nature biotechnology

Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*

Alexander Bolotin[1], Benoît Quinquis[1], Pierre Renault[1], Alexei Sorokin[1], S Dusko Ehrlich[1], Saulius Kulakauskas[2], Alla Lapidus[3,5], Eugene Goltsman[3,5], Michael Mazur[3,6], Gordon D Pusch[3,5], Michael Fonstein[3,5], Ross Overbeek[3,5], Nikos Kyprides[3,5], Bénédicte Purnelle[4], Deborah Prozzi[4], Katrina Ngui[4,5], David Masuy[4,6], Frédéric Hancy[4], Sophie Burteau[4,5], Marc Boutry[4], Jean Delcour[4], André Goffeau[4] & Pascal Hols[4]

The lactic acid bacterium *Streptococcus thermophilus* is widely used for the manufacture of yogurt and cheese. This dairy species of major economic importance is phylogenetically close to pathogenic streptococci, raising the possibility that it has a potential for virulence. Here we report the genome sequences of two yogurt strains of *S. thermophilus*. We found a striking level of gene decay (10% pseudogenes) in both microorganisms. Many genes involved in carbon utilization are nonfunctional, in line with the paucity of carbon sources in milk. Notably, most streptococcal virulence-related genes that are not involved in basic cellular processes are either inactivated or absent in the dairy streptococcus. Adaptation to the constant milk environment appears to have resulted in the stabilization of the genome structure. We conclude that *S. thermophilus* has evolved mainly through loss-of-function events that remarkably mirror the environment of the dairy niche resulting in a severely diminished pathogenic potential.

The genus *Streptococcus* comprises several harmful pathogenic species such as *Streptococcus pyogenes* or *Streptococcus pneumoniae*, together with a single 'Generally Recognized As Safe' species, *S. thermophilus*. Assessing the innocuous nature of *S. thermophilus* as a food microorganism is of major importance since this bacterium is widely used for the manufacture of dairy products[1–3] (annual market value of ~$40 billion)[3]. In consequence, over 10^{21} live cells are ingested annually by the human population. The dairy streptococcus must have followed a divergent evolutionary path from that of its pathogenic congeners, as it has adapted to a rather narrow, well-defined and constant ecological niche, milk. To obtain insight into this path and to assess the potential for virulence of this bacterium, we sequenced the genomes of two yogurt strains of *S. thermophilus*, and compared them to those of previously sequenced pathogenic streptococci[4–9].

RESULTS

Divergence of *S. thermophilus* strains

S. thermophilus CNRZ1066 and LMG13811 were isolated from yogurt manufactured in France and in the United Kingdom, respectively. Both strains contain a single circular chromosome of 1.8 Mb, containing about 1,900 coding sequences (**Supplementary Fig. 1** online and **Table 1**). Out of these, about 1,500 (80%) are orthologous (defined as BLASTP reciprocal best hits) to other streptococcal genes, which indicates that *S. thermophilus* and its pathogenic relatives still share a substantial part of their overall physiology and metabolism. The two *S. thermophilus* genomes reported here display about 3,000 single nucleotide differences (0.15% polymorphism). Taking into account the estimated natural mutation rate[10], and assuming a growth rate between one and ten divisions per day, their common ancestor would have lived about 10^7 generations ago, that is, 3,000–30,000 years back, roughly fitting the duration of human dairy activity, believed to have begun about 7,000 years ago[1]. The two genomes differ by 170 single nucleotide shifts, mostly in mononucleotide ($n > 3$) stretches, and 42 regions of sequence differences >50 base pairs (indels) that represent about 4% of genome length (**Supplementary Table 1** online). The two strains have >90% of coding sequences in common (**Table 1**), suggesting a similar lifestyle, as expected from their involvement in the same dairy process. The main differences concern genes for extracellular polysaccharide biosynthesis (*eps, rps*), bacteriocin synthesis and immunity, a remnant prophage and a locus known as 'clustered regularly interspaced short palindromic repeats' (denoted CRISPR2 here; CRISPR1 is present in both strains), closely linked to genes of unknown function (*cas*, **Supplementary Fig. 2** online)[11].

