Supplementary material

Identification of Isolated or Mixed Strains from Long Reads:

A Challenge met on *Streptococcus thermophilus* Using a MinION Sequencer

Grégoire Siekaniec [1,2,†], Emeline Roux [1,3], Téo Lemane [1], Eric Guédon [2], and Jacques Nicolas [1,*]

[1] Univ Rennes, INRIA, Campus de Beaulieu 35042 Rennes cedex, France
[2] INRAE, Institut Agro, STLO, F-35000, Rennes, France
[3] CALBINOTOX (Composés ALimentaire BIofonctionnalités et risques NeuTOXiques) EA7488 Université de Lorraine, France; emeline.roux@univ-lorraine.fr

* Correspondence: jacques.nicolas@inria.fr
† mail address: gregoire.siekaniec@inria.fr

Strain	Size (Mb)	GC%	Contig	Status	Accession	WGS	Source	Year	Country	Reference
05-34	1.80032	38.9	43	Scaffold		QFLC01	kefir grains	2010	China	Qin et al., 2011
1F8CT	1.74694	39.1	60	Chromosome	NZ_CM003138	AZTK01	curd from Grana Padano cheese	2012	Italy	Treu et al., 2014c
ACA-DC 2	1.73184	39.2	1	Complete Genome	NZ_LT604076		Greek yogurt		Greece	Alexandraki et al., 2017
APC151	1.83913	39.1	1	Complete Genome	NZ_CP019935		digestive tract of marine fish	2015	Ireland	Linares et al., 2016, 2017
ASCC 1275	1.8455	39.1	1	Complete Genome	NZ_CP006819				Australia	Wu et al., 2014
ATCC 19258	2.0747	38.9	88	Scaffold		PRKU01	culture	2017	Denmark	
B59671	1.82117	39.1	1	Complete Genome	NZ_CP022547		raw milk		USA	Renye et al., 2017
BIO1488	1.74857	39.0	52	Contig		WMLD01				
C106	1.76901	38.9	85	Contig		LGRS01	cheese	2010	Ireland	Wels et al., 2015
CAG:236	1.63196	39.0	66	Scaffold	GCA_000434755.1	CBBT01	distal human gut microbiota	2013		
CNRZ1066	1.79623	39.1	1	Complete Genome	NC_006449		yogurt		France	Bolotin et al., 2004
CS8	1.79166	39.0	1	Complete Genome	NZ_CP016439		rubin	2015	China	
DGCC 7710	1.85121	39.0	1	Complete Genome	CP025216		dairy culture			Hatmaker et al., 2018
EPS	1.81231	39.0	1	Complete Genome	CP025400		milk	2016	China	
FAM 13496	1.83635	38.9	62	Contig		VBTK01	cheese starter culture	1988	Switzerland	
GABA	1.85747	39.1	1	Complete Genome	CP025399		milk	2016	China	
IDCC2201	1.79483	39.2	1	Complete Genome	NZ_CP035306		home-made cheese	2018	South Korea	
JIM 8232	1.9299	38.9	1	Complete Genome	NC_017581		Milk	2002	France	Delorme et al., 2010, 2011
KLDS 3.1003	1.89996	38.9	1	Complete Genome	NZ_CP016877		traditional yogurt culture		Inner Mongolia	Evivie et al., 2017
KLDS3.1012	1.86802	39.2	10	Scaffold		LHSK01	traditional sour milk	2009	Inner Mongolia	Tian et al., 2018
KLDS SM	1.85679	39.1	1	Complete Genome	NZ_CP016026		naturally fermented yogurt		Inner Mongolia	Li et al., 2018; Tian et al., 2018
KU30	1.79076	38.9	43	Contig		QFFS01	milk	2009	China	
LMD-9	1.85637	39.1	3	Complete Genome	NC_008532					Makarova et al., 2006
Plasmid 1	4449 bp	37.0								
Plasmid 2	3361 bp	35.1								
LMG 18311	1.79685	39.1	1	Complete Genome	NC_006448		commercial yogurt	1974	UK	Bolotin et al., 2004
LMG S-29186	1.79829	38.9	42	Contig		NIGT01		2017	Italy	
M17PTZA496	2.06546	38.8	70	Chromosome	NZ_CM002372	AZJT01	Fontina cheese	1996	Italy	Treu et al., 2014a
MN-BM-A01	1.87652	39.1	1	Complete Genome	NZ_CP012588		traditional yogurt block	2012	China	Bai et al., 2016
MN-BM-A02	1.85043	39.0	1	Complete Genome	NZ_CP010999		dairy Fan	2008	China	Shi et al., 2015
MN-ZLW-002	1.84852	39.1	1	Complete Genome	NC_017927		traditional yogurt block		China	Kang et al., 2012
MTCC 5460	1.60936	39.3	143	Contig		ALIK01	fermented milk product (curd)	1984	India	
MTCC 5461	1.6199	39.3	144	Contig		ALIL01	fermented milk product (curd)	1984	India	Prajapati et al., 2013
MTH17CL396	1.82822	38.9	49	Chromosome	NZ_CM002371	AZJS01	Fontina cheese	1996	Italy	Treu et al., 2014a
N4L	1.83175	39.1	1	Complete Genome	NZ_LS974444					Proust et al., 2018
NCTC12958	2.10227	39.0	1	Complete Genome	NZ_LS483339		milk			
ND03	1.83195	39.0	1	Complete Genome	NC_017563		traditional dairy product	2010	China	Sun et al., 2010,2011
ND07	1.86951	39.0	1	Complete Genome	NZ_CP016394		fermented Yak milk	2005	China	
S9	1.78744	39.1	1	Complete Genome	NZ_CP013939		traditional dairy product	2013	China	
SMQ-301	1.86179	39.1	1	Complete Genome	NZ_CP011217		industrial cheese starter	1996		Labrie et al., 2015
ST106	1.85608	39.3	1	Complete Genome	NZ_CP031881		raw milk		USA	Renye et al., 2019
ST109	1.78886	39.2	1	Complete Genome	NZ_CP031545		raw milk		USA	Renye et al., 2019
ST3	1.86506	39.0	1	Complete Genome	NZ_CP017064		commerc. Diet. Supplements		South Korea	
TH1435	1.75346	38.8	36	Chromosome	NZ_CM002369	AYSG01	raw goat milk	2011	Italy	Treu et al., 2014b
TH1436	1.78213	38.8	28	Chromosome	NZ_CM002370	AYTT01	raw goat milk	2011	Italy	Treu et al., 2014b
TH1477	1.88956	38.9	56	Chromosome	NZ_CM003135	AZTJ01	cow milk	2012	Italy	Treu et al., 2014c
TH982	1.78931	38.8	52	Chromosome	NZ_CM003136	AZTL01	buffalo mozzarella curd	2003	Italy	Treu et al., 2014c
TH985	1.84443	39.0	84	Chromosome	NZ_CM003139	AZTM01	buffalo mozzarella whey	2003	Italy	Treu et al., 2014c

