Complete genome sequence of the dairy isolate *Streptococcus macedonicus* ACA-DC 198

Papadimitriou K.[1,*], Papandreou N.C.[2], Ferreira S.[3], Supply P.[3,4], Pot B.[4] and Tsakalidou E.[1]

[1]*Laboratory of Dairy Research, Department of Food Science and Technology, Agricultural University of Athens, Iera Odos 75, 118 55 Athens, Greece,*
[2]*Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, Athens 157 01, Greece,*
[3]*Genoscreen, Genomic Platform and R&D, Campus de l'Institut Pasteur, 1 rue du Professeur Calmette, 59000 Lille, France,*
[4]*Institut Pasteur de Lille, Center for Infection and Immunity of Lille (CIIL), F-59019 Lille, France*
[*]*Correspondence to: kpapadimitriou@aua.gr*

Abstract

Within the *Streptococcus* genus, only *Streptococcus thermophilus* is considered to be non-pathogenic due to its adaptation to the milk environment. *Streptococcus macedonicus* is also an intriguing streptococcal species since its most frequent source of isolation to date is fermented foods, mainly of dairy origin. Sequencing of *S. macedonicus* ACA-DC 198 genome was performed using a combination of 454 GS-FLX pyrosequencing and HiSeq 2000 Illumina sequencing. The hybrid assembly between 454 and HiSeq2000 data (>200x coverage) resulted in one continuous genomic scaffold of 2,130,034 bp and a plasmid of 12,728 bp. The genome assembly was validated against a *Nhe*I optical map of the *S. macedonicus* genome. Sequences were annotated with the BaSys and the RAST pipelines and manually curated using Kodon. Final corrections were made based on the quality assessment of the annotation using GenePRIMP. We found 2,192 protein-coding genes on the chromosome, 192 of which were identified as potential pseudogenes, indicating an ongoing genome decay process. This hypothesis is also supported by the approximately 220 kb-smaller genome size of *S. macedonicus* compared to the *S. gallolyticus* genomes, despite the high level of gene synteny between the two species. Such a reductive evolutionary process is common for lactic acid bacteria domesticated to the food environment, which in the case of *S. thermophilus* was also accompanied by the loss of pathogenicity traits. With our *in silico* analysis we attempt to investigate whether *S. macedonicus* shows traits that would support its adaptation to the dairy environment at the genomic level.

Sequencing the genome of *S. macedonicus* ACA-DC 198

- **1st step: shotgun pyrosequencing with 454 GS-FLX titanium (>100 contigs)**
- **2nd step: 3kb paired-end pyrosequencing with 454 GS-FLX titanium (7 scaffolds)**
- **3rd step: gap-closure and polishing with Illumina sequencing using the HiSeq 2000 (1 chromosome and 1 plasmid)**
- **4th step: validation of the overall assembly (>200X coverage) with an *Nhe*I optical map**

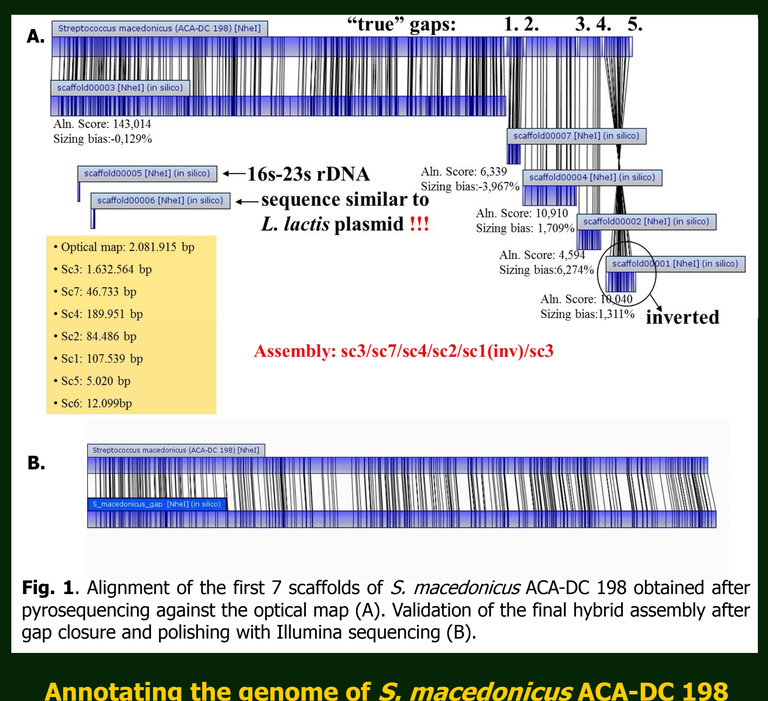

Fig. 1. Alignment of the first 7 scaffolds of *S. macedonicus* ACA-DC 198 obtained after pyrosequencing against the optical map (A). Validation of the final hybrid assembly after gap closure and polishing with Illumina sequencing (B).