[1]Génétique Microbienne, [2]Unité de Recherche Latière et Génétique Appliquée, Centre de Recherche de Jouy en Josas, Institut National de la Recherche Agronomique, 78352 Jouy en Josas Cedex, France. [3]Integrated Genomics, Chicago, Illinois 60612, USA. [4]Institut des Sciences de la Vie, Université Catholique de Louvain, 1348, Louvain-la-Neuve, Belgium. [5]Present addresses: Microbial Genomics, Department of Energy Joint Genome Institute, 2800 Mitchell Drive, B400, Walnut Creek, California 94598, USA (A.L., E.G. & N.K.), Fellowship for Interpretation of Genomes, 15W155 81st Street, Burr Ridge, Illinois 60527, USA (G.D.P. & R.O.), Cleveland BioLabs, Inc., 10265 Carnegie Ave., Cleveland, Ohio 44106 (M.F.), Department Anatomy and Cell Biology, University of Melbourne, Victoria 3010, Australia (K.N.), Unité de Recherche en Biologie Cellulaire, Facultés Universitaires Notre-Dame de la Paix, 61 Rue de Bruxelles, 5000, Namur, Belgium (S.B.). [6]Deceased. Correspondence should be addressed to S.D.E. (ehrlich@diamant.jouy.inra.fr).

Published online 14 November 2004; doi:10.1038/nbt1034

Inactive *S. thermophilus* genes

Unexpectedly, 10% of the *S. thermophilus* genes are not functional due to frameshift, nonsense mutation, deletion or truncation (globally named pseudogenes). This proportion is the highest among the sequenced streptococcal genomes (**Supplementary Table 2** online). A nearly identical set of pseudogenes is shared between the two strains. Different functional categories are affected to various extents, ranging from ~60% truncated coding sequences for "Other Functions" (atypical conditions, phages, transposons), which mainly include insertion sequences known to be prone to inactivation[5] (**Supplementary Table 2** online), to only 3.5% or even none, for 'Translation' and 'Transcription', respectively (**Table 2**). Remarkably, two of the most highly decaying functional groups, 'Transport Proteins' and 'Energy Metabolism' (~30% truncated coding sequences) relate to carbohydrate degradation, uptake and fermentation. Notably, half of the genes dedicated to sugar uptake, including four of the seven sugar phosphotransferase system (PTS) transporters, are pseudogenes in *S. thermophilus* (**Supplementary Tables 3** and **4** online). To substantiate this finding, we sequenced *ptsG* (glucose), *fruA* (fructose), *bglP* (β-glucoside) and *treP* (trehalose) PTS transporter genes in eight different *S. thermophilus* strains and in four strains of the closely related oral commensal *Streptococcus salivarius* (trehalose PTS was not analyzed in the latter). We found them to be pseudogenes in all *S. thermophilus* strains (with a single exception of the fructose PTS in one strain), whereas they appeared intact in the four *S. salivarius* strains. Inactivation of two other genes involved in carbon metabolism, *butA* (acetoin reductase) and *adhE* (alcohol-acetaldehyde dehydrogenase), in the *S. thermophilus* but not *S. salivarius* strains also took place. Some genes dedicated to carbohydrate uptake may have also been lost, as *S. thermophilus* has only a minor fraction (19–36%) of the genes present in other streptococci (**Supplementary Table 3** online). Conversely, a specific symporter for lactose (the main milk carbohydrate) is present in the *S. thermophilus* genome but absent in other streptococci (**Supplementary Tables 3** and **4** online). Thus, probably because mammals have emerged relatively recently (60 million years ago) in comparison to the remote lactic acid bacteria group (1.5–2 billion years ago)[12], numerous genes encoding proteins dispensable in the milk niche have become pseudogenes, paving the way towards gene loss.

