**Streptococcus thermophilus** Core Genome: Comparative Genome Hybridization Study of 47 Strains

Thomas Bovbjerg Rasmussen,1* Morten Danielsen,2 Ondrej Valina,3 Christel Garrigues,1 Eric Johansen,4 and Martin Bastian Pedersen1

Physiology, Cultures & Enzymes Division,1 Identification, Cultures & Enzymes Division,2 Discovery, Human Health & Nutrition Division,3 and Innovation Management, Cultures & Enzymes Division,4

Chr. Hansen A/S, Hørsholm, Denmark

Received 15 January 2008/Accepted 16 April 2008

A DNA microarray platform based on 2,200 genes from publicly available sequences was designed for *Streptococcus thermophilus*. We determined how single-nucleotide polymorphisms in the 65- to 75-mer oligonucleotide probe sequences affect the hybridization signals. The microarrays were then used for comparative genome hybridization (CGH) of 47 dairy *S. thermophilus* strains. An analysis of the exopolysaccharide genes in each strain confirmed previous findings that this class of genes is indeed highly variable. A phylogenetic tree based on the CGH data showed similar distances for most strains, indicating frequent recombination or gene transfer within *S. thermophilus*. By comparing genome sizes estimated from the microarrays and pulsed-field gel electrophoresis, the proportion of unknown DNA in each strain was estimated. A core genome comprised of 1,271 genes detected in all 47 strains was identified. Likewise, a set of noncore genes detected in only some strains was identified. The concept of an industrial core genome is proposed. This is comprised of the genes in the core genome plus genes that are necessary in an applied industrial context.

The genome includes all genetic material of an organism. It includes a number of core genes, which are detected in all strains of a species, as well as variable genes, which are detected in only some of the strains. The core genes are necessary for overall viability, while the variable genes enable the organism to thrive in particular niches. The properties of a strain are the result of both types of genes.

*Streptococcus thermophilus* is an industrially important species used globally in the production of fermented milk products and cheese. This species has two industrial roles, i.e., the production of lactic acid in cheese and yogurt and the formation of texture and flavor during yogurt production (11, 23). To ensure production of lactic acid in cheese and yogurt and the formation of texture and flavor during yogurt production and to provide product diversity, a large number of different strains are used commercially (23).

Even though *S. thermophilus* is part of the genus *Streptococcus*, which includes several pathogens, it has a generally recognized as safe status in the United States and a qualified presumption of safety status in the European Union due to a long history of safe use in food production.

Comparison of complete genome sequences has revealed the absence of genes associated with pathogenicity in three sequenced strains of *S. thermophilus*, and it is postulated that extensive genome evolution has taken place due to the use of this species in a milk environment for several millennia (4). Here we further explore the gene content and genome evolution of *S. thermophilus* by comparative genome hybridization (CGH) of 47 strains. The microarray platform harvests probes for the genes identified in three published *S. thermophilus* genome sequences combined with additional *S. thermophilus* genes found in GenBank (www.ncbi.nlm.nih.gov). This analysis gives information about the presence or absence of genes in the genomes of the investigated strains, thereby refining the definition of the core genome of this species (13).

**MATERIALS AND METHODS**

**Strains and culture conditions.** All *S. thermophilus* strains were obtained from the Chr. Hansen culture collection (Hørsholm, Denmark). They were confirmed to be *S. thermophilus* by sequence analysis of the 16S rRNA gene (data not shown). Strains were grown without shaking at 37°C in M17 broth (Oxoid A/S, Greve, Denmark) supplemented with 2% lactose.

**Design of oligonucleotides and preparation of microarrays.** The design of oligonucleotides for the *S. thermophilus* microarray platform was carried out using OligoWiz 1.0 (17), with the relevant coding sequences (CDSs) as input and 65- to 75-mer oligonucleotides as output. The platform was intended to cover all publicly available CDSs of *S. thermophilus* (as of August 2005). LGM18311 (4) was arbitrarily chosen as the primary strain, and all CDSs were collected in FASTA format. One copy of the 16S and 23S rRNA genes was included, whereas tRNA genes, which are relatively short, were excluded. CDSs present in CNRZ1066 (4) but not LGM18311 were identified using nucleotide BLAST, with an E value cutoff of 10−50 (1). Subsequently, CDSs present only in the 199 cons of the M9 draft genome sequence (4) were included. Additional entries for *S. thermophilus* CDSs were downloaded from GenBank. CDSs were extracted using FeatureExtract (www.cbs.dtu.dk/services/FeatureExtract), and unique genes were identified. Finally, unique CDSs from phages DT1 (26) and O1205 (21) were downloaded from GenBank.