Annotating the genome of *S. macedonicus* ACA-DC 198

- **1st step: initial annotation was performed with the BaSys and the RAST pipelines**
- **2nd step: annotations were manually compiled in one using Kodon software**
- **3rd step: final corrections and quality assessment was performed using GenePRIMP (including predictions for potential pseudogenes)**

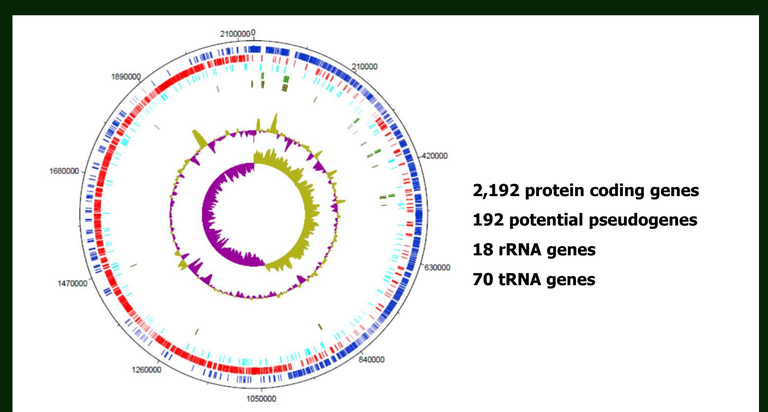

Fig. 2. Genome map of *S. macedonicus* ACA-DC 198

Comparative genomics of *S. macedonicus* ACA-DC 198

- **The complete genome sequence of *S. macedonicus* offered new opportunities to investigate the properties of the species at the genomic scale**
- **The inclusion of *S. macedonicus* and *S. pasteurianus* as subspecies of *S. gallolyticus* has been previously suggested (Schlegel et al. Int J Syst Evol Microbiol. 2003), but this taxonomic reappraisal has not been formally accepted due to low DNA-DNA hybridization relatedness values (<70%) (Whiley et al. Int J Syst Evol Microbiol. 2003)**

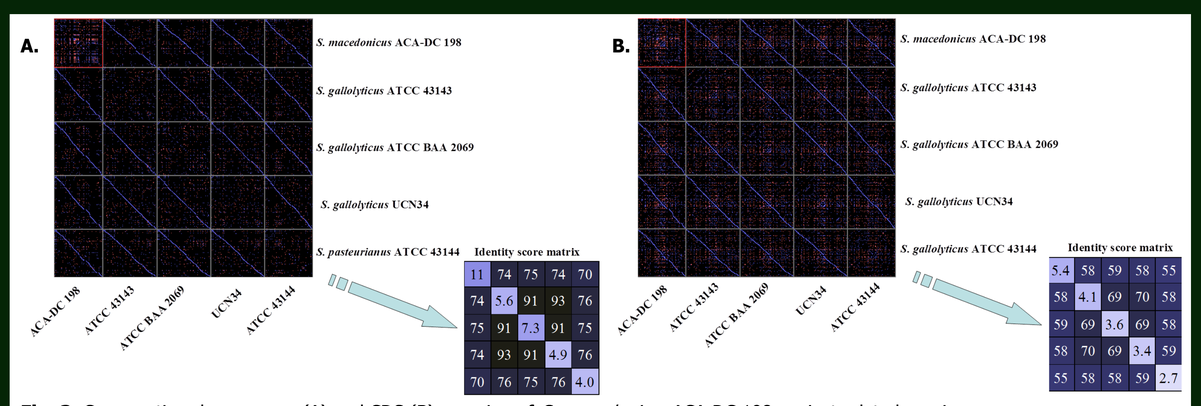

Fig. 3. Comparative chromosome (A) and CDS (B) mapping of *S. macedonicus* ACA-DC 198 against related species

Pairwise alignments of the chromosomes at the nucleotide or the CDS level revealed the degree of synteny between each pair (Fig. 3). The identity score at the nucleotide level of *S. macedonicus* against S. *gallolyticus* and *S. pasteurianus* was around 76% and 70%, respectively. Even more, the identity score at the CDS level dropped radically, reaching 58% in the case of *S. macedonicus* against *S. gallolyticus* and 55% in the case of *S. macedonicus* against *S. pasteurianus*. These values can not be used to directly determine the actual taxonomy of the three species. However, it is a fact that they are quiet low and they coincide with the low (≤ 70%) relatedness values of interspecies DNA–DNA hybridization experiments reported previously, reinforcing the notion that *S. macedonicus* and *S. gallolyticus* should remain separate species.