Absence of virulence-related genes in *S. thermophilus*

The availability of the *S. thermophilus* genome sequence allowed us to systematically search the chromosome for potential genetic virulence determinants. The ability to use an extended range of carbohydrates is reported to be important for the virulence of pathogenic streptococci, possibly by allowing maintenance of these bacteria in their ecological niche[5,9]. The observed impairment of this function in *S. thermophilus* is likely to reduce the virulence potential. Antibiotic resistance is another important facet of pathogen virulence. The *S. thermophilus* genome does not contain any obvious antibiotic modification genes such as those found in the *Streptococcus agalactiae* pathogen[8] and it is reported to be sensitive to a wide range of antimicrobial compounds[2]. Many streptococcal virulence-related genes (VRGs) are absent from the *S. thermophilus* genome or are present only as pseudogenes, unless they code for proteins performing basic cellular functions (**Supplementary Tables 5–7** online). Have some of the absent genes been lost from *S. thermophilus*, rather than being acquired by the pathogenic streptococci? Over a quarter of virulence-related genes absent in *S. thermophilus* are present in both *S. pyogenes* and *S. pneumoniae* (25/92, using BLASTP with a cut-off value of 10^{-10}) and almost 40% of these (9/25) are found in regions that are colinear in the two genomes. This suggests that they were present in the strain ancestral to both pathogenic species and presumably *S. thermophilus* and that they were lost from the latter.

Pathogenic streptococci exploit surface-exposed proteins to achieve adhesion to mucosal surfaces and escape host defenses[4]. Among the 28 *S. pneumoniae* virulence-related genes coding surface-exposed proteins, only 4 have orthologs in *S. thermophilus* (**Supplementary Table 7** online). A global analysis of surface proteins revealed a major decay in specialized surface proteins (excluding lipoproteins) with a high proportion of pseudogenes (8/13, **Supplementary Tables 8–10** online). The lipoprotein class, which includes a large number of

Table 1 General features of *S. thermophilus* CNRZ1066 and LMG18311 genomes

	CNRZ1066	LMG18311
Genome size (bp)	1,796,226	1,796,846
G+C content	39%	39%
Ribosomal RNA operons	6	6
Transfer RNAs	67	67
Prophages	1	0
Coding sequences	1,915	1,890
Coding sequences present in both genomes	1,785 (93.2%)	1,785 (94.4%)
Identical coding sequences	663 (34.6%)	663 (35.1%)
Coding sequences with assigned functions	1,372 (71.6%)	1,376 (72.8%)
IS-related regions (hybrids of more than one IS)	55 (9)	54 (7)
IS-associated coding sequences	105	107
Pseudogenes (associated coding sequences)	182 (352, 18.4%)	180 (358, 18.9%)
Coding density	84%	84%
Number of insertion-deletion regions (indels) >50 bp between the two genomes	25	30
Total size of indels (bp)	72,180	71,692
Number of CRISPR loci (number of repeats in each locus)	1 (42)	2 (34 and 5)
Total size of indels <50 bp and single nucleotide differences (bp)	3,923	3,923
Single nucleotide differences (SNPs)	2,905	2,905
Short (1–3 nucleotide) sequence shifts	362	362

IS, insertion sequence.

Table 2 Truncated coding sequences in different functional categories

	CNRZ1066			LMG18311		
		Truncated CDSs			Truncated CDSs	
Functional category	Total CDSs	No.	%	Total CDSs	No.	%
Amino-acid biosynthesis	75	4	5.3	74	4	5.4
Biosynthesis of cofactors, prosthetic groups and carriers	45	8	17.8	45	6	13.3
Cell envelope	93	18	19.4	93	16	17.2
Cellular processes	82	7	8.5	86	7	8.1
Central intermediary metabolism	25	3	12.0	20	1	5.0
Energy metabolism	115	34	29.6	119	39	32.8
Fatty acid and phospholipid metabolism	33	2	6.1	34	2	5.9
Purines, pyrimidines, nucleosides and nucleotides	72	12	16.7	70	12	17.1
Regulatory functions	101	23	22.8	100	23	23.0
Replication, DNA metabolism	93	15	16.1	93	13	14.0
Transcription, RNA metabolism	37	0	0.0	37	0	0.0
Translation, protein metabolism	146	5	3.4	144	5	3.5
Transport and binding proteins	253	76	30.0	246	77	31.3
Other functions (atypical conditions, phages, transposons)	147	90	61.2	138	82	59.4
Hypothetical	413	55	13.3	425	71	16.7
Unknown[a]	185			166		
Total	1,915	352	18.4	1,890	358	18.9