All CDSs were used as input to OligoWiz, with LGM18311 set as the reference genome. The numbers of input CDSs were as follows: for LGM18311, 1,893; for CNRZ1066, 103; for LMG18311, 156; for GenBank sequences, 219; and for phages, 45. The resulting oligonucleotides were analyzed against the combined list of CDSs by using BLAST. Some oligonucleotides showed multiple matches (a perfect match over the entire 65- to 75-mer oligonucleotide would give an E value of about 10−52). For these, we attempted to redesign the oligonucleotide. When this was not possible, we removed the respective oligonucleotides, i.e., 87 for LGM18311, 13 for CNRZ1066, 34 for LMG18311, 30 for GenBank sequences, and 2 for phages. One oligonucleotide was designed for each of 2,250 genes.
The oligonucleotides are designated as follows, where "#" indicates an integer relating to a gene identifier for the relevant genome or a position: for LMG18311, stl#; for CNRZ1066, stc#; for LMD-9, StelK#; for GenBank sequences, relevant gene identifier; and for phages, gi23455848_# and gi29165635_#).

We designed multiple oligonucleotides for three important S. thermophilus CDSs for use as controls. These are designated with the suffixes -a, -b, -c, etc. The selected CDSs were the urea amidohydrolase (urease) α subunit (five oligonucleotides), pyruvate kinase (five oligonucleotides), and β-galactosidase (eight oligonucleotides). Twenty negative control oligonucleotides were designed from CDSs of the Leuconostoc mesenteroides ATCC 8293 draft genome sequence (14). These oligonucleotides had no hits among the S. thermophilus CDSs even when the E value was set very high (10^-9) in a BLAST comparison and are designated negLM#.

For determining the sensitivity to one or more sequence variations, we designed oligonucleotides with point mutations, insertions, and deletions, all with one of the pyruvate kinase oligonucleotides (stl1196c) as the basis. These control oligonucleotides were named apm-pyk#, ins-pyk#, and del-pyk#, where "#" indicates an integer, 1 to 8, 1 to 4, and 1 to 4, respectively, and designates the number of sequence changes. The designation "apm" stands for "accumulated point mutations," "ins" designates insertions, and "del" designates deletions.

To determine the number of bases needed to obtain a signal, oligonucleotides with increasing numbers of base substitutions in a negative control oligonucleotide were designed. The oligonucleotides for one such series were named delLM0001pyk#, where "#" equals 0, 14, 16, 18, 20, 22, 19/20, or 21/22. LM0001 oligonucleotide was a negative control oligonucleotide from L. mesenteroides, and "#" designates the number of bases from pyk of S. thermophilus that were substituted in the middle of the oligonucleotide for bases of LM0001. The designations "19/20" and "21/22" indicate one point mutation in 20- and 22-base regions, respectively. A similar series named delLM0356gap was designed, where gap is glyceraldehyde-3-dehydrogenase of S. thermophilus and LM0356 is the negative control oligonucleotide.

In total, 2,304 oligonucleotides were designed and purchased from Bioneer Corporation (Daedeok-gu, Daejeon, South Korea). Lyophilized oligonucleotides were provided following 50-nmol-scale synthesis and BioRP purification. The printing, or spotting, of microarrays was carried out as described previously (Pedersen et al., submitted). Normalization was done in Acuity 4.0, using the default parameters (Axon Instruments Inc., Union City, CA). When an array was of poor quality (<70% of the spots/features designated "found" in generating a data set in Acuity), the preparation of the respective array was repeated.

Generation of dendrograms. The log2 values were exported from Acuity to Excel, and genes determined to be present (those with a log2 value of >-2) were assigned a value of 1 and those determined to be absent were assigned a value of 0. The data were then reimported into Acuity and used to generate dendrograms.

Microarray data accession number. The platform specifications for the microarrays are available at the NCBI Gene Expression Omnibus (GEO) under accession number GPL6369.

RESULTS AND DISCUSSION

Design of microarray platform. Our aim was to design an array platform with all publicly available CDSs of S. thermophilus. The following three genome sequences of S. thermophilus were available: S. thermophilus CNRZ1066 (4), LMG18311 (4), and LMD-9 (14). All nonredundant S. thermophilus CDSs in GenBank were also added. In total, probes for more than 2,200 S. thermophilus genes are represented on the microarrays.