Acknowledgments

The present work was cofinanced by the European Social Fund and the National resources EPEAEK and YPEPTH through the Thales project.

Fig. 4. Complete alignment of the *S. macedonicus* ACA-DC 198 genome against those of related species using local collinear blocks (LCBs)

Full chromosome alignments were performed using local collinear blocks (LCBs) among the three species. The analysis revealed a mosaic pattern of homology (Fig. 4). Evidently, a significant portion of the genetic information has been overall conserved, since the majority of the LCBs are shared by all species over most of their genome length. It should be noted that numerous strain-specific differences can also be detected. Furthermore, there are LCBs common only among some of the strains, while there are regions divergent enough so as not to be placed within a LCB. These findings indicate that, apart from gene loss through genome decay, gene gain events like lateral gene transfer (LGT) must have played a role during the evolution of the three species. In addition, chromosomal rearrangements seem to have been rather minimal, as the number of "movable" LCBs was low and their length was short. Inclusion of the *S. infantarius* genome in the analysis increased significantly the number of LCBs and reduced drastically the level of conservation among the genomes (data not shown), indicating that this particular genome is fairly different from the rest.

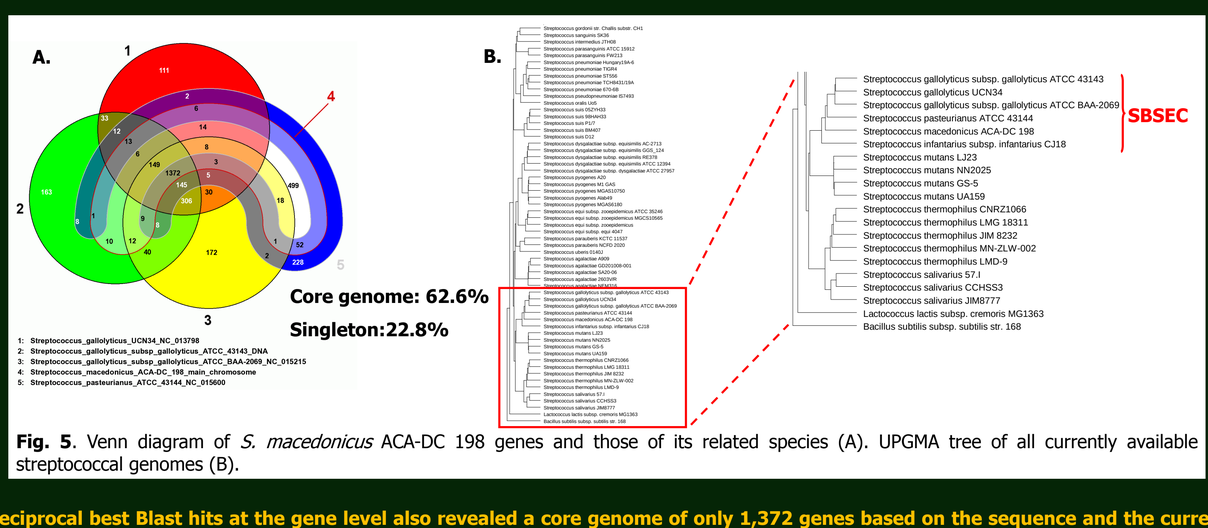

Fig. 5. Venn diagram of *S. macedonicus* ACA-DC 198 genes and those of its related species (A). UPGMA tree of all currently available streptococcal genomes (B).

Reciprocal best Blast hits at the gene level also revealed a core genome of only 1,372 genes based on the sequence and the current annotation of the three species. This allows for a significant percentage of variable genes within the species that must have evolved during the adaptation to their specific environment. Still, *S. macedonicus*, *S. gallolyticus*, *S. pasteurianus* and *S. infantarius* form a single branch in the phylogenetic tree constructed based on the currently available complete streptococcal genome sequences providing extra evidence for the taxonomic integrity of the SBSEC

