strains and that their chromosomal locations are largely conserved among strains. Eight LPSM amplicons obtained from these strains were cloned in *Escherichia coli* using the pGEM-T easy vector system (Promega Corp., Madison, WI), and transformants were readily obtained, suggesting that these LPSM sequences can be stably cloned in this host organism. Sequencing of the inserts was performed (BaseClear, Leiden, The Netherlands) using

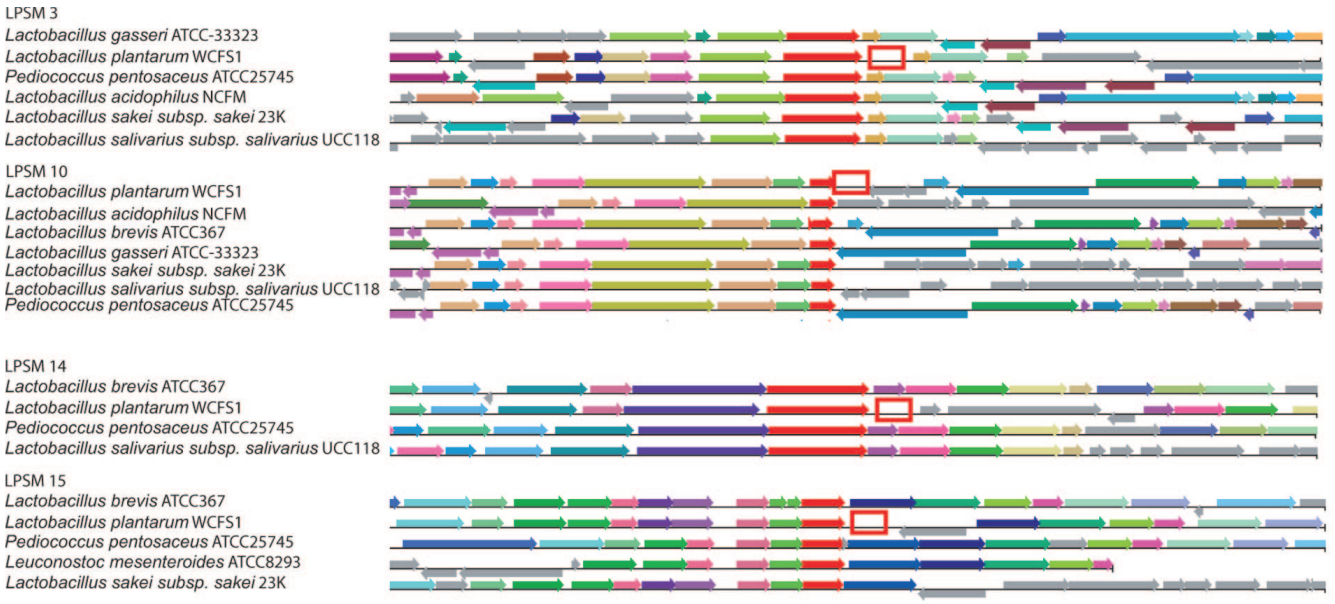

FIG. 2. Genome context of four LPSM locations in *L. plantarum* WCFS1 compared to those in other lactic acid bacteria. LPSM position in *L. plantarum* is denoted by the gray box. Orthologous genes in different organisms have the same color. Figure generated using the ERGO Bioinformatics suite (19).