In the following, we use the log2 value (log2 ratio), i.e., the log2 of the Cy5 "test strain" hybridization signal (minus background) divided by the Cy3 "reference strain" hybridization signal (minus background). LMG18311 was always the reference strain.

Microarray platform validation. The microarray platform was tested in a self-hybridization with LMG18311, i.e., this strain was used as both the reference and test strains. A plot of the log2 values of the data is presented in Fig. 1. LMG18311 did not hybridize to all probes on the microarrays, as the arrays contain CDSs not present in this strain. The majority (99.4%)
of genes present in the strain produced a log₂ value of between −0.5 and 0.5 (a ratio of 1.4 up or down); a few gave greater differences, but all genes present in LM18311 had log₂ values of between −1 and 1.

Hols et al. (12) identified 3,000 single-nucleotide polymorphisms (SNPs) between LMG18311 and CNRZ1066. We investigated the effect of SNPs on the hybridization signal. The presence of one SNP in the oligonucleotide sequence, as determined from the genome sequences, resulted in log₂ values ranging from approximately 0 to about −1 on the CGH array. In the case of six, seven, or more SNPs in the oligonucleotide sequence, the log₂ value was around −2 (Fig. 2A). When more SNPs were present, the log₂ value dropped even further. The exact effect of a SNP on hybridization is dependent on its position within the oligonucleotide sequence and also on the specific base involved (data not shown).

Comparing the genome sequences of the two strains allows prediction of which oligonucleotides should return a “present” signal from CNRZ1066 and which should return an “absent” signal. From Fig. 2B, it can be seen that absent and present genes (based on the sequenced DNA) were distributed relatively well around a log₂ value of about −2. We therefore designated genes with log₂ values of <−2 as absent from the test genome. In addition, the genes absent in both the test and reference strains yielded no hybridization data. Such genes were assigned a log₂ value of −2, marking them as not present.

We have applied this method to 47 S. thermophilus strains. Genes present in the investigated strains but not on the microarray could naturally not be detected. A single array was done for each strain. While this slightly reduces the accuracy of the results, it allows the inclusion of a greater number of strains. We have previously shown that the general array platform gives highly reliable data using single arrays, i.e., a detection level of 1.5- to 2-fold, up or down, without the use of replicate arrays (10, 18). This is further substantiated by the data presented in Fig. 1.

Phylogenetic tree. To establish the relatedness of the different strains, a phylogenetic tree was generated using a hierarchical clustering algorithm (Pearson centered algorithm; Acuity) (Fig. 3), following conversion of the microarray data to absent (0) and present (1). Some distinct subgroups appear within the tree. Strain S39-20 clusters in a group of its own. Except for the relatively few strains in the subgroups, most strains have similar phylogenetic distances to most other strains.

Delorme et al. (6) concluded that recombination is frequent in Streptococcus salivarius and Streptococcus vestibularis, which are part of the S. salivarius group, which also contains S. thermophilus (8a). The lack of a clear evolutionary path (Fig. 3) indicates that recombination or gene transfer is also frequent in S. thermophilus. This confirms previous observations based on the sequence diversity of exopolysaccharide (EPS) genes (4a) but is in contrast to observations based on multilocus sequence typing results (12).

Interestingly, the strains in the largest subgroup (marked with a dashed square in Fig. 3) all lack the prtS protease gene, important for rapid acidification by S. thermophilus in milk (20). In total, 20 genes that are present in >50% of the other strains are missing from the strains in the prtS-negative group. These include genes for efflux pumps, antimicrobial peptide transporters, ion channels, and response regulators. Likewise, 39 genes are detected in all of the strains of the prtS-negative
Some noncore genes are detected in only a few strains, whereas others are detected in almost all strains. There were 69 noncore genes detected in 45 strains and 233 noncore genes detected in 46 strains (Fig. 4). These 302 genes we designate “conserved genes.” They are probably core genes which have been lost in a few strains (or, in a few cases, genes for which the signal from the respective oligonucleotide is below the threshold due to noise). Similarly, there were between 27 and 58 noncore genes detected in only one to five genomes (Fig. 4). These are likely to be genes which were recently acquired and are designated “recently acquired genes.” There are 183 recently acquired genes in this data set. Finally, the group of noncore genes detected in 6 to 44 genomes we term “variable genes.” Curiously, 14 genes on the array were not detected in any of these strains; these are also likely to belong to the group of “recently acquired genes.”

