

## SUPPLEMENTARY FIGURES AND TABLES

Bruno Contreras-Moreira and Pablo Vinuesa (2013).

GET\_HOMOLOGUES, a versatile software package for scalable and robust microbial pan-genome analysis.

The package, license and supplementary data can be downloaded at:

<http://www.eead.csic.es/compbio/soft/gethoms.php>

and

<http://maya.ccg.unam.mx/soft/gethoms.php>

### LICENSE AND USAGE TERMS

GET\_HOMOLOGUES is designed, created and maintained both at the Laboratory of Computational Biology at Estación Experimental de Aula Dei (EEAD/CSIC, <http://www.eead.csic.es/compbio>) in Zaragoza, Spain and at the Center for Genomic Sciences of Universidad Nacional Autónoma de Mexico (CCG/UNAM, <http://www.ccg.unam.mx/~vinuesa>).

The code has been written by Bruno Contreras-Moreira and Pablo Vinuesa and we kindly ask you to report errors or bugs in the program to the authors ( [bcontreras@eead.csic.es](mailto:bcontreras@eead.csic.es) / [vinuesa@ccg.unam.mx](mailto:vinuesa@ccg.unam.mx) ) and to acknowledge the use of the program in scientific publications:

Contreras-Moreira,B. and Vinuesa,P. (2013) GET\_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis (2013). Appl. Environ. Microbiol. *In press*.

This software is copyright 2013, Bruno Contreras-Moreira and Pablo Vinuesa.

GET\_HOMOLOGUES is made freely available for academic research and shall not be used for any tasks related to military purposes. Commercial users interested in GET\_HOMOLOGUES should contact the authors.

This package includes BLAST, HMMER, COGtriangles, OrthoMCL, mcl, PHYLIP source code/executables, which are distributed with their own licenses. The user of GET\_HOMOLOGUES agrees to respect those as well.

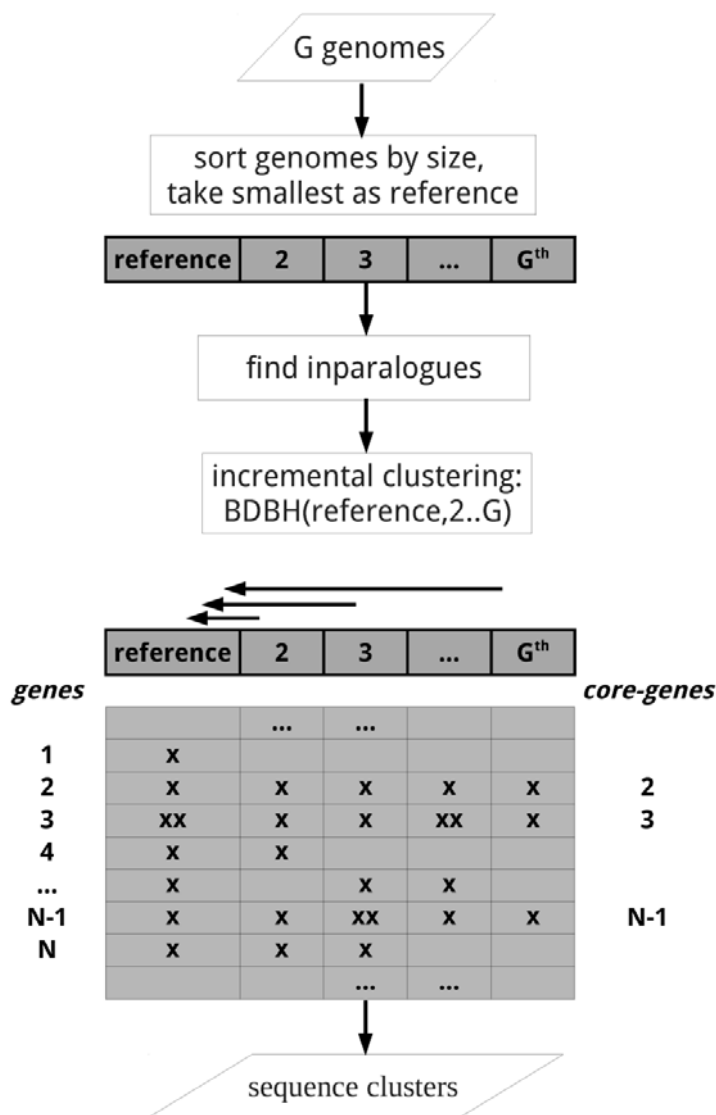
This package is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License provided at the package download link for more details.

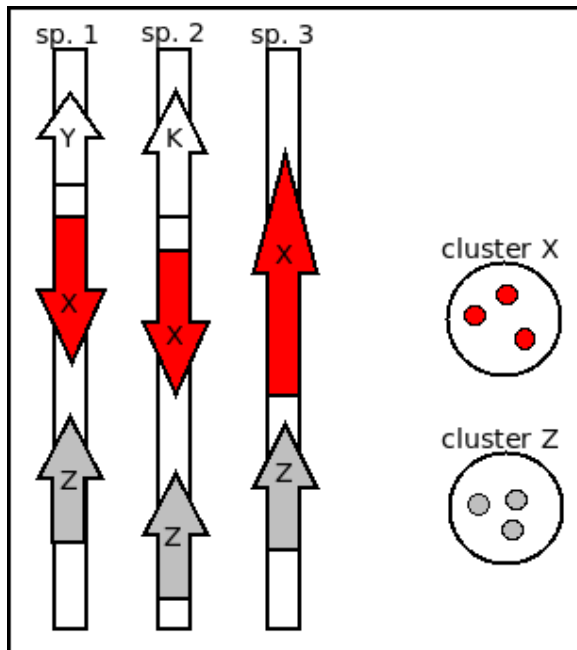
## **ACKNOWLEDGEMENTS**

We thank Romualdo Zayas and Víctor del Moral at CCG-UNAM for their technical support. We also thank Dr. David M. Kristensen and the development team of OrthoMCL for their permission to use their code in our project. Funding: Fundación ARAID, Consejo Superior de Investigaciones Científicas [grant number 200720I038]. DGAPA-PAPIIT UNAM-México [grant number IN212010], CONACyT-México [grant numbers P1-60071 and 179133].