Table S1: ***Streptococcus thermophilus*** **genomes extracted from public databases.**

Strain	Short name	Size (Mb)	GC%	Contig	Status	Accession	Source	Year	Country
CIRM-BIA1035	CIRM1035	1.81944	39.1	1	Complete Genome	NZ_LR822029	artisanal lactic starter (Gruyere de Comte)	1973	France
CIRM-BIA1046	CIRM1046	1.88980	39.0	1	Complete Genome	NZ_LR822030	yoghurt	1971	Japan
CIRM-BIA1047	CIRM1047	1.81490	39.0	1	Complete Genome	NZ_LR822031	lactic starter (for yoghurt making)	1978	France
CIRM-BIA1048	CIRM1048	1.86536	39.0	1	Complete Genome	NZ_LR822033	lactic starter (for yoghurt making)	1978	Bulgaria
CIRM-BIA1049	CIRM1049	1.86535	39.0	1	Complete Genome	NZ_LR822034	lactic starter	1978	France
CIRM-BIA1050	CIRM1050	1.84723	39.0	1	Complete Genome	NZ_LR822032	yoghurt (ewe milk)	1973	Greece
CIRM-BIA1051	CIRM1051	1.85757	39.1	1	Complete Genome	NZ_LR822035	yoghurt (Tarag)	1974	Mongolia
CIRM-BIA1055	CIRM1055	1.83916	39.1	1	Complete Genome	NZ_LR822036	yoghurt	1990	France
CIRM-BIA1116	CIRM1116	1.80437	39.1	1	Complete Genome	NZ_LR822039	yoghurt	1963	France
CIRM-BIA1121	CIRM1121	1.78180	39.2	2	Complete Genome	NZ_LR822037	lactic starter (for yoghurt making)	1978	France
Plasmid		3530 bp	32.3						
CIRM-BIA1122	CIRM1122	1.80842	39.0	1	Complete Genome	NZ_LR822041	lactic starter	¡1979	France
CIRM-BIA1125	CIRM1125	1.76728	39.2	1	Complete Genome	NZ_LR822040	yoghurt	1987	Greece
CIRM-BIA1358	CIRM1358	1.80090	39.1	1	Complete Genome	NZ_LR822042	artisanal lactic starter (Gruyere de Comte)	1974	France
CIRM-BIA16	CIRM16	1.80404	39.0	1	Complete Genome	NZ_LR822006	lactic starter (for yoghurt making)	1954	France
CIRM-BIA18	CIRM18	1.86007	39.1	1	Complete Genome	NZ_LR822008	artisanal lactic starter (Gruyere de Comte)	1964	France
CIRM-BIA19	CIRM19	1.82804	39.0	1	Complete Genome	NZ_LR822009	artisanal lactic starter (Gruyere de Comte)	1963	France
CIRM-BIA2101	CIRM2101	1.78894	39.1	1	Complete Genome		commercial lactic starter	1975	Germany
CIRM-BIA23	CIRM23	1.79608	39.0	1	Complete Genome	NZ_LR822011	yoghurt	¡1955	France
CIRM-BIA29	CIRM29	1.80465	39.0	1	Complete Genome	NZ_LR822010	yoghurt	1960	Netherlands
CIRM-BIA30	CIRM30	1.80141	39.1	1	Complete Genome	NZ_LR822012	artisanal lactic starter (Gruyere de Comte)	1963	France
CIRM-BIA32	CIRM32	1.86007	39.1	1	Complete Genome	NZ_LR822013	artisanal lactic starter (Emmental)	1964	France
CIRM-BIA336	CIRM336	1.87117	39.1	2	Complete Genome	NZ_LR822017	milk after contact with traditional sicilian wooden vat (Tina)	2007	Italy
Plasmid		3823 bp	33.4						
CIRM-BIA368	CIRM368	2.02646	38.8	3	Complete Genome	NZ_LR822023	milk after contact with traditional sicilian wooden vat (Tina)	2007	Italy
Plasmid 1		4451 bp	39.4						
Plasmid 2		2162 bp	36.8						
CIRM-BIA36	CIRM36	1.78366	39.1	1	Complete Genome	NZ_LR822014	yoghurt	1971	France
CIRM-BIA65	CIRM65	1.78829	39.1	2	Complete Genome	NZ_LR822015	lactic starter (for yoghurt making)	1971	France
Plasmid		3345 bp	35.5						
CIRM-BIA67	CIRM67	1.77325	39.1	2	Complete Genome	NZ_LR824002	lactic starter (for yoghurt making)	1971	France
Plasmid		3345 bp	35.5						
CIRM-BIA772	CIRM772	1.77249	39.2	1	Complete Genome	NZ_LR822019	fermented milk	1963	France
CIRM-BIA956	CIRM956	1.99901	38.7	3	Complete Genome	NZ_LR822020	cow raw milk	2009	Italy
Plasmid 1		4403 bp	39.4						
Plasmid 2		2162 bp	36.8						
CIRM-BIA961	CIRM961	2.01958	38.9	1	Complete Genome	NZ_LR822025	cow raw milk	2009	Italy
CIRM-BIA967	CIRM967	2.01753	38.9	1	Complete Genome	NZ_LR822026	cow raw milk	2009	Italy
CIRM-BIA998	CIRM998	1.97776	38.9	2	Complete Genome	NZ_LR822027	cow raw milk	2009	Italy
Plasmid		4395 bp	39.3						

Table S2: ***Streptococcus thermophilus*** **genomes sequenced from the CIRM-BIA collection (see Sample Availibility chapter).**