The distribution of core genes and the various noncore genes within S. thermophilus is illustrated in Fig. 5. The noncore genes consist mainly of genes encoding bacteriocins, ef-
flux/uptake pumps, and proteins involved in EPS biosynthesis and peptide metabolism, phage genes, and phage resistance genes.

The definition of the size of the core genome depends on the number of strains investigated, as illustrated in Fig. 6. When a small number of strains are taken into account, the number of core genes is high but decreases rapidly as genomes are added. When more strains are taken into account, the rate of gene exclusion decreases. In this study, we used 47 strains and ended up with a core genome of 1,271 genes. Since the curve on Fig. 6 has not flattened out, it appears that we have not yet reached the point where the core genome no longer decreases when more strains are included in the calculation. Lefèbure and Stanhope (13) defined the core genome of *S. thermophilus* to be 1,487 genes, based on three sequenced strains. They also defined the core genome of the genus *Streptococcus* to be around 600 genes, indicating that a substantial number of genes are present which give *S. thermophilus* its unique characteristics.

*S. thermophilus*, together with *S. salivarius* and *S. vestibularis*, forms the salivarius group of the viridans group streptococci. These species are closely related (8a), and based on the genome sequences of two *S. thermophilus* strains, it has been suggested that *S. thermophilus* has evolved during the last 3,000 to 30,000 years, perhaps as part of human dairy activities, which began around 7,000 years ago. Several of the acquired genes found in *S. thermophilus* appear to originate from other dairy species, such as *Lactococcus lactis* and *Lactobacillus delbrueckii*, and thus contribute to its adaptation to the milk environment (7).

**Estimation of chromosome size.** The CGH microarrays can be used to determine an approximate chromosome size for the strains. First, the starting position of each gene on the chromosome can be used to estimate the approximate size of each gene. When the gene size is calculated from the start position of one gene to the start position of the next gene, the size includes intergenic/noncoding regions associated with the gene. For genes that do not have an identified position on the main genome sequences, we estimated the size based on the average number of base pairs per *S. thermophilus* gene. From the information presented by Hols et al. (12), we calculated that *S. thermophilus* has 944 bp/gene. Furthermore, when the microarrays were designed, all tRNA and rRNA genes were excluded, except for one 23S and one 16S RNA gene. Bolotin et al. (4) identified 67 tRNAs and 6 rRNA operons in the genomes of CNRZ1066 and LMG13811. To compensate for this, we added 14.2 kb to the values calculated from the microarrays, as tRNA genes have an average size of about 80 bp.

**FIG. 5.** Relative proportions of different gene groups in 47 *S. thermophilus* strains.

**FIG. 6.** Size of the core genome plotted against the number of strains used for determination. Error bars are for 10 different random input sequences of strains.
bp and rRNA genes have an average size of 2.2 kb (1.5 kb for 16S rRNA and 2.9 kb for 23S rRNA). Furthermore, if plasmid DNA is present in the strains, its size has to be excluded from the size calculations.

To get an estimate of how well the method determines chromosome sizes, the values obtained from the CGH data for the three sequenced strains were compared to the published sizes (12, 14) (Table 1). It is evident that the error for LMD-9 is quite large, probably partly because the oligonucleotides were designed before the LMD-9 genome sequence was completed.

Using only the data for LMG18311 and CNRZ1066, we estimate that our CGH-based method underestimates chromosome sizes by about 8 kb (−0.4%). The largest chromosome is some sizes by about 8 kb (Table 1). It is evident that the error for LMD-9 is quite large, probably partly because the oligonucleotides were designed before the LMD-9 genome sequence was completed.

Table 1. Comparison of genome sizes determined with microarrays and sequencing

<table>
<thead>
<tr>
<th>Strain</th>
<th>No. of genes</th>
<th>No. of bp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Published</td>
<td>CGH Difference</td>
</tr>
<tr>
<td>LMG18311</td>
<td>1,890</td>
<td>1,862</td>
</tr>
<tr>
<td>CNRZ1066</td>
<td>1,915</td>
<td>1,850</td>
</tr>
<tr>
<td>LMD-9</td>
<td>1,998</td>
<td>1,854</td>
</tr>
</tbody>
</table>

* Value from CGH data minus the published value (12, 14).