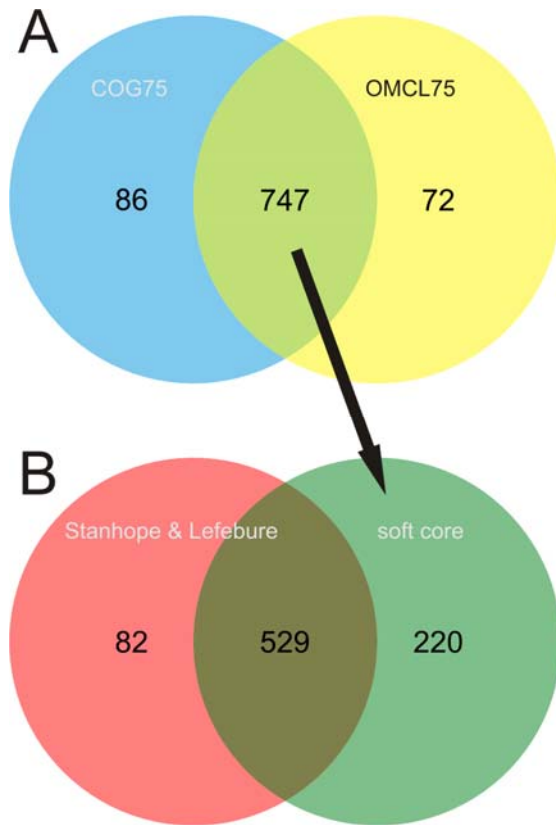
**Figure S1.** Flowchart of the BDBH algorithm as implemented in *get\_homologues.pl* with default parameters and G genomes. First, inparalogues, defined as intra-specific bidirectional best hits (BDBHs), are identified in each genome; second, new genomes are incrementally compared to the reference genome and their BDBHs annotated; finally, clusters containing at least 1 sequence per genome are conserved. Note that the results of the BDBH clustering algorithm (as with COGtriangles and OMCL) can be fine-tuned by customizing diverse parameters, as detailed in the manual and summarized in Fig. 1 of the main text. This figure is taken from the manual.



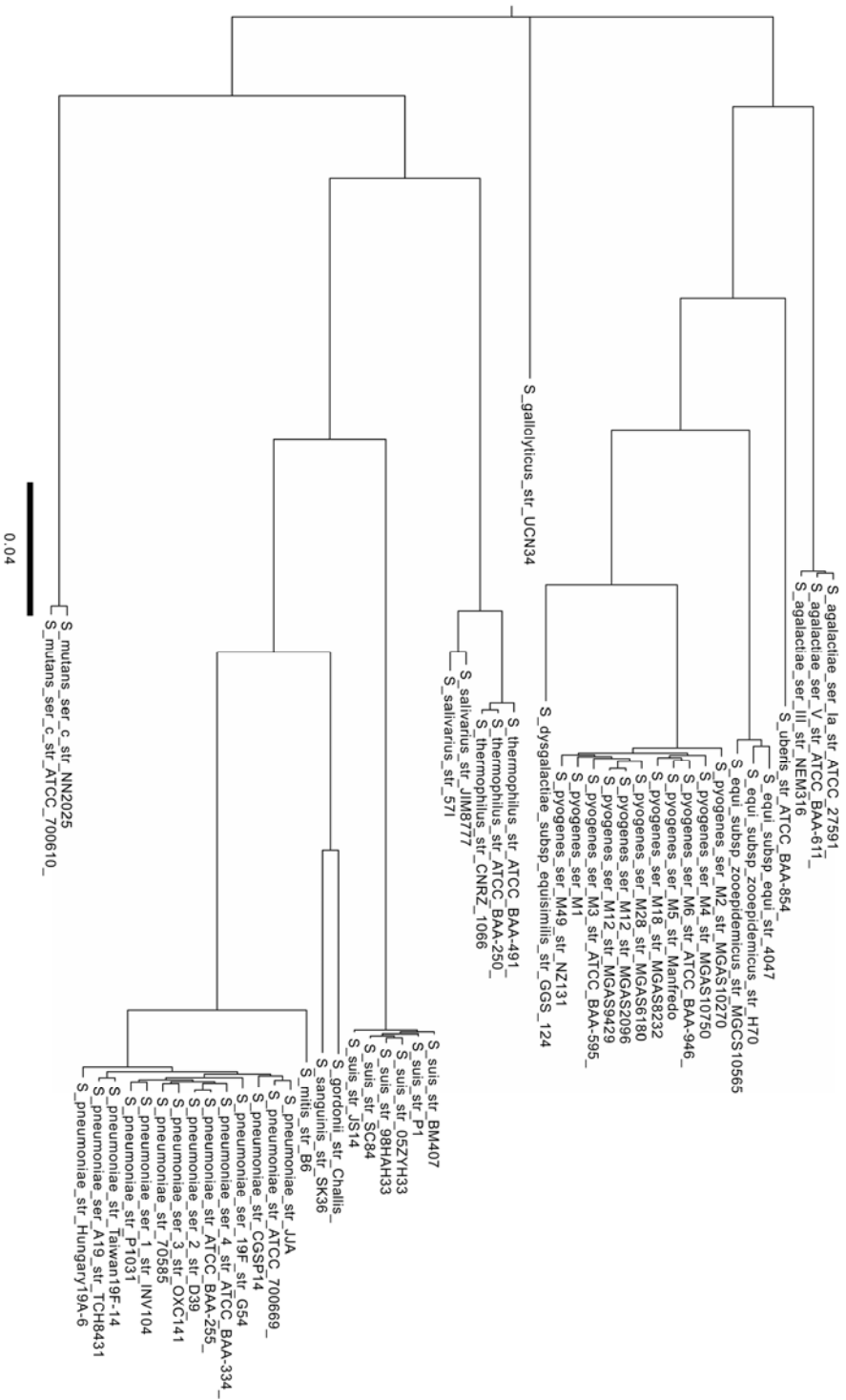
**Figure S2.** Definition: a cluster is called syntenic when it contains neighboring genes which are also contained in other single clusters. In this example gene X and gene Z are found to be syntenic, regardless of their orientation. This figure is taken from the manual.



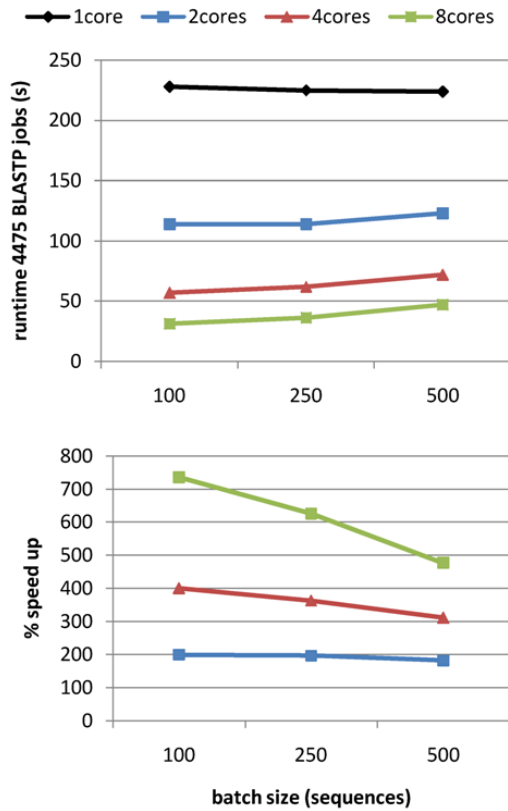
**Figure S3.** A) Overlap of soft-core clusters generated by OMCL and COGtriangles algorithms in GET\_HOMOLOGUES with 75% alignment coverage and 26 *Streptococcus* genomes (from 6 species) first analyzed by Lefébure & Stanhope. B) Venn analysis of the consensus soft-core genes produced by OMCL & COGtriangles and the core gene collection published by Lefébure & Stanhope, comprising 611 genes.