Additional characteristics of the genomes under investigation

Species	Genome size (Mb)	No. of protein coding genes	No. of potential pseudogenes/ (% percentage)
S. gallolyticus ATCC BAA 2069	2.35	2329	nr*/(nr)
S. gallolyticus ATCC 43143	2.36	2287	41(1.8)
S. gallolyticus UCN34	2.35	2251	28/(1.2)
S. macedonicus ACA-DC 198	2.13	2192	192/(8.7)
S. pasteurianus ATCC 43144	2.10	1869	157/(7.7)
S. infantarius CJ18	1.98	1964	nr/(4.6)

* not reported

- **S. macedonicus, S. pasteurianus and S. infantarius genomes are being shaped by selective pressures that favor extensive gene loss events and genome decay processes when compared to the *S. gallolyticus* genome**
- **This property (i.e. genome decay) has been linked to the adaptation of bacteria to rich in nutrients environments as in the case of *S. thermophilus* adaptation to the milk environment**

Niche-specific and pathogenicity genes presence/absence

S. gallolyticus UCN 34 locus_tag	gene	function	*S. gallolyticus* ATCC BAA 2069	*S. gallolyticus* ATCC 43143	*S. macedonicus* ACA-DC 198	*S. pasteurianus* ATCC 43144	*S. infantarius* CJ18
gallo_0112	fruA	fructan hydrolase	√	√	-	-	-
gallo_0330	-	beta-1,4-endoglucanase (cellulase)	√	√	-	-	-
gallo_0757	-	α-amylase	√	√	-	-	-
gallo_0162	-	mannase	√	√	-	-	pseudo
gallo_0189	-	endo-beta-1,4-galactanase	√	√	pseudo	√	pseudo
gallo_1577	-	pectate lyase	√	√	√	pseudo	√
gallo_1578	-	pectate lyase	√	√	√	pseudo	√
gallo_1632	amyE	α-amylase	√	√	√	pseudo	√
gallo_0933	tanA	tanins degradation	√	√	-	-	-
gallo_1609	similar to tanA	tanins degradation	√	√	-	-	-
gallo_2106	padC	gallic acid decarboxylation	√	√	√	√	-
gallo_0906		gallic acid decarboxylation	√	√	-	√	-
gallo_0818	bsh	bile salt hydrolase	√	√	pseudo	√	√
gallo_2179	-	accessory pilin (pil1)	√	√	-	-	-
gallo_2178	-	major pilin (pil1)	√	√	-	-	-
gallo_2177	-	sortase C (pil1)	√	√	-	-	-
gallo_1570	-	accessory pilin (pil2)	√	√	-	-	-
gallo_1569	-	major pilin (pil2)	√	√	-	-	-
gallo_1568	-	sortase C (pil2)	√	√	-	-	-
gallo_2040	-	accessory pilin (pil3)	√	√	√	√	√
gallo_2039	-	major pilin (pil3)	√	√	√	√	√
gallo_2038	-	sortase C (pil3)	√	√	√	√	√

Our findings clearly suggest that not only *S. macedonicus*, but also *S. pasteurianus* and *S. infantarius* have deviated from *S. gallolyticus* in their potential to catabolize complex plant carbohydrates and to cope with the harsh environment of the GI tract of herbivores. Furthermore, *in silico* analysis of *S. gallolyticus* has revealed that it contains three pilus gene clusters (*pil1*, *pil2*, *pil3*), which may mediate its binding to the extracellular matrix (ECM), although variations of pilus genes presence/absence within strains have also been reported. Each gene cluster consists of three genes. The first two genes encode two adhesins belonging to the MSCRAMM (microbial surface recognizing adhesive matrix molecules) family, one being the major and one being the minor (or accessory) pilus subunit. Pilus attachment to the peptidoglycan, as well as polymerization of adhesin filaments are catalyzed by a sortase C encoded by the third gene of the cluster. *pil1* and *pil2* loci are absent in *S. macedonicus*, *S. pasteurianus* and *S. infantarius* indicating a diminished tendency to adhere to ECM that could probably influence their ability to colonize host tissues and to produce infections when compared to *S. gallolyticus*.

Conclusions

1. *S. macedonicus* is most probably a separate species from *S. gallolyticus*

2. *In silico* analysis of *S. macedonicus* ACA-DC 198 suggests that:

- **The strain is at the process of adapting to a rich in nutrients environment**
- **It shows a diminished capacity to live and survive in the GI tract of herbivores**
- **It has a diminished pathogenic potential compared to *S. gallolyticus***

Bibliography

• Papadimitriou K., S. Ferreira, N. C. Papandreou, E. Mavrogonatou, P. Supply, B. Pot, and E. Tsakalidou (2012) Complete genome sequence of the dairy isolate *Streptococcus macedonicus* ACA-DC 198. J Bacteriol 194:1838-9.