[a]Excluded from analysis. CDSs, coding sequences.

substrate-binding subunits of ABC transporters (16 out of 27–28 predicted lipoproteins) and contains a low number of virulence-related genes (2 out of 27–28), is not massively affected. Globally, the most important virulence determinants that are exposed on the cell surface of pathogenic streptococci are absent or inactivated, such as the pneumococcal surface protein A and C (PspA, PspC), the pneumococcal manganese ABC transporter lipoprotein PsaA, IgA proteases, adhesins and a majority of pneumococcal choline-binding proteins. One homolog to a choline-binding protein (CbpD) was found in each *S. thermophilus* genome (**Supplementary Table 9** online) but neither contains the domain necessary for binding to teichoic acids substituted with phosphorylcholine, in line with the lack of the *lic* gene cluster required for phosphorylcholine metabolism[13] in the *S. thermophilus* genome. The two *S. thermophilus* genomes lack genes coding for sortase-anchored surface proteins[14]; moreover, the single sortase gene itself is a pseudogene. Some of the important virulence determinants in pathogenic streptococci (*S. pyogenes, Streptococcus mutans*)[6,9] are sortase-anchored proteins. Furthermore, sortase mutants of pathogenic Gram-positive bacteria, including streptococci (*S. mutans*, *Streptococcus gordonii*) are attenuated in animal models[15]. In spite of the presence of homologs of the *cps* genes, which are involved in the synthesis of the capsule that is essential for virulence in pathogenic streptococci such as *S. pneumoniae*, the two *S. thermophilus* strains are not encapsulated. Their *cps* homologs, also known as *eps*, are involved in the synthesis of exopolysaccharides, important for the industrial use of *S. thermophilus*, as they confer the desired texture to yogurt[16].

RecQ inhibits symmetrical genome inversions in bacteria

Genome plasticity is another important feature for evolutionary adaptation of pathogens to host defense mechanisms[17], as opposed to genome stability, which is expected to better fit the sedate life style of a dairy bacterium. To estimate genome instability we analyzed symmetrical inversions around the chromosome origin/termination axis, which result from recombination events between the replication forks[18]. X-alignment analysis of pathogenic streptococci versus *S. thermophilus* revealed a much higher score of chromosomal inversions within the *Streptococcus* genus than in pairwise comparisons of closely related *Bacillus* species (see **Supplementary Fig. 3a** and **b** online for two selected comparisons with similar G+C content). What might be the reason for this high inversion frequency? We examined replication and recombination-related genes likely to play a role in recombination between the replication forks, and found that streptococci lack the *recQ* gene whereas *B. subtilis* has it. RecQ helicases are present in most living cells, from bacteria to man, and contribute in several ways to genome stability[19]. We found a negative correlation between the frequency of symmetrical chromosomal inversions and the presence of the *recQ* helicase gene in Gram-positive bacteria (**Supplementary Fig. 3c** online), suggesting that RecQ stabilizes the genome of these bacteria. However, as all streptococci lack RecQ, this protein does not increase the stability of the *S. thermophilus* genome relative to its pathogenic relatives and its X-alignment with other streptococci does not appear more conserved than that between pathogenic streptococci (not shown). It is interesting that a phylogenetically related bacterium used in dairy fermentations, *Lactococcus lactis* (previously *S. lactis*) possesses *recQ*[20]. We noted that pathogenic streptococci, but not *S. thermophilus* and *L. lactis*, lack yet another potential genome-stabilizing function, encoded by the *sbcC* and *sbcD* genes and thought to participate in the repair of recombinogenic double-stranded DNA breaks[21]. These genes are adjacent to a remnant transposase in *S. thermophilus*, suggesting they may have been introduced by lateral gene transfer at a later evolutionary stage to counteract the destabilizing consequences of RecQ deprivation. However, as is often the case with the putative LGT, we cannot rule out the possibility that the genes were originally present in all streptococci and were lost subsequently by deletion from the pathogenic species.