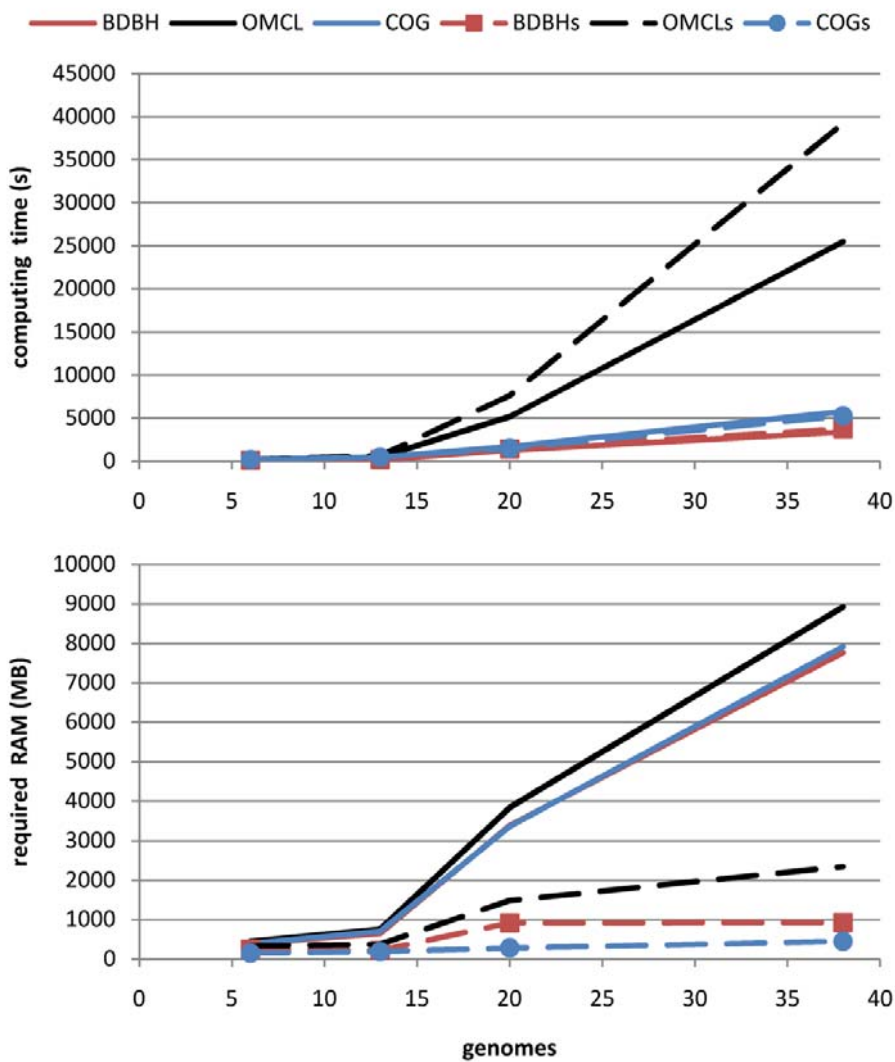
**Figure S4.** Maximum likelihood tree out of 364 concatenated alignments of single-copy consensus core genes of 50 *Streptococcus* species resulting from the intersection of those found by the BDBH, COGtriangles and OMCL algorithms, each run with the Pfam domain-scanning option turned on (-D).



**Figure S5.** Runtime of a typical bacterial genomic BLASTP job as calculated with 1 to 8 CPU cores when splitting the task in batches of 100, 250 and 500 protein sequences. From these analyses, we suggest splitting BLAST (and Pfam HMMER searches, required with parameter `-D`) in batches of 100 sequences when *get\_homologues.pl* is run on a multi-core machine. This figure is taken from the manual.



**Figure S6.** Computing time and RAM requirements of BDBH, COG and OMCL algorithms when processing increasing sets of microbial genomes. Performance is measured also when using the optional BerkeleyDB embedded database (-s parameter), which temporarily stores sequences in files. Broken lines correspond to jobs run with the -s parameter, revealing how computing time is sacrificed in trade of a significantly reduced RAM requirement. This figure is taken from the manual, where further details can be found.





**Figure S7.** For  $G$  genomes, a typical *get\_homologues.pl* job requires running  $G^2$  BLAST searches in order to compare all against all sequences, including against itself to help infer inparalogues. Therefore, the resources required for calculating BLAST jobs grow quadratically. Instead, the BDBH algorithm with option *-b* requires only  $3G$  BLAST searches (in grey) for any reference genome. That is, each genome is blasted only against itself (to detect inparalogs) and against the selected reference genome. This figure is taken from the manual.

	ref	2	3	4	5	6	...	$G^{\text{th}}$
reference								
2								
3								
4								
5								
6								
...								
$G^{\text{th}}$								