To determine how well the method determines chromosome sizes, the values obtained from the CGH data for the three sequenced strains were compared to the published sizes (12, 14) (Table 1). It is evident that the error for LMD-9 is quite large, probably partly because the oligonucleotides were designed before the LMD-9 genome sequence was completed.

By comparing closely related strains, using the CGH data and PFGE with four different restriction enzymes on gels including the three reference strains sequenced, we found that most strains analyzed in this study do contain the conditional core genome, but that is not necessarily applicable in an industrial context. Certain features are needed for strains to be useful in the dairy industry. The strains must (i) be able to grow in laboratory medium when they are first isolated, (ii) grow in production medium to high density, (iii) grow well in milk, (iv) be stable and maintain their activity when stored, and (v) be resistant to phages. Hence, the industrial minimum genome (presented here) is probably substantially larger than the theoretical minimum genome.

**EPS genes.** The ability of *S. thermophilus* to produce EPS is important for the dairy industry, as it enhances the texture of fermented milk products such as yogurt. *S. thermophilus* EPS consists of heterosaccharide polymers of primarily galactose, glucose, and rhamnose monomers (8). EPS synthesis in *S. thermophilus* involves binding of sugar monomers to a lipid carrier, using amino sugars as precursors. This reaction is performed by a galactose-1-phosphate or glucose-1-phosphate transferase, and subsequent attachment of different monomers is performed by glycosyl transferases. In addition to this, enzymes for polymerization and transmembrane translocation are needed (9, 19).

More than 10 EPS clusters in *S. thermophilus* have been identified and sequenced, indicating a large degree of variability. It was suggested by Broadbent et al. (5) that the organization of structural and regulatory genes within the EPS clusters is modular. All EPS clusters identified so far contain the *deoD-epsABCD* genes at the 5′ end (see reference 5 and references therein). We found that most strains analyzed in this study do contain the *deoD-epsABCD* genes. Seven of the strains are missing *epsE*, encoding a galactose-1-phosphate or glucose-1-phosphate transferase, and two strains are missing *epsA*, encoding a transcriptional regulator. These missing genes could be compensated for by other EPS genes with similar functions.

The *S. thermophilus* microarray platform contains 118 putative genes involved in EPS synthesis. A few of these genes form clusters where the entire cluster is always either absent or present. The EPS clusters of the different strains seem to be almost “random” assemblies of different EPS genes. The relatedness of the different strains based on their EPS gene contents can be seen in Fig. 7.

In accordance with the work of Broadbent et al. (5), we speculate that the different EPS genes are harbored in a few
physical locations in the chromosome (as in the case of the sequenced strains) and that diversification of the clusters proceeds via horizontal gene transfer and recombination of EPS genes.

**Conclusion.** In the present work, we used CGH to identify 1,271 genes belonging to the core genome of the 47 investigated industrial *S. thermophilus* strains. Interestingly, only a few strains cluster phylogenetically, indicating that *S. thermophilus* evolves mainly via recombination with other *S. thermophilus* strains.

The microarrays were also used to estimate the sizes of the chromosomes. The largest chromosome found was about 1,814 kb. In contrast, the smallest chromosome among the 47 investigated strains was 118 kb smaller, containing approximately 135 fewer genes. The size estimates were confirmed by PFGE and could be used to reveal the presence of up to 50 kb of novel DNA in the investigated genomes. The genome size estimates indicate that even the smallest identified genomes are considerably larger than the minimal core genome. We propose that there is a conditional core genome consisting of the core genes plus a subset of genes drawn from a pool of genes encoding essential functions.

CGH provides a detailed picture of the evolution of *S. thermophilus* strains and might give clues to how the strains are constantly evolving. An understanding of the evolution of *S. thermophilus* might be used in the search for new industrial strains or the development of new derivatives with improved technological functions.

**ACKNOWLEDGMENTS**

We thank Henrik Bjørn Nielsen (CBS, DTU) for adding *S. thermophilus* LMG18311 and CNRZ1066 to the homology database of OligoWiz. We also thank Karen Fuglede Appel and Helle Schack Andersen for excellent technical work.

**REFERENCES**

2. Reference deleted.
3. Reference deleted.

**FIG. 7.** Phylogenetic tree based on EPS gene content. The gray squares indicate the presence of various EPS genes.
15. Reference deleted.
16. Reference deleted.
22. Reference deleted.
25. Reference deleted.