Lateral gene transfer in *S. thermophilus*

In addition to gene decay and loss, lateral gene transfer has contributed to the shaping of the *S. thermophilus* genome. There are >50

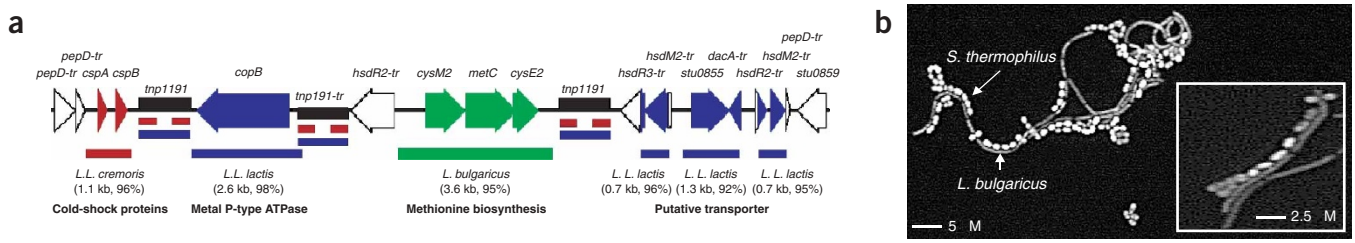

Figure 1 Lateral gene transfer between *S. thermophilus* and dairy bacteria. (**a**) Schematic representation of a 17-kb mosaic region of lateral gene transfer encompassing DNA fragments with more than 90% DNA/DNA identity with *Lactococcus lactis* subsp. *lactis*[20] (*L.L. lactis*, blue), *Lactococcus lactis* subsp. *cremoris* (*L.L. cremoris*, red; Joint Genome Institute, http://www.jgi.doe.gov) and *Lactobacillus bulgaricus* (green; Joint Genome Institute, http://www.jgi.doe.gov). Rectangular boxes in color correspond to exchanged DNA regions; species, DNA fragment size and percentage of DNA identity are indicated below. IS*1191* are shown as black boxes. Extension '*-tr*' indicates genes inactive because of a truncation or one or more frameshifts, *pep*D, endopeptidase; *tnp*, transposase; *hsd*R, restriction endonuclease; *hsd*M, methylase; *dac*A, carboxypeptidase; IS, insertion sequences. (**b**) Adhesion of *S. thermophilus* CNRZ1066 and *L. bulgaricus*. The two organisms were cultivated together in liquid broth to mid-exponential phase; a glass slide was deposited in the culture for 1 h, withdrawn and rinsed five times with water and observed under an optical microscope. Inset, higher magnification.

insertion sequences in the two genomes, some with anomalous G+C content and associated with genes of relevance to milk adaptation. About 75% of insertion sequences are associated with the change in *S. thermophilus* gene order relative to *S. pyogenes*, suggesting that these sequences play an important role in the shaping of the genome. A particularly interesting case of LTG is a 17-kb region found within a truncated *pepD* gene, that is present in both *S. thermophilus* strains. It could be considered as a hot spot of lateral gene transfer, as it contains three of the six insertion sequence *1191* copies present in the LMG18311 strain and constitutes a mosaic of fragments with more than 90% identity to DNA of *Lactobacillus bulgaricus* and two subspecies of *L. lactis* (*lactis* and *cremoris*), three other bacteria also growing in milk (**Fig. 1a**). Interestingly, the leftward flanking region is conserved in two streptococcal species (*Streptococcus equii* and *S. mutans*). Similarly, the rightward flanking region is conserved in *S. equii*, starting about 3.5 kb from the end of the 17 kb region. This conservation supports the hypothesis that insertions took place in the *S. thermophilus* genome. The *L. bulgaricus* fragment (3.6 kb) brings a unique copy of *metC* allowing methionine biosynthesis, a rare amino acid in milk[2]. The high level of identity (95%) of the respective *metC* regions reveals a recent lateral gene transfer event between these two rather distant species used in association in yogurt manufacture[2] and suggests that ecological proximity rather than a phylogenetic one is a prerequisite for lateral gene transfer. We observed that the two species adhere to each other (**Fig. 1b**), which could facilitate gene transfer between them.