**Table T1.** List of 50 *Streptococcus* proteomes downloaded from OMA browser (dec2012 version).

Name	TaxonID	Scientific name	Genome source	Version/Release
STRA1	205921	<i>Streptococcus agalactiae</i> serotype Ia (strain ATCC 27591 / A909 / CDC SS700)	EBI	30-SEP-2005 (Rel. 85, Last updated, Version 1)
STRA3	211110	<i>Streptococcus agalactiae</i> serotype III (strain NEM316)	Genome Reviews	02-MAR-2010 (Rel. 118, Last updated, Version 117)
STRA5	208435	<i>Streptococcus agalactiae</i> serotype V (strain ATCC BAA-611 / 2603 V/R)	NCBI	NC_004116.1 GI:22536185
STRDG	486410	<i>Streptococcus dysgalactiae</i> subsp. equisimilis (strain GGS_124)	Genome Reviews	03-NOV-2009 (Rel. 113, Last updated, Version 3)
STRE4	553482	<i>Streptococcus equi</i> subsp. equi (strain 4047)	EBI	27-MAR-2009 (Rel. 100, Last updated, Version 2)
STRE5	1046629	<i>Streptococcus salivarius</i> (strain 57.I)	EBI	22-SEP-2011 (Rel. 110, Last updated, Version 2)
STRE8	347253	<i>Streptococcus salivarius</i> (strain JIM8777)	EBI	04-NOV-2011 (Rel. 110, Last updated, Version 3)
STREJ	945704	<i>Streptococcus suis</i> (strain JS14)	EBI	01-MAY-2011 (Rel. 108, Last updated, Version 2)
STREM	552526	<i>Streptococcus equi</i> subsp. zoepidemicus (strain MGCS10565)	Genome Reviews	25-NOV-2008 (Rel. 99, Last updated, Version 2)
STRG3	637909	<i>Streptococcus gallolyticus</i> (strain UCN34)	Genome Reviews	15-JUN-2010 (Rel. 122, Last updated, Version 2)
STRGC	467705	<i>Streptococcus gordonii</i> (strain Challis / ATCC 35105 / CH1 / DL1 / V288)	Genome Reviews	26-FEB-2008 (Rel. 87, Last updated, Version 2)
STRM6	365659	<i>Streptococcus mitis</i> (strain B6)	Genome Reviews	08-MAR-2011 (Rel. 129, Last updated, Version 9)
STRMN	511691	<i>Streptococcus mutans</i> serotype c (strain NN2025)	EBI	02-AUG-2009 (Rel. 101, Last updated, Version 1)
STRMU	210007	<i>Streptococcus mutans</i> serotype c (strain ATCC 700610 / UA159)	Genome Reviews	02-MAR-2010 (Rel. 118, Last updated, Version 118)
STRP0	525381	<i>Streptococcus pneumoniae</i> serotype A19 (strain TCH8431)	Genome Reviews	08-MAR-2011 (Rel. 129, Last updated, Version 5)
STRP1	160490	<i>Streptococcus pyogenes</i> serotype M1	Genome Reviews	24-NOV-2009 (Rel. 114, Last updated, Version 114)
STRP2	373153	<i>Streptococcus pneumoniae</i> serotype 2 (strain D39 / NCTC 7466)	EBI	20-DEC-2006 (Rel. 90, Last updated, Version 2)
STRP3	198466	<i>Streptococcus pyogenes</i> serotype M3 (strain ATCC BAA-595 / MGAS315)	Genome Reviews	02-MAR-2010 (Rel. 118, Last updated, Version 118)
STRP4	512566	<i>Streptococcus pneumoniae</i> serotype 19F (strain G54)	Genome Reviews	16-DEC-2008 (Rel. 100, Last updated, Version 2)
STRP6	286636	<i>Streptococcus pyogenes</i> serotype M6 (strain ATCC BAA-946 / MGAS10394)	EBI	17-APR-2005 (Rel. 83, Last updated, Version 2)
STRP7	488221	<i>Streptococcus pneumoniae</i> (strain 70585)	Genome Reviews	25-MAY-2010 (Rel. 121, Last updated, Version 14)
STRP8	186103	<i>Streptococcus pyogenes</i> serotype M18 (strain MGAS8232)	NCBI	NC_003485.1 GI:19745201

STRPB	370553	<i>Streptococcus pyogenes</i> serotype M12 (strain MGAS2096)	EBI	31-MAY-2006 (Rel. 88, Last updated, Version 2)
STRPC	370551	<i>Streptococcus pyogenes</i> serotype M12 (strain MGAS9429)	EBI	31-MAY-2006 (Rel. 88, Last updated, Version 2)
STRPD	370552	<i>Streptococcus pyogenes</i> serotype M2 (strain MGAS10270)	EBI	31-MAY-2006 (Rel. 88, Last updated, Version 2)
STRPF	370554	<i>Streptococcus pyogenes</i> serotype M4 (strain MGAS10750)	EBI	31-MAY-2006 (Rel. 88, Last updated, Version 2)
STRPG	160491	<i>Streptococcus pyogenes</i> serotype M5 (strain Manfredo)	Genome Reviews	17-APR-2007 (Rel. 72, Last updated, Version 3)
STRPI	487214	<i>Streptococcus pneumoniae</i> (strain Hungary19A-6)	Genome Reviews	01-JUL-2008 (Rel. 93, Last updated, Version 3)
STRPJ	561276	<i>Streptococcus pneumoniae</i> (strain ATCC 700669 / Spain 23F-1)	Genome Reviews	24-NOV-2009 (Rel. 114, Last updated, Version 9)
STRPM	319701	<i>Streptococcus pyogenes</i> serotype M28 (strain MGAS6180)	EBI	19-AUG-2005 (Rel. 84, Last updated, Version 2)
STRPN	170187	<i>Streptococcus pneumoniae</i> serotype 4 (strain ATCC BAA-334 / TIGR4)	Genome Reviews	02-MAR-2010 (Rel. 118, Last updated, Version 118)
STRPS	516950	<i>Streptococcus pneumoniae</i> (strain CGSP14)	Genome Reviews	01-JUL-2008 (Rel. 93, Last updated, Version 1)
STRPZ	471876	<i>Streptococcus pyogenes</i> serotype M49 (strain NZ131)	Genome Reviews	16-DEC-2008 (Rel. 100, Last updated, Version 1)
STRR6	171101	<i>Streptococcus pneumoniae</i> (strain ATCC BAA-255 / R6)	NCBI	NC_003098.1 GI:15902044
STRS2	391296	<i>Streptococcus suis</i> (strain 98HAH33)	Genome Reviews	26-JUN-2007 (Rel. 76, Last updated, Version 1)
STRS4	568814	<i>Streptococcus suis</i> (strain BM407)	Genome Reviews	15-JUN-2010 (Rel. 122, Last updated, Version 12)
STRS7	553483	<i>Streptococcus equi</i> subsp. zoepidemicus (strain H70)	EBI	27-MAR-2009 (Rel. 100, Last updated, Version 2)
STRSE	218494	<i>Streptococcus suis</i> (strain P1/7)	Genome Reviews	13-OCT-2009 (Rel. 112, Last updated, Version 2)
STRSV	388919	<i>Streptococcus sanguinis</i> (strain SK36)	Genome Reviews	17-APR-2007 (Rel. 72, Last updated, Version 1)
STRSX	568813	<i>Streptococcus suis</i> (strain SC84)	Genome Reviews	13-OCT-2009 (Rel. 112, Last updated, Version 2)
STRSY	391295	<i>Streptococcus suis</i> (strain 05ZYH33)	Genome Reviews	26-JUN-2007 (Rel. 76, Last updated, Version 1)
STRT1	299768	<i>Streptococcus thermophilus</i> (strain CNRZ 1066)	EBI	17-APR-2005 (Rel. 83, Last updated, Version 4)
STRT2	264199	<i>Streptococcus thermophilus</i> (strain ATCC BAA-250 / LMG 18311)	EBI	17-APR-2005 (Rel. 83, Last updated, Version 5)
STRTD	322159	<i>Streptococcus thermophilus</i> (strain ATCC BAA-491 / LMD-9)	EBI	04-NOV-2006 (Rel. 89, Last updated, Version 3)
STRU0	218495	<i>Streptococcus uberis</i> (strain ATCC BAA-854 / 0140J)	Genome Reviews	24-NOV-2009 (Rel. 114, Last updated, Version 8)
STRZI	869269	<i>Streptococcus pneumoniae</i> serotype 1 (strain INV104)	Genome Reviews	08-MAR-2011 (Rel. 129, Last updated, Version 1)
STRZJ	488222	<i>Streptococcus pneumoniae</i> (strain JJA)	Genome Reviews	07-JUL-2009 (Rel. 108, Last updated, Version 2)

STRZO	869215	<i>Streptococcus pneumoniae</i> serotype 3 (strain OXC141)	Genome Reviews	07-JUN-2011 (Rel. 130, Last updated, Version 2)
STRZP	488223	<i>Streptococcus pneumoniae</i> (strain P1031)	Genome Reviews	16-JUN-2009 (Rel. 107, Last updated, Version 1)
STRZT	487213	<i>Streptococcus pneumoniae</i> (strain Taiwan19F-14)	Genome Reviews	24-NOV-2009 (Rel. 114, Last updated, Version 5)

**Table T2.** Genome list formatted for *download\_genomes\_ncbi.pl* in order to download the corresponding sequences used in the work of Lefebure and Stanhope (2007) Genome Biol 8:R71.