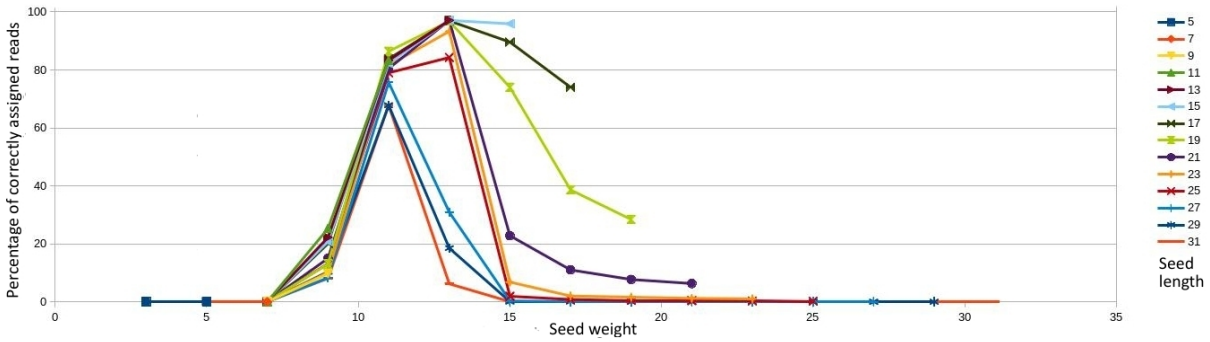

Figure S1: **Identification results of 4000 reads of *S. thermophilus* JIM8232 strain using different seed length and seed weight.** The seed patterns were calculated using the iedera software (33, 34).

Table S3: **Composition of the 180 strains identification experiments.**

Strains proximity	Close		Moderately close		Distant	
Strains number	4	6	4	6	4	6
Experiment 1	CIRM65 CIRM23 CIRM1049 CIRM1048	CIRM65 CIRM36 CIRM23 CIRM67 CIRM32 CIRM18	JIM 8232 CIRM1358 CIRM1047 CIRM1051	CIRM2101 CIRM1047 CIRM65 CIRM1049 CIRM772 CIRM1035	CIRM1047 CIRM368 CIRM336 CIRM956	CIRM1121 CIRM1055 CIRM1047 CIRM1116 CIRM772 CIRM336
Experiment 2	CIRM67 CIRM36 CIRM29 CIRM1122	CIRM18 CIRM32 CIRM23 CIRM67 CIRM2101 CIRM65	CIRM1050 CIRM65 CIRM1035 CIRM1116	CIRM32 CIRM1055 CIRM67 CIRM1125 CIRM1048 CIRM30	CIRM1050 CIRM2101 CIRM30 CIRM1051	CIRM30 CIRM961 JIM 8232 CIRM368 CIRM1125 CIRM772
Experiment 3	CIRM1048 CIRM1049 CIRM1122 CIRM29	CIRM1116 CIRM1122 CIRM961 CIRM967 CIRM67 CIRM2101	CIRM30 CIRM1055 CIRM1051 CIRM67	CIRM19 CIRM2101 CIRM1050 CIRM772 CIRM1116 CIRM23	CIRM1125 CIRM1121 CIRM772 CIRM18	CIRM1050 CIRM1055 CIRM772 CIRM32 CIRM1121 CIRM1116
Experiment 4	CIRM2101 CIRM36 CIRM967 CIRM961	CIRM1116 CIRM29 CIRM65 CIRM2101 CIRM23 CIRM67	CIRM1055 CIRM36 CIRM1046 JIM 8232	CIRM30 CIRM32 CIRM1358 CIRM65 CIRM1116 CIRM2101	CIRM1125 CIRM998 CIRM30 CIRM961	CIRM1121 CIRM19 CIRM1046 CIRM961 CIRM1050 CIRM29
Experiment 5	CIRM1116 CIRM29 CIRM961 CIRM967	CIRM36 CIRM2101 CIRM67 CIRM65 CIRM961 CIRM967	CIRM18 CIRM36 CIRM772 CIRM32	CIRM368 CIRM961 CIRM967 CIRM336 CIRM998 CIRM956	CIRM1046 CIRM961 JIM 8232 CIRM18	CIRM336 CIRM1049 JIM 8232 CIRM32 CIRM772 CIRM998

	Reads number					
Distribution	Uniform			Dominant and subdominant		
4 strains	1000	4000	16000	1000	4000	16000
1	250	1000	4000	500	2000	8000
2	250	1000	4000	250	1000	4000
3	250	1000	4000	125	500	2000
4	250	1000	4000	125	500	2000
6 strains	1000	4000	16000	1000	4000	16000
1	167	667	2667	500	2000	8000
2	167	667	2667	250	1000	4000
3	167	667	2667	125	500	2000
4	167	667	2667	62	250	1000
5	166	666	2666	32	125	500
6	166	666	2666	31	125	500