DISCUSSION

Comparative genomics leads us to the view that the dairy streptococcus genome may have been shaped mainly through loss-of-function events, even if lateral gene transfer played an important role. This is the first instance where regressive evolution is observed in a food niche rather than in pathogen- or symbiont-host situations[22,23]. The massive gene decay resulted in inactivation and loss of most of the virulence determinants. This provides a strong genomic argument in support of the 'Generally Recognized As Safe' status of the dairy streptococcus, indicating that massive consumption of this bacterium by humans likely entails no health risk.

METHODS

Strains. *S. thermophilus* strains CNRZ1066 and LMG18311 are yogurt isolates, deposited in Institut National de la Recherche Agronomique (INRA) and Laboratorium voor Microbiologie Gent (LMG) collections. Other *Streptococcus* strains used in this study are from the INRA collection: *S. salivarius* JIM 14, 15, 16 and 17; *S. thermophilus* CNRZ 302, 385, 388, 389, 703, 1100,1202 and 1575.

Genome sequencing and assembly. The complete sequences were determined by the random shotgun sequencing strategy followed by multiplex PCR as described earlier[20]. Two sets of random libraries containing 2- to 3-kb inserts were constructed from chromosomal DNA from *S. thermophilus* strains LMG18311 and CNRZ1106. Assembling of 20,000 and 28,000 sequences gave 350 and 300 contigs, respectively, for the two strains. We carried out 1,500 multiplex PCR reactions for final assembling of CNRZ1066 in mixtures of 48 primers, according to the one-step protocol[20], which led to a single circular contig. Subsequently, fragments representing the boundaries of repetitive regions were flagged with respect to partial mismatches at the ends of alignments and were independently assembled before the final sequence polishing. We used a similar finishing strategy for strain LMG18311. In summary, sequences of the two strains were determined by construction of two independent sequence data sets containing 28,000 random and 2,000 primer-directed reads for CNRZ1066 strain and 21,000 random and 1,500 primer-directed reads for LMG18311 strain.

Gene prediction and annotation. A combination of CRITICA[24], Glimmer[25] and an open reading frame calling program developed at Integrated Genomics was used to identify coding sequences. The assembled genomes were analyzed using the ERGO (http://ergo.integratedgenomics.com/IGwit/) bioinformatics suite. The complete DNA sequence and the predicted coding sequences were added into the integrated environment for genome annotation and metabolic reconstruction as described[26]. Protein identifiers (PIDs) sth0001 and stu0001 were assigned to *dnaA* in CNRZ1066 and LMG18311, respectively.

Strain polymorphism. Nucleotide sequences of internal fragments of the genes from different *Streptococcus* strains were determined from PCR products amplified from chromosomal DNAs using selected primers. For each gene fragment the nucleotide sequences were compared and clustered using CLUSTALW program.

Genome comparisons. MUMmer[27] was used for detailed comparative analysis of the two *S. thermophilus* genomes. Comparative genome alignments were based on results of BLASTP Reciprocal Best Hits (RBH)[28], identification of conserved gene order and construction of chromosome gene clusters[29]. The number of ori-symmetrical genome rearrangements[17] was computed using the syntheny groups[30] identified in the RBH genome comparison. We computed the number of inversions by first determining the synthenic regions composed of RBH and then counting the number of regions equidistant from the origin (within 10% tolerance, allowing us to eliminate effects of most insertions and deletions in the compared genomes) but carried on different chromosome arms. Only genomes pairs that have a homology greater than 50% were

selected. The homology was defined as the mean of BLASTP identity of all RBH in the pair of genomes that were compared.

Nucleotide sequence accession number. The *S. thermophilus* genome sequences have been deposited in GenBank with accession no. CP000024 (CNRZ1061) and CP000023 (LMG18311).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The *S. thermophilus* LMG18311 chromosome sequence was supported by funding from the Walloon Region (Bioval no. 981/3866 and First Europe no. EPH3310300R0082) and FNRS (grant no. 2.4586.02). P.H. is Research Associate at FNRS.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 8 July; accepted 21 September 2004
Published online at http://www.nature.com/naturebiotechnology/