Accession GenomeName	CDSs reported by Lefebure and Stanhope (2007)	CDSs with current GenBank genome annotations, if different
NC_008022 <i>S.pyogenes</i> _MGAS10270	1987	1986
NC_008024 <i>S.pyogenes</i> _MGAS10750	1979	1978
NC_008023 <i>S.pyogenes</i> _MGAS2096	1898	equal
NC_008021 <i>S.pyogenes</i> _MGAS9429	1877	equal
NC_002737 <i>S.pyogenes</i> _M1GAS	1697	1696
NC_007297 <i>S.pyogenes</i> _MGAS5005	1865	equal
NC_003485 <i>S.pyogenes</i> _MGAS8232	1845	1839
NC_007296 <i>S.pyogenes</i> _MGAS6180	1894	equal
NC_004070 <i>S.pyogenes</i> _MGAS315	1865	equal
NC_004606 <i>S.pyogenes</i> _SSI-1	1861	1859
NC_006086 <i>S.pyogenes</i> _MGAS10394	1886	equal
NC_003098 <i>S.pneumoniae</i> _R6	2043	2042
NC_003028 <i>S.pneumoniae</i> _TIGR4	2094	2105
NC_004350 <i>S.mutans</i> _UA159	1960	equal
NC_004116 <i>S.agalactiae</i> _2603VR	2124	equal
NC_007432 <i>S.agalactiae</i> _A909	1996	equal
NC_004368 <i>S.agalactiae</i> _NEM316	2094	equal
AAJP <i>S.agalactiae</i> _515	2275	equal
AAJQ <i>S.agalactiae</i> _CJB111	2197	equal

AAJR <i>S.agalactiae</i> _COH1	2376	equal
AAJS <i>S.agalactiae</i> _H36B	2376	equal
AAJO <i>S.agalactiae</i> _18RS21	2146	equal
AAFA <i>S.suis</i> _891591	1896	2119
NC_006449 <i>S.thermophilus</i> _CNRZ1066	1915	equal
NC_006448 <i>S.thermophilus</i> _LMG 18311	1889	1888
NC_008532 <i>S.thermophilus</i> _LMD-9	1835	1709

**Table T3.** Consensus core-genes of 50 *Streptococcus* genomes produced by all BDBH, COGtriangles and OMCL with 75% alignment coverage. Gene identifiers correspond to *Streptococcus pyogenes* serotype M1, release 114, as downloaded from OMAbrowser. A file with the corresponding full sequence clusters is available from <http://www.eead.csic.es/compbio/soft/gethoms.php>.

STRP100001,STRP100005,STRP100009,STRP100010,STRP100011,STRP100014,STRP100015,STRP100016,STRP100017,STRP100018,STRP100019,STRP100020,STRP100021,STRP100024,STRP100025,STRP100026,STRP100030,STRP100031,STRP100032,STRP100042,STRP100043,STRP100044,STRP100045,STRP100046,STRP100047,STRP100048,STRP100049,STRP100051,STRP100052,STRP100053,STRP100054,STRP100055,STRP100056,STRP100058,STRP100061,STRP100063,STRP100066,STRP100067,STRP100070,STRP100072,STRP100073,STRP100074,STRP100079,STRP100081,STRP100084,STRP100085,STRP100089,STRP100090,STRP100119,STRP100128,STRP100140,STRP100152,STRP100159,STRP100163,STRP100164,STRP100179,STRP100191,STRP100192,STRP100195,STRP100196,STRP100197,STRP100200,STRP100201,STRP100202,STRP100207,STRP100208,STRP100210,STRP100211,STRP100212,STRP100213,STRP100214,STRP100221,STRP100223,STRP100224,STRP100225,STRP100226,STRP100227,STRP100230,STRP100233,STRP100234,STRP100237,STRP100239,STRP100242,STRP100244,STRP100245,STRP100246,STRP100247,STRP100249,STRP100250,STRP100253,STRP100258,STRP100259,STRP100260,STRP100262,STRP100263,STRP100264,STRP100265,STRP100266,STRP100267,STRP100270,STRP100274,STRP100275,STRP100276,STRP100283,STRP100289,STRP100291,STRP100295,STRP100300,STRP100303,STRP100316,STRP100317,STRP100319,STRP100326,STRP100328,STRP100330,STRP100331,STRP100332,STRP100334,STRP100338,STRP100339,STRP100341,STRP100352,STRP100353,STRP100357,STRP100358,STRP100359,STRP100368,STRP100370,STRP100371,STRP100379,STRP100380,STRP100381,STRP100382,STRP100408,STRP100415,STRP100421,STRP100422,STRP100426,STRP100427,STRP100429,STRP100431,STRP100432,STRP100439,STRP100441,STRP100442,STRP100445,STRP100448,STRP100464,STRP100465,STRP100468,STRP100469,STRP100470,STRP100471,STRP100472,STRP100520,STRP100523,STRP100524,STRP100525,STRP100526,STRP100528,STRP100530,STRP100531,STRP100533,STRP100547,STRP100548,STRP100550,STRP100552,STRP100553,STRP100554,STRP100556,STRP100557,STRP100559,STRP100562,STRP100563,STRP100569,STRP100574,STRP100575,STRP100576,STRP100592,STRP100593,STRP100594,STRP100595,STRP100596,STRP100598,STRP100599,STRP100600,STRP100601,STRP100605,STRP100606,STRP100610,STRP100611,STRP100614,STRP100616,STRP100617,STRP100618,STRP100623,STRP100624,STRP100631,STRP100642,STRP100643,