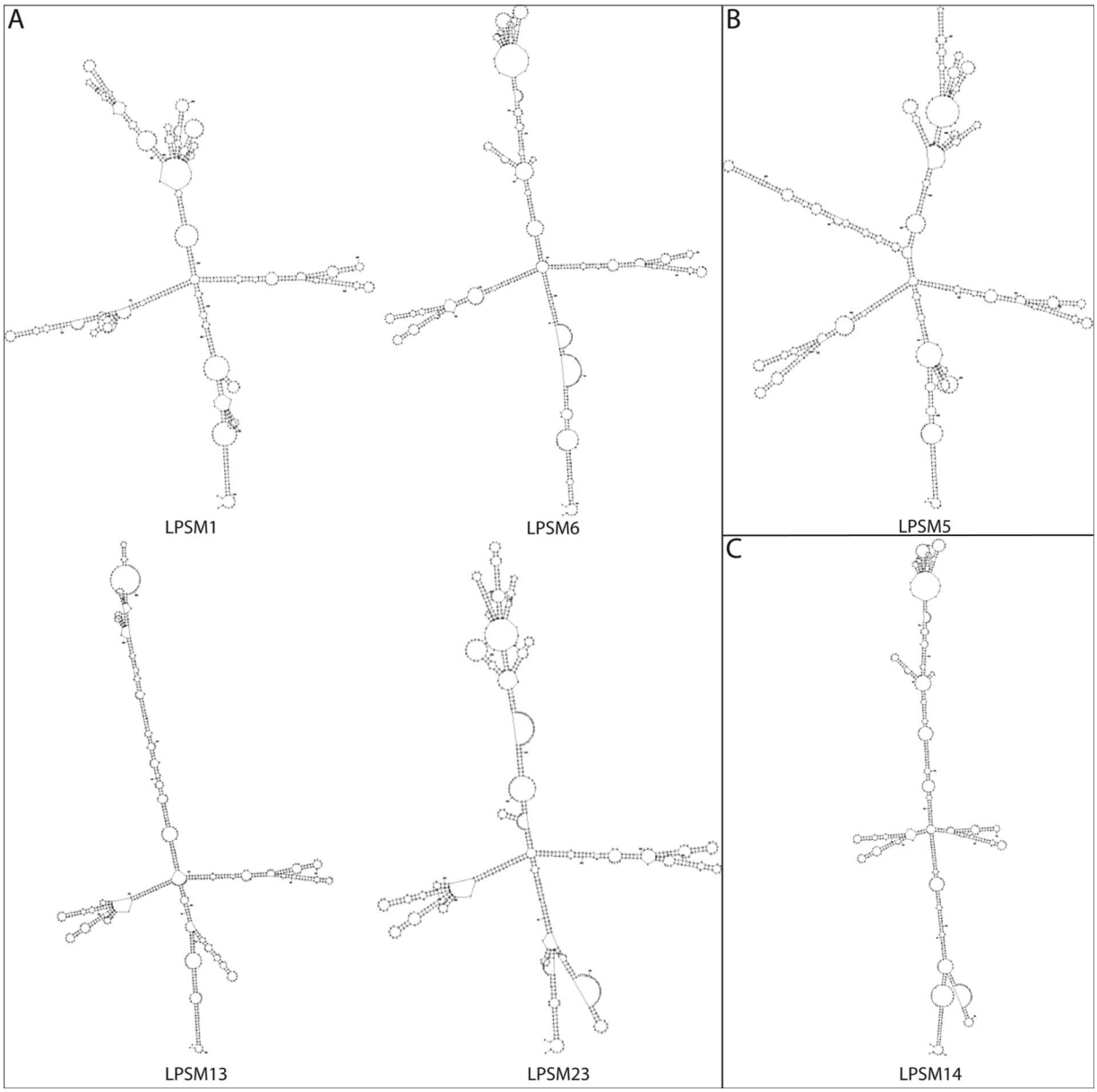

FIG. 3. Exemplary two-dimensional structure predictions for LPSM sequences. (A) The most commonly found motif resembles the structure of a cruciform. (B to C) Other structures clearly resemble the cruciform structure but have certain specific variations; the structure shown in panel B contains an additional hairpin in the "head" region of the LPSM structure, while the structure shown in panel C lacks the "arms" of the LPSM structure.

standard pUC-based forward and reverse sequencing primers. The sequences obtained were aligned with the corresponding *L. plantarum* WCFS1 sequences using Muscle software (7). These sequences displayed a high level (>88%) of sequence identity with the corresponding region found in strain WCFS1. In addition, a maximum likelihood phylogenetic tree (Phyml [9]) of the sequenced LPSMs in combination with the 33 identified *L. plantarum* WCFS1 LPSMs showed clustering for seven out of eight of the sequenced LPSMs with their locus-specific counterparts in *L. plantarum* WCFS1 (see Fig. S1 in the supplemental material), suggesting that the motifs were present in the last common ancestor of the *L. plantarum* strains tested.