Table S4: **Hamming distance between *S. thermophilus* strains** for strain proximity experiments (identification of strains more or less close)

Distance x 10^4	close	medium	distant
Average	7.16	7.72	11.94
Mean std	5.21	2.23	3.81
Average Min/Max	0.0/11.6	2.7/10.3	6.4/16.4

Table S5: **Clusters of sibling strains**. Cluster numbers also appear in the heatmap (Figure S2)

Cluster id	Strains
1	05-34 ; ASCC_1275 ; DGCC_7710 ; KLDS_SM ; MN-BM-A02 ; ND07 ; C106
2	CIRM1048 ; CIRM1049
3	CIRM1051 ; MN-BM-A01 ; MN-ZLW-002
4	CIRM1055 ; APC151 ; KLDS3.1012 ; KU30 ; ND03
5	CIRM1116 ; CIRM1122 ; CIRM16 ; CIRM29 ; CNRZ1066 ; CS8 ; EPS ; S9
6	CIRM18 ; CIRM32
7	CIRM2101 ; CIRM23 ; CIRM36 ; CIRM65 ; CIRM67 ; LMG_18311
8	CIRM961 ; CIRM967
9	IDCC2201 ; BIO1488
10	LMD-9 ; LMG_S-29186 ; SMQ-301
11	ATCC_19258 ; NCTC12958
12	MTCC_5460 ; MTCC_5461