STRP100649,STRP100650,STRP100651,STRP100652,STRP100654,STRP100655,STRP100656,STRP100658,STRP100662,STRP100663,  
STRP100665,STRP100670,STRP100671,STRP100675,STRP100676,STRP100677,STRP100689,STRP100690,STRP100691,STRP100692,  
STRP100693,STRP100695,STRP100696,STRP100698,STRP100759,STRP100761,STRP100762,STRP100763,STRP100767,STRP100777,  
STRP100779,STRP100781,STRP100783,STRP100785,STRP100806,STRP100810,STRP100830,STRP100831,STRP100833,STRP100846,  
STRP100847,STRP100848,STRP100852,STRP100853,STRP100861,STRP100862,STRP100864,STRP100865,STRP100868,STRP100869,  
STRP100870,STRP100874,STRP100881,STRP100883,STRP100884,STRP100886,STRP100909,STRP100911,STRP100912,STRP100913,  
STRP100915,STRP100923,STRP100931,STRP100935,STRP100936,STRP100937,STRP100940,STRP100941,STRP100942,STRP100948,  
STRP100949,STRP100955,STRP100956,STRP100971,STRP100979,STRP100983,STRP101005,STRP101010,STRP101019,STRP101028,  
STRP101030,STRP101035,STRP101037,STRP101038,STRP101040,STRP101050,STRP101051,STRP101054,STRP101059,STRP101061,  
STRP101063,STRP101064,STRP101065,STRP101066,STRP101072,STRP101073,STRP101076,STRP101083,STRP101084,STRP101088,  
STRP101089,STRP101090,STRP101143,STRP101146,STRP101148,STRP101149,STRP101150,STRP101151,STRP101159,STRP101160,  
STRP101161,STRP101164,STRP101165,STRP101169,STRP101170,STRP101174,STRP101181,STRP101182,STRP101206,STRP101218,  
STRP101236,STRP101238,STRP101241,STRP101242,STRP101243,STRP101246,STRP101248,STRP101249,STRP101250,STRP101253,  
STRP101254,STRP101261,STRP101263,STRP101265,STRP101266,STRP101267,STRP101268,STRP101269,STRP101270,STRP101275,  
STRP101276,STRP101277,STRP101278,STRP101279,STRP101280,STRP101293,STRP101294,STRP101295,STRP101298,STRP101303,  
STRP101316,STRP101317,STRP101318,STRP101319,STRP101320,STRP101322,STRP101323,STRP101324,STRP101326,STRP101328,  
STRP101330,STRP101331,STRP101335,STRP101337,STRP101338,STRP101341,STRP101343,STRP101344,STRP101345,STRP101347,  
STRP101348,STRP101350,STRP101351,STRP101352,STRP101354,STRP101360,STRP101361,STRP101362,STRP101364,STRP101366,  
STRP101371,STRP101383,STRP101384,STRP101385,STRP101390,STRP101391,STRP101394,STRP101396,STRP101398,STRP101400,  
STRP101401,STRP101402,STRP101403,STRP101405,STRP101409,STRP101410,STRP101412,STRP101413,STRP101416,STRP101417,  
STRP101419,STRP101421,STRP101424,STRP101426,STRP101429,STRP101436,STRP101439,STRP101440,STRP101441,STRP101444,  
STRP101447,STRP101449,STRP101452,STRP101456,STRP101458,STRP101459,STRP101486,STRP101488,STRP101491,STRP101501,  
STRP101504,STRP101505,STRP101507,STRP101508,STRP101517,STRP101522,STRP101524,STRP101527,STRP101573,STRP101575,  
STRP101581,STRP101595,STRP101596,STRP101597,STRP101603,STRP101604,STRP101606,STRP101610,STRP101611,STRP101612,  
STRP101615,STRP101636,STRP101643,STRP101644,STRP101645,STRP101659,STRP101660,STRP101662,STRP101664,STRP101668,  
STRP101675,STRP101677,STRP101678,STRP101683,STRP101685,STRP101691

**Table T4.** Sets of core-genome protein families estimated by BDBH, OMCL and COG strategies for 50 *Streptococcus* proteomes downloaded from OMA browser. All 177 core genes annotated by OMA (OMAc, <http://omabrowser.org/All/oma-groups.txt.gz>) were successfully retrieved by OMCL and COGS, in 176 clusters, after merging two paralogous copies of topoisomerase 4 subunit A in a single cluster. The same result was obtained by BDBH after lowering alignment coverage to 50% from default 75%. However, 6 sequences from OMAc were not found among consensus clusters, due to different groupings of inparalogs by BDBH (5 cases) and OMCL (1 case). Of the 611 core clusters published by Lefébure & Stanhope in 2007 for 26 genomes, 343, 365 and 373 exactly matched GET\_HOMOLOGUES clusters, respectively. Note that their

clustering procedure included a post-processing step, as they produced global CLUSTAL alignments of nucleotides sequences and applied sequence identity cutoffs to call orthologues. Furthermore, the dataset used by Lefébure & Stanhope contained only 26 genomes from 6 species, making their core-genome size estimate necessarily larger.

	BDBH	OMCL	COG	consensus
<i>get_homologues.pl</i> (50 genomes, 14 species, 75 % alignment coverage)	487	521	538	456
same string of Pfam domains	391	413	428	364
OMAc (177 core genes)	176	176	176	170
Lefébure & Stanhope (26 genomes, 611 genes, 6 species)	343	365	373	324

**Table T5.** AutoFACT annotation of 286 consensus core-genes of 50 *Streptococcus* genomes, produced by BDBH, COGtriangles and OMCL with 75% alignment coverage, which are not included in OMAc.

OMA ID	Gene	Subject title	E-score	Source DB
STRA100001	dnaA	Chromosomal replication initiator protein DnaA n=10 Tax=Streptococcus agalactiae ReplD=DNAA_STRA1	0	uniref90
STRA100007	PTH	Peptidyl-tRNA hydrolase n=10 Tax=Streptococcus agalactiae ReplD=PTH_STRA1	1.00E-105	uniref90
STRA100013	TILS	tRNA(Ile)-lysine synthase n=11 Tax=Streptococcus agalactiae ReplD=TILS_STRA3	0	uniref90
STRA100015	ftsH	ATP-dependent zinc metalloprotease FtsH n=10 Tax=Streptococcus agalactiae ReplD=E7S155_STRAG	0	uniref90
STRA100017	KPRS1	Ribose-phosphate pyrophosphokinase 1 n=36 Tax=Streptococcus ReplD=KPRS1_STRA3	1.00E-164	uniref90
STRA100022	acyl carrier protein, putative	Acyl carrier protein 2 n=10 Tax=Streptococcus agalactiae ReplD=E7S7A7_STRAG	2.00E-36	uniref90
STRA100023	PUR7	Phosphoribosylaminoimidazole-succinocarboxamide synthase n=24 Tax=Streptococcus ReplD=PUR7_STRA1	1.00E-133	uniref90
STRA100025	purF	Amidophosphoribosyltransferase n=26 Tax=Streptococcus ReplD=Q1J937_STRPF	0	uniref90
STRA100026	PUR5	Phosphoribosylformylglycinamide cyclo-ligase n=8 Tax=Streptococcus ReplD=PUR5_STRPB	1.00E-176	uniref90
STRA100043	purE	Phosphoribosylaminoimidazole carboxylase, catalytic subunit n=23 Tax=Streptococcus ReplD=Q3D3T6_STRAG	4.00E-80	uniref90
STRA100049		Phosphotyrosine protein phosphatase n=10 Tax=Streptococcus agalactiae ReplD=E7S775_STRAG	5.00E-82	uniref90
STRA100065	50S ribosomal protein L29	50S ribosomal protein L29 n=84 Tax=Streptococcus ReplD=RL29_STRMU	1.00E-17	uniref90
STRA100075	rpmD	50S ribosomal protein L30 n=73 Tax=Streptococcus ReplD=RL30_STRE4	1.00E-23	uniref90
STRA100078	KAD	Adenylate kinase n=24 Tax=Streptococcus ReplD=KAD_STRP3	1.00E-117	uniref90
STRA100080	rpsM	30S ribosomal protein S13 n=260 Tax=Streptococcus ReplD=RS13_STRE4	5.00E-53	uniref90
STRA100082	RPOA	DNA-directed RNA polymerase subunit alpha n=68 Tax=Streptococcus ReplD=RPOA_STRP1	1.00E-173	uniref90
STRA100094	dnaK	Chaperone protein DnaK n=267 Tax=Streptococcus	0	uniref90