The putative secondary structures of the LPSM DNA regions were predicted using the Mfold program (22). The folded structures had free energy values (ΔG) ranging from -93 kJmol^{-1} to -149 kJmol^{-1} and an average value of -132 kJmol^{-1} compared to -73 kJmol^{-1} for scrambled LPSM sequences. Mfold predicted similar secondary structures for 24 out of 33 LPSM sequences, structures that were not predicted from the scrambled sequences. The structures were compared manually and divided into different subfamilies. Nineteen LPSMs displayed conserved and typical secondary structures (Fig. 3A). Notably, these structures share remarkable similarity with DNA structures of eukaryotic origin with unknown functions called cruciforms (12). Of the remaining five LPSM sequences, two (LPSM 5 and LSPM 15) appeared to contain an

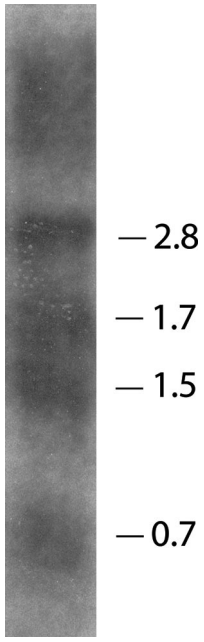

FIG. 4. Northern blot analysis of RNA isolated from *L. plantarum* grown on MRS broth and sampled during exponential growth (optical density at 600 nm [OD_{600}] of 2.5). The LPSM2-specific PCR product was used as a probe. Four transcripts of different sizes (0.7, 1.5, 1.7, and 2.8) were found. The same transcripts were detected at other time points (OD_{600} values of 0.5, 1, and 3.5) during growth (data not shown).

additional hairpin loop in the top of the cruciform structure (Fig. 3B), while three others (LPSM 14, 20, and 33) were characterized by the lack of "arms" (Fig. 3C). In all cases, these aberrations of the general LPSM structure were the result of deletion or insertion in part of the sequence and are clearly apparent from the multiple alignment of all LPSMs (see file S1 in the supplemental material). The predicted common fold of the LPSM sequence elements provided an alternative structure-based method rather than a sequence-based method for searching similar LPSM structures in other bacterial genomes. Therefore, all intergenic regions of lengths larger than 500 nt present in all completely sequenced genomes (i.e., HMM searching for source) were subjected to Mfold structure prediction, and the results were scanned for ΔG levels comparable to those measured for the LPSM structures in *L. plantarum*. No comparable ΔG levels were observed for any of the intergenic regions of the selected species, which corroborates the lack of LPSM-like sequences in other species.

In this study, 33 LPSMs were identified in the intergenic regions of the *L. plantarum* genome. These motifs are highly conserved and show a conserved and distinct secondary structure. The main question is what is the function of these LPSMs? Although more conserved intergenic sequences have been found in different bacteria (5, 15, 16, 21), the function has been described for only a few, like the RNAIII in *S. aureus* and the CRISPR loci (found in, e.g., *Streptococcus thermophilus*). Although the CRISPR-CAS machinery appears to be absent in *L. plantarum*, the LPSMs identified here might be transcribed and be active as a transcript analogous to the CRISPR system. Indeed, Northern blot analysis for RNA isolated from *L. plantarum* WCFS1 shows that an LPSM-specific probe hybridizes to several different RNAs, suggesting that at least one of the LPSM sequences is transcribed into RNA (Fig. 4). Therefore, we hypothesize that the LPSM is functional as an RNA molecule. However, the physiological function of the LPSM transcript(s) remains to be established.

Thanks to Christof Francke for insightful discussions and John van der Oost for critically reading the manuscript.