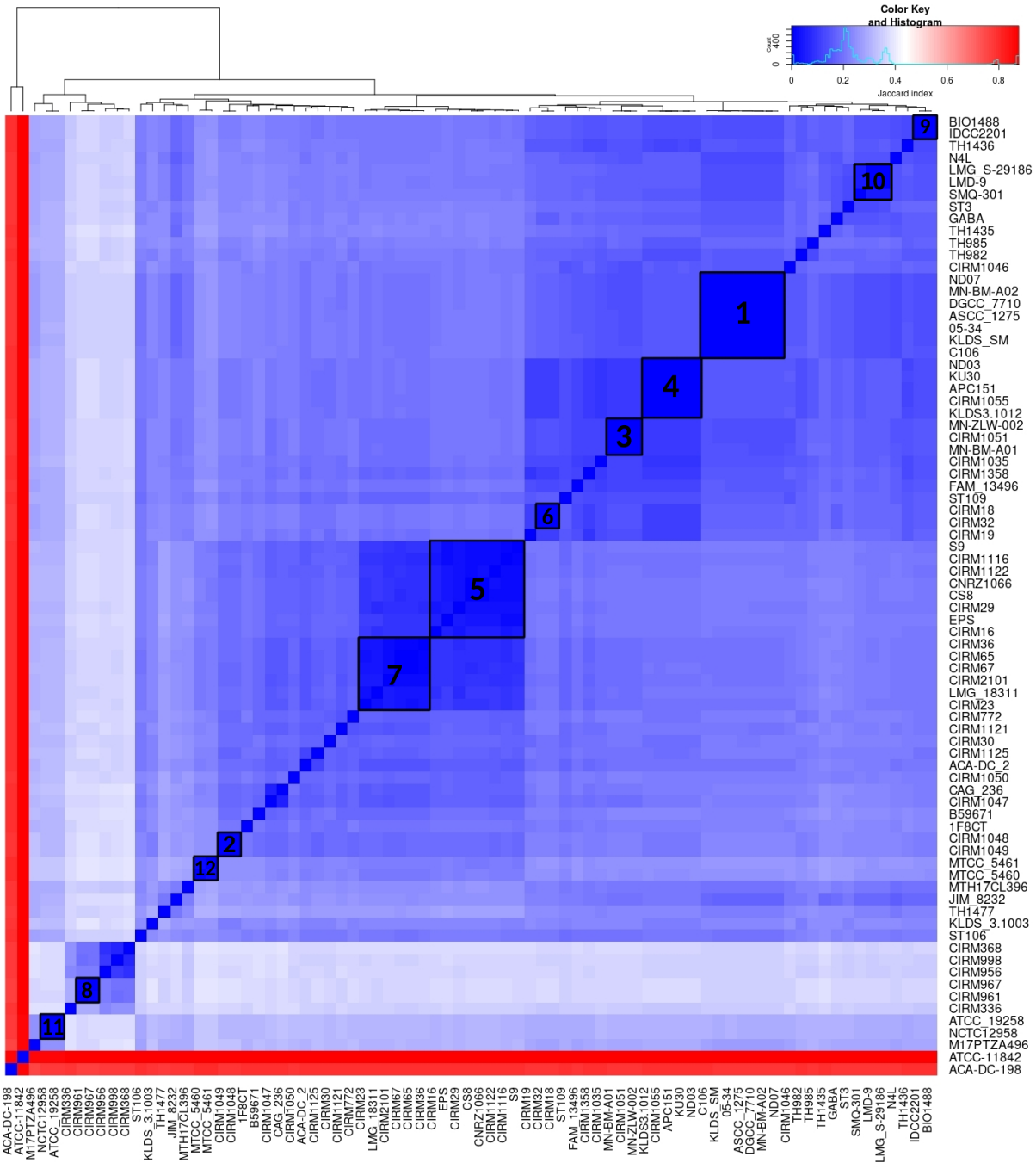

Figure S2: **Heatmap of the Jaccard distance for 77 *S.thermophilus* strains + *S. macedonicus* ACA-DC 198 + *L. delbrueckii* subsp. *bulgaricus* ATCC 11842. The clusters of Table S5 are represented by black squares numbered from 1 to 12**

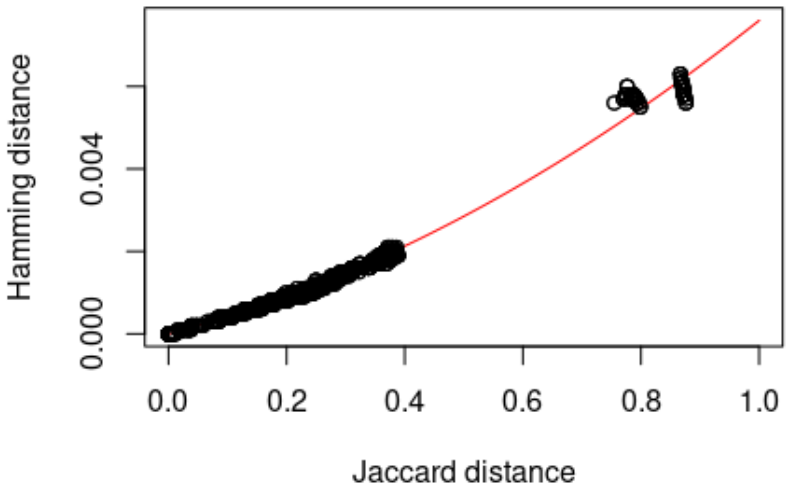

Figure S3: **Correlation between Jaccard and Hamming distances with quadratic regression curve.**

Table S6: **Precision of the *Streptococcus thermophilus* strains identification using ORI with and without clustering of sibling strains for all the experiments.** The clusters are the same as in Table S5. The percentage represents the number of times the strain was correctly identified divided by the number of times it was identified. This corresponds to the precision: TP/(TP+FP) with TP = true positives and FP = false positives.

Cluster id	Strain	Precision without merge (%)	Precision with merge (%)
2	CIRM1048	54.5	100.0
	CIRM1049	81.2	
3	CIRM1051	100.0	100.0
4	CIRM1055	100.0	100.0
5	CIRM29	100.0	91.8
	CIRM1116	100.0	
	CIRM1122	100.0	
6	CIRM18	33.3	100.0
	CIRM32	80.6	
7	CIRM67	50.0	100.0
	CIRM36	60.0	
	CIRM65	77.3	
	CIRM2101	100.0	
	CIRM23	100.0	
8	CIRM967	100.0	100.0
	CIRM961	100.0	
No cluster	CIRM1125	62.5	76.0
	CIRM19	83.3	100.0
	CIRM368	89.5	94.4
	CIRM1050	100.0	100.0
	CIRM30	100.0	100.0
	CIRM1046	100.0	100.0
	JIM_8232	100.0	100.0
	CIRM998	100.0	100.0
	CIRM772	100.0	100.0
	CIRM1121	100.0	100.0
	CIRM1035	100.0	100.0
	CIRM336	100.0	100.0
	CIRM1047	100.0	100.0
	CIRM1358	100.0	100.0
	CIRM956	100.0	100.0