		RepID=DNAK_STRMU		
STRA100095	dnaJ	Chaperone protein DnaJ n=11 Tax=Streptococcus agalactiae RepID=DNAJ_STRAG3	0	uniref90
STRA100097	TRUA	tRNA pseudouridine synthase A n=10 Tax=Streptococcus agalactiae RepID=TRUA_STRAG1	1.00E-151	uniref90
STRA100098		Phosphomethylpyrimidine kinase n=10 Tax=Streptococcus agalactiae RepID=E7S721_STRAG	1.00E-143	uniref90
STRA100139	UPPP	Undecaprenyl-diphosphatase n=10 Tax=Streptococcus agalactiae RepID=UPPP_STRAG1	1.00E-149	uniref90
STRA100140		Adapter protein mecA n=9 Tax=Streptococcus agalactiae RepID=MECA_STRAG1	1.00E-131	uniref90
STRA100144	sufS	Cysteine desulfurase n=10 Tax=Streptococcus agalactiae RepID=E7S6Y0_STRAG	0	uniref90
STRA100145	NifU family protein	NifU family SUF system FeS assembly protein n=10 Tax=Streptococcus agalactiae RepID=E7S6X9_STRAG	4.00E-82	uniref90
STRA100156	zinc ABC transporter, ATP-binding protein	Zinc ABC superfamily ATP binding cassette transporter, ABC protein n=10 Tax=Streptococcus agalactiae RepID=E7S6W7_STRAG	1.00E-138	uniref90
STRA100160	penicillin-binding protein 1B	Penicillin-binding protein 1B n=14 Tax=Streptococcus agalactiae RepID=B2NHG3_STRAG	0	uniref90
STRA100163		DUF1033 multi-domain protein	5.00E-39	CDD
STRA100164	cglA	Competence protein ComGA n=9 Tax=Streptococcus agalactiae RepID=E7S6V8_STRAG	0	uniref90
STRA100169	competence protein ComGF	Competence protein ComGF n=10 Tax=Streptococcus agalactiae RepID=F8XY10_STRAG	4.00E-66	uniref90
STRA100171		Adenine-specific methyltransferase n=10 Tax=Streptococcus agalactiae RepID=E7S6V1_STRAG	1.00E-170	uniref90
STRA100177	proC	Pyrroline-5-carboxylate reductase n=10 Tax=Streptococcus agalactiae RepID=E7S6U7_STRAG	1.00E-144	uniref90
STRA100178	pepA	Glutamyl aminopeptidase n=10 Tax=Streptococcus agalactiae RepID=E7S6U6_STRAG	0	uniref90
STRA100206	cysE	Serine acetyltransferase n=10 Tax=Streptococcus agalactiae RepID=E7S6R7_STRAG	4.00E-95	uniref90
STRA100209		Ribonuclease III n=10 Tax=Streptococcus agalactiae RepID=E7S6R4_STRAG	4.00E-69	uniref90
STRA100210		TrmH family RNA methyltransferase n=10 Tax=Streptococcus agalactiae RepID=E7S6R3_STRAG	1.00E-143	uniref90
STRA100260	nagA	N-acetylglucosamine-6-phosphate deacetylase n=10 Tax=Streptococcus agalactiae RepID=E7S6J9_STRAG	0	uniref90
STRA100278	PROA	Gamma-glutamyl phosphate reductase n=11 Tax=Streptococcus agalactiae RepID=PROA_STRAG1	0	uniref90
STRA100280		Cell division protein FtsL n=10 Tax=Streptococcus agalactiae RepID=E7S6E4_STRAG	6.00E-55	uniref90
STRA100281	pbpX	Penicillin-binding protein 2X n=31 Tax=Streptococcus agalactiae RepID=B2NHG6_STRAG	0	uniref90
STRA100282	MRAY	Phospho-N-acetylmuramoyl-pentapeptide-transferase n=10 Tax=Streptococcus agalactiae RepID=MRAY_STRAG1	1.00E-177	uniref90
STRA100283		DEAD-box ATP dependent DNA helicase n=5 Tax=Streptococcus agalactiae RepID=F8XYE2_STRAG	0	uniref90
STRA100290	NADE	NH(3)-dependent NAD(+) synthetase n=10 Tax=Streptococcus agalactiae RepID=NADE_STRAG1	1.00E-156	uniref90
STRA100291	pepC	Amino peptidase C n=10 Tax=Streptococcus agalactiae RepID=E7S6D2_STRAG	0	uniref90
STRA100292	pbp1A	Penicillin-binding protein 1A n=21 Tax=Streptococcus agalactiae RepID=Q8E1Q5_STRAG5	0	uniref90
STRA100293	recU	Holliday junction resolvase recU n=11 Tax=Streptococcus agalactiae RepID=RECU_STRAG1	1.00E-106	uniref90
STRA100296		Site-specific DNA-methyltransferase n=8 Tax=Streptococcus agalactiae RepID=E7S6C7_STRAG	0	uniref90
STRA100298	LUXS	S-ribosylhomocysteine lyase n=24 Tax=Streptococcus RepID=LUXS_STRAG1	4.00E-81	uniref90
STRA100308	FMT	Methionyl-tRNA formyltransferase n=10 Tax=Streptococcus agalactiae RepID=FMT_STRAG1	1.00E-179	uniref90
STRA100309	sun	RNA methyltransferase Sun n=11 Tax=Streptococcus agalactiae RepID=E7S6B3_STRAG	0	uniref90
STRA100310		Protein phosphatase 2C n=12 Tax=Streptococcus agalactiae RepID=F8XYG9_STRAG	1.00E-138	uniref90