REFERENCES

1. **Bailey, T. L., and C. Elkan.** 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. **2:**28–36.
2. **Bailey, T. L., and M. Gribskov.** 1998. Combining evidence using p-values: application to sequence homology searches. Bioinformatics **14:**48–54.
3. **Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, and P. Horvath.** 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science **315:**1709–1712.
4. **Brouns, S. J., M. M. Jore, M. Lundgren, E. R. Westra, R. J. Slijkhuis, A. P. Snijders, M. J. Dickman, K. S. Makarova, E. V. Koonin, and J. van der Oost.** 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science **321:**960–964.
5. **Cramton, S. E., N. F. Schnell, F. Gotz, and R. Bruckner.** 2000. Identification of a new repetitive element in *Staphylococcus aureus*. Infect. Immun. **68:** 2344–2348.
6. **Durbin, R., S. Eddy, A. Krogh, and G. Mitchison.** 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, United Kingdom.
7. **Edgar, R. C.** 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5:**113.
8. **Grundy, F. J., and T. M. Henkin.** 2003. The T box and S box transcription termination control systems. Front. Biosci. **8:**d20–31.
9. **Guindon, S., and O. Gascuel.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52:**696–704.
10. **Horvath, P., A. C. Coûté-Monvoisin, D. A. Romero, P. Boyaval, C. Fremaux, and R. Barrangou.** 13 May 2008, posting date. Comparative analysis of CRISPR loci in lactic acid bacteria genomes. Int. J. Food Microbiol. doi: 10.1016/j.ijfoodmicro.2008.05.030.
11. **Kerkhoven, R., F. H. van Enckevort, J. Boekhorst, D. Molenaar, and R. J. Siezen.** 2004. Visualization for genomics: the Microbial Genome Viewer. Bioinformatics **20:**1812–1814.
12. **Kurahashi, H., H. Inagaki, K. Yamada, T. Ohye, M. Taniguchi, B. S. Emanuel, and T. Toda.** 2004. Cruciform DNA structure underlies the etiology for palindrome-mediated human chromosomal translocations. J. Biol. Chem. **279:**35377–35383.
13. **Makarova, K., A. Slesarev, Y. Wolf, A. Sorokin, B. Mirkin, E. Koonin, A. Pavlov, N. Pavlova, V. Karamychev, N. Polouchine, V. Shakhova, I. Grigoriev, Y. Lou, D. Rohksar, S. Lucas, K. Huang, D. M. Goodstein, T. Hawkins, V. Plengvidhya, D. Welker, J. Hughes, Y. Goh, A. Benson, K. Baldwin, J. H. Lee, I. Diaz-Muniz, B. Dosti, V. Smeianov, W. Wechter, R. Barabote, G. Lorca, E. Altermann, R. Barrangou, B. Ganesan, Y. Xie, H. Rawsthorne, D. Tamir, C. Parker, F. Breidt, J. Broadbent, R. Hutkins, D. O'Sullivan, J. Steele, G. Unlu, M. Saier, T. Klaenhammer, P. Richardson, S. Kozyavkin, B. Weimer, and D. Mills.** 2006. Comparative genomics of the lactic acid bacteria. Proc. Natl. Acad. Sci. USA **103:**15611–15616.
14. **Molenaar, D., F. Bringel, F. H. Schuren, W. M. de Vos, R. J. Siezen, and M. Kleerebezem.** 2005. Exploring *Lactobacillus plantarum* genome diversity by using microarrays. J. Bacteriol. **187:**6119–6127.
15. **Moszer, I.** 1998. The complete genome of *Bacillus subtilis*: from sequence annotation to data management and analysis. FEBS Lett. **430:**28–36.
16. **Mrázek, J., L. H. Gaynon, and S. Karlin.** 2002. Frequent oligonucleotide motifs in genomes of three streptococci. Nucleic Acids Res. **30:**4216–4221.
17. **Novick, R. P.** 2003. Autoinduction and signal transduction in the regulation of staphylococcal virulence. Mol. Microbiol. **48:**1429–1449.
18. **Novick, R. P., H. F. Ross, S. J. Projan, J. Kornblum, B. Kreiswirth, and S. Moghazeh.** 1993. Synthesis of staphylococcal virulence factors is controlled by a regulatory RNA molecule. EMBO J. **12:**3967–3975.
19. **Overbeek, R., N. Larsen, T. Walunas, M. D'Souza, G. Pusch, E. Selkov, Jr., K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, A. Bhattacharyya, H. Burd, W. Gardner, P. Hanke, V. Kapatral, N. Mikhailova, O. Vasieva, A. Osterman, V. Vonstein, M. Fonstein, N. Ivanova, and N. Kyrpides.** 2003. The ERGO genome analysis and discovery system. Nucleic Acids Res. **31:**164–171.
20. **Winkler, W. C.** 2005. Riboswitches and the role of noncoding RNAs in bacterial metabolic control. Curr. Opin. Chem. Biol. **9:**594–602.
21. **Woods, S. A., and S. T. Cole.** 1990. A family of dispersed repeats in *Mycobacterium leprae*. Mol. Microbiol. **4:**1745–1751.
22. **Zuker, M.** 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31:**3406–3415.