Table S7: **Precision of the *Streptococcus thermophilus* subdominant strains identification using ORI without clustering of sibling strains for experiments with dominant and subdominant strains.** The clusters are the same as in Table S5. The percentage represents the number of times the strain was correctly identified divided by the number of times it was identified. This corresponds to the precision: TP/(TP+FP) with TP = true positives and FP = false positives. Unpredicted means that the strain is not identified in any experiment.

Precision without merge (%)				
Cluster id	**Strain**	**1000 reads**	**4000 reads**	**16000 reads**
2	CIRM1048	0.0	0.0	0.0
	CIRM1049	50.0	100.0	33.33
3	CIRM1051	100.0	100.0	100.0
4	CIRM1055	Unpredicted	Unpredicted	Unpredicted
5	CIRM1116	0.0	0.0	0.0
	CIRM1122	Unpredicted	66.67	66.67
	CIRM29	50.0	50.0	33.33
6	CIRM18	0.0	50.0	0.0
	CIRM32	33.33	0.0	42.86
7	CIRM2101	0.0	0.0	0.0
	CIRM23	Unpredicted	0.0	33.33
	CIRM65	0.0	14.29	0.0
	CIRM67	50.0	Unpredicted	Unpredicted
8	CIRM961	40.0	50.0	55.56
	CIRM967	75.0	100.0	100.0
No cluster	CIRM1125	0.0	0.0	0.0
	CIRM1121	0.0	0.0	0.0
	CIRM1050	0.0	0.0	20.0
	CIRM30	0.0	25.0	25.0
	CIRM1047	0.0	25.0	25.0
	CIRM998	0.0	50.0	66.67
	CIRM956	Unpredicted	Unpredicted	100.0
	CIRM1035	Unpredicted	100.0	100.0
	CIRM368	25.0	33.33	33.33
	JIM_8232	40.0	40.0	40.0
	CIRM1046	50.0	33.33	33.33
	CIRM336	66.67	66.67	75.0
	CIRM772	100.0	50.0	75.0

Table S8: ***S. thermophilus*** **strains identification by ORI, with and without merge index, in a balanced mixture of 4 or 6 strains more or less genetically close, by using 1000, 4000 or 16000 Nanopore sequencing reads.** Best results are in bold. Values of Hamming distance (zero = perfect identification); MCC: Matthews Correlation Coefficient (1 = perfect correlation); ratio: number of strains identified / number of strains present; std: standard deviation.

(a) Heterogeneity results (variable number of strains mixed):

Method		**ORI**		**ORI_merge**	
Number of strains		**4**	**6**	**4**	**6**
Distance	**median**	**0**	**0**	**0**	**0**
	mean	0.73	0.31	**0.53**	**0.29**
	std	2.64	**1.13**	**1.78**	1.19
	[min,max]	[0,15]	**[0,7]**	**[0,7]**	**[0,7]**
MCC	**mean**	0.70	0.65	**0.94**	**0.96**
	std	0.18	0.15	**0.10**	**0.08**
Ambiguity	**mean**	0.65	0.56	**0.93**	**0.96**
	std	0.25	0.21	**0.19**	**0.15**

(b) Data quantity (different number of .fastq reads):

Method		**ORI**			**ORI_merge**		
Number of reads		**1000**	**4000**	**16000**	**1000**	**4000**	**16000**
Distance	**median**	**0**	**0**	**0**	**0**	**0**	**0**
	mean	0.17	0.80	**0.60**	**0**	**0.43**	0.80
	std	0.73	2.93	**1.80**	**0**	**1.43**	2.14
	[min,max]	[0,4]	[0,15]	**[0,7]**	**[0,0]**	**[0,7]**	**[0,7]**
MCC	**mean**	0.55	0.64	0.78	**0.86**	**0.93**	**0.98**
	std	0.20	0.18	0.14	**0.14**	**0.12**	**0.05**
Ambiguity	**mean**	0.44	0.64	0.80	**0.77**	**0.92**	**1.05**
	std	**0.17**	0.20	0.20	0.21	**0.12**	**0.13**