1. Fox, P.F. *Cheese: Chemistry, Physics and Microbiology* (Chapman & Hall, London, 1993).
2. Tamine, A.Y. & Deeth, H.C. Yogurt: technology and biochemistry. *J. Food Protection* **43**, 939–977 (1980).
3. Chausson, F. & Maurisson, E. *L'économie Laitière en chiffres* (Centre National Interprofessionnel de l'Economie Laitière, Paris, France, 2002).
4. Mitchell, T.J. The pathogenesis of streptococcal infections: from tooth decay to meningitis. *Nat. Rev. Microbiol.* **1**, 219–230 (2003).
5. Tettelin, H. *et al.* Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293**, 498–450 (2001).
6. Ferretti, J.J. *et al.* Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl. Acad. Sci. USA* **98**, 4658–4663 (2001).
7. Tettelin, H. *et al.* Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proc. Natl. Acad. Sci. USA* **99**, 12391–12396 (2002).
8. Glaser, P. *et al.* Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. *Mol. Microbiol.* **45**, 1499–1513 (2002).
9. Ajdic, D. *et al.* Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. *Proc. Natl. Acad. Sci. USA* **99**, 14434–14439 (2002).
10. Ochman, H., Elwyn, S. & Moran, N.A. Calibrating bacterial evolution. *Proc. Natl. Acad. Sci. USA* **96**, 12638–12643 (1999).
11. Jansen, R., van Embden, J.D., Gaastra, W. & Schouls, L.M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 (2002).
12. Stackebrandt, E. & Teuber, M. Molecular taxonomy and phylogenic position of lactic acid bacteria. *Biochimie* **70**, 317–324 (1988).
13. Zhang, J.-R., Idanpaan-Heikkila, I., Fischer, W. & Tuomanen, E.I. Pneumococcal *licD2* gene is involved in phosphorylcholine metabolism. *Mol. Microbiol.* **31**, 1477–1488 (1999).
14. Comfort, D. & Clubb, R.T. A comparative genome analysis identifies distinct sorting pathways in Gram-positive bacteria. *Infect. Immun.* **72**, 2710–2722 (2004).
15. Paterson, G.K. & Mitchell, T.J. The biology of Gram-positive sortase enzymes. *Trends in Microbiol.* **12**, 89–95 (2004).
16. Broadbent, J.R., McMahon, D.J., Welker, D.L., Oberg, C.J. & Moineau, S. Biochemistry, genetics, and applications of exopolysaccharide production in *Streptococcus thermophilus*: a review. *J. Dairy Sci.* **86**, 407–423 (2003).
17. Dobrint, U. & Hacker, J. Whole genome plasticity in pathogenic genomes. *Curr. Opinion Microbiol.* **4**, 550–557 (2001).
18. Eisen, J.A., Heidelberg, J.F., White, O. & Salzberg, S.L. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* **1**, 1101–1109 (2000).
19. Hickson, I.D. RecQ helicases: caretakers of the genome. *Nat. Rev. Cancer.* **3**, 169–178 (2003).
20. Bolotin, A. *et al.* The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. *Genome Res.* **11**, 731–753 (2001).
21. Bidnenko, V. *et al. sbcB sbcC* null mutations allow RecF-mediated repair of arrested replication forks in *rep recBC* mutants. *Mol. Microbiol.* **33**, 846–857 (1999).
22. Wren, B.W. Microbial genome analysis: insights into virulence, host adaptation and evolution. *Nat. Rev. Genetics,* **1**, 30–39 (2000).
23. Cole, S.T. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011 (2001).
24. Badger, J.H. & Olsen, G.J. CRITICA: Coding Region Identification Tool Invoking Comparative Analysis. *Mol. Biol. Evol.* **16**, 512–524 (1999).
25. Delcher, A.L., Harmon, D., Kasif, S., White, O. & Salzberg, S.L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).
26. Overbeek, R. *et al.* WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28**, 123–125 (2000).
27. Delcher, A.L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* **27**, 2369–2376 (1999).
28. Hirsh, A.E. & Fraser, H.B. Protein dispensability and rate of evolution. *Nature* **411**, 1046–1049 (2001).
29. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96**, 2896–2901 (1999).
30. Huyen, M. & Bork, P. Measuring genome evolution. *Proc. Natl. Acad. Sci. USA* **96**, 5849–5856 (1998).