(c) Resolution power (proximity between strains within the mixture):

Method		**ORI**			**ORI_merge**		
Proximity		**distant**	**medium**	**close**	**distant**	**medium**	**close**
Distance	**median**	**0**	**0**	**0**	**0**	**0**	**0**
	mean	0.10	1.17	**0.30**	**0**	**0.90**	0.33
	std	0.54	3.16	**1.27**	**0**	**2.17**	1.35
	[min,max]	[0,3]	[0,15]	**[0,7]**	**[0,0]**	**[0,7]**	**[0,7]**
MCC	**mean**	0.75	0.61	0.6	**0.93**	**0.87**	**0.97**
	std	0.17	0.24	0.14	**0.09**	**0.15**	**0.09**
Ambiguity	**mean**	0.73	0.68	0.47	**0.89**	**0.85**	**1**
	std	0.20	0.23	0.19	**0.15**	**0.22**	**0.18**

Table S9: **Subdominant *S. thermophilus* strains identification by ORI, without/with merge in a mixture of 4 or 6 strains, by using 1000, 4000 or 16000 Nanopore sequencing reads.** Best results are in bold. Values of Hamming distance: in all experiments, minimum value is 0 (perfect identification); MCC: Matthews Correlation Coefficient (1 = perfect correlation); Ambiguity ratio: number of strains identified / number of strains present; std: standard deviation.

Number of strains		**4 (ORI/ORI_merge)**			**6 (ORI/ORI_merge)**		
Number of reads		**1000**	**4000**	**16000**	**1000**	**4000**	**16000**
Distance	**median**	**0/0**	**0/0**	**0/0**	3/**0**	**0/0**	**0/0**
	mean	19.8/19.8	9.9/**0**	**0/0**	30.3/**15.4**	26.7/**0**	**0/0**
	std	24.2/24.2	19.8/**0**	**0/0**	35.9/**29.4**	36.3/**0**	**0/0**
	max	49/49	49/**0**	**0/0**	74.2/74.2	74.2/**0**	**0/0**
MCC	**mean**	0.28/**0.38**	0.42/**0.78**	0.57/**0.9**	0.22/**0.38**	0.34/**0.65**	0.63/**0.8**
	std	0.35/**0.32**	0.35/**0.2**	0.34/**0.21**	**0.28**/0.34	0.28/**0.17**	0.16/**0.14**
Ambiguity	**mean**	0.4/0.4	0.5/**0.8**	0.7/**1**	0.2/**0.33**	0.2/**0.47**	0.53/**0.67**
	std	0.37/0.37	0.32/**0.24**	0.24/**0**	**0.16**/0.21	**0.16**/0.27	**0.16**/0.21

$$J(G1, G2) = \frac{|G1 \cup G2| - |G1 \cap G2|}{|G1 \cup G2|} \tag{1}$$

$$H(G1, G2) = \frac{\text{Number of positions where filters are different}}{\text{Total number of positions}} \tag{2}$$

Equation S1: **Jaccard (1) and Hamming (2) distance equations.** The Jaccard distance is calculated between the qgrams of two genomes (G1 & G2) and the Hamming distance is calculated between two Bloom filters containing these qgrams.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Equation S2: **Matthews correlation coefficient equation.** TP = number of true positives, FP = number of false positives, TN = number of true negatives and FN = number of false negatives.

Text S1: *S. thermophilus* strains cannot be differentiated using 16S-23S markers or MLST

The proximity between genomes makes the taxonomic attribution of bacterial strains difficult, and these noisy long reads make it even more challenging. Indeed, the comparison of their 16S-23S genomic regions showed a maximum of 11 mutations of discrepancies over a length of about 1420 bases (e.g. strains ST106 vs CIRM-BIA967) and no difference in many. The identification of strains of *S. thermophilus* is thus a difficult problem. Indeed, since these strains have very close core genomes, they cannot be differentiated using MLST either (40).

Text S2: Theoretical point of view of the strain identification part

Find a subset of strains that best explain the observed reads may be seen as a set cover optimization problem: find a minimal subset of strains allowing to fully explain the reads. It is one of Karp's NP-complete problems and people usually limit themselves to approximated solution through a statistical approach (see e.g. (41)). It is however possible to search for an exact solution of moderate size (say a few dozens of OTU's in the sample) using modern combinatorial solvers.