STRA100326	cysK	Cysteine synthase n=10 Tax=Streptococcus agalactiae ReplD=E7S696_STRAG	1.00E-156	uniref90
STRA100327		Xaa-Pro dipeptidase n=10 Tax=Streptococcus agalactiae ReplD=E7S695_STRAG	1.00E-118	uniref90
STRA100330	yfiA	Ribosomal subunit interface protein n=10 Tax=Streptococcus agalactiae ReplD=E7S692_STRAG	1.00E-99	uniref90
STRA100333		Enoyl-CoA hydratase/isomerase n=9 Tax=Streptococcus agalactiae ReplD=E7S689_STRAG	1.00E-120	uniref90
STRA100335	FABH	3-oxoacyl-[acyl-carrier-protein] synthase 3 n=9 Tax=Streptococcus agalactiae ReplD=FABH_STRAG1	0	uniref90
STRA100338		Malonyl CoA-acyl carrier protein transacylase n=10 Tax=Streptococcus agalactiae ReplD=E7S684_STRAG	1.00E-175	uniref90
STRA100339	fabG	3-oxoacyl-[acyl-carrier-protein] reductase n=10 Tax=Streptococcus agalactiae ReplD=E7S683_STRAG	1.00E-135	uniref90
STRA100340	fabF	3-oxoacyl-[acyl-carrier-protein] synthase 2 n=10 Tax=Streptococcus agalactiae ReplD=E7S682_STRAG	0	uniref90
STRA100342	fabZ	(3R)-hydroxymyristoyl-[acyl-carrier-protein] dehydratase n=14 Tax=Streptococcus ReplD=FABZ_STRE4	2.00E-69	uniref90
STRA100347	SYS	Serine--tRNA ligase n=11 Tax=Streptococcus agalactiae ReplD=SYS_STRAG1	0	uniref90
STRA100350	PTS system, mannose- specific IID component	PTS system, mannose-specific IID component n=9 Tax=Streptococcus agalactiae ReplD=Q3D755_STRAG	1.00E-151	uniref90
STRA100356	xanthine/uracil permease family protein	xanthine/uracil permease family protein; K06901 putative MFS transporter, AGZA family, xanthine/uracil permease	0	kegg
STRA100357		ATP/GTP hydrolase n=8 Tax=Streptococcus agalactiae ReplD=E7S664_STRAG	3.00E-84	uniref90
STRA100359		LytR family transcriptional regulator n=11 Tax=Streptococcus agalactiae ReplD=E7S662_STRAG	0	uniref90
STRA100361	hit	HIT family protein n=10 Tax=Streptococcus agalactiae ReplD=E7S660_STRAG	3.00E-76	uniref90
STRA100365		Phosphotransferase enzyme family protein n=10 Tax=Streptococcus agalactiae ReplD=E7S656_STRAG	1.00E-155	uniref90
STRA100368	nusA	Transcription termination factor NusA n=10 Tax=Streptococcus agalactiae ReplD=E7S653_STRAG	0	uniref90
STRA100370		50S ribosomal protein L7/L12 n=10 Tax=Streptococcus agalactiae ReplD=E7S651_STRAG	1.00E-50	uniref90
STRA100371	infB	Translation initiation factor IF-2 n=19 Tax=Streptococcus agalactiae ReplD=IF2_STRAG1	0	uniref90
STRA100372	rbfA	Ribosome-binding factor A n=11 Tax=Streptococcus agalactiae ReplD=RBFA_STRAG1	6.00E-61	uniref90
STRA100378		hydrolase, haloacid dehalogenase-like family	1.00E-156	kegg
STRA100391	G6PI	Glucose-6-phosphate isomerase n=38 Tax=Streptococcus ReplD=G6PI_STRPF	0	uniref90
STRA100395	galU	Glucose-1-phosphate-uridylyltransferase n=20 Tax=Streptococcus ReplD=DORSA2_9STRE	1.00E-146	uniref90
STRA100396	GPDA	Glycerol-3-phosphate dehydrogenase [NAD(P)+] n=9 Tax=Streptococcus agalactiae ReplD=GPDA_STRAG1	0	uniref90
STRA100398	membrane protein OxaA, putative	Membrane protein insertase YidC 1 n=9 Tax=Streptococcus agalactiae ReplD=YIDC1_STRAG3	1.00E-138	uniref90
STRA100408		Ribonucleoside-diphosphate reductase n=9 Tax=Streptococcus agalactiae ReplD=E7S5U6_STRAG	0	uniref90
STRA100445	SYV	Valine--tRNA ligase n=20 Tax=Streptococcus ReplD=SYV_STRAG1	0	uniref90
STRA100471	glk	Glucokinase n=9 Tax=Streptococcus agalactiae ReplD=F8XYV1_STRAG	1.00E-166	uniref90
STRA100475	MURD	UDP-N-acetylmuramoylalanine--D-glutamate ligase n=11 Tax=Streptococcus agalactiae ReplD=MURD_STRAG1	0	uniref90
STRA100480		Cell division protein YlmE n=10 Tax=Streptococcus agalactiae ReplD=E7S5K1_STRAG	1.00E-120	uniref90
STRA100481		[S] COG1799 Uncharacterized protein conserved in bacteria	3.00E-53	cog
STRA100484	cell division protein DivIVA, putative	Cell division protein DivIVA n=10 Tax=Streptococcus agalactiae ReplD=E7S5J7_STRAG	1.00E-143	uniref90
STRA100493	FOLD	Bifunctional protein FOLD n=10 Tax=Streptococcus agalactiae ReplD=FOLD_STRAG1	1.00E-163	uniref90
STRA100495	EX7L	Exodeoxyribonuclease 7 large subunit n=10 Tax=Streptococcus	0	uniref90

		agalactiae RepID=EX7L_STR1		
STRA100496	EX7S	Exodeoxyribonuclease 7 small subunit n=11 Tax=Streptococcus agalactiae RepID=EX7S_STR1	4.00E-31	uniref90
STRA100497		Geranyltransferase n=10 Tax=Streptococcus agalactiae RepID=E7S5M9_STRAG	1.00E-164	uniref90
STRA100501		DegV family protein n=8 Tax=Streptococcus agalactiae RepID=E7S5M4_STRAG	1.00E-158	uniref90
STRA100555	PYRDA	Putative dihydroorotate dehydrogenase A (fumarate) n=9 Tax=Streptococcus agalactiae RepID=PYRDA_STR1	0	uniref90
STRA100559		PRK10513 Putative haloacid dehalogenase-like hydrolase [Erwinia billingiae Eb661]	1.00E-80	NCBI clusters
STRA100566	ftsE	Cell division ATP-binding protein FtsE n=10 Tax=Streptococcus agalactiae RepID=E7S5J1_STRAG	1.00E-130	uniref90
STRA100567	ftsX	Cell division protein FtsX n=10 Tax=Streptococcus agalactiae RepID=E7S5J0_STRAG	1.00E-139	uniref90
STRA100572	aspC	Aspartate transaminase n=10 Tax=Streptococcus agalactiae RepID=E7S5I4_STRAG	0	uniref90
STRA100578		Transporter n=10 Tax=Streptococcus agalactiae RepID=Q3DTC0_STRAG	1.00E-150	uniref90
STRA100583		DHH family protein n=10 Tax=Streptococcus agalactiae RepID=E7S5H3_STRAG	0	uniref90
STRA100601	gyrB	DNA gyrase subunit B n=44 Tax=Streptococcus RepID=E4SQY5_STRTN	0	uniref90
STRA100606	ENO	Enolase n=350 Tax=Streptococcus RepID=ENO_STRP2	0	uniref90
STRA100609	aroK	Shikimate kinase n=8 Tax=Streptococcus agalactiae RepID=F8XZ46_STRAG	1.00E-72	uniref90
STRA100720	ccpA	Catabolite control protein A n=52 Tax=Streptococcus RepID=E7S562_STRAG	0	uniref90
STRA100723		Group 1 glycosyl transferase n=10 Tax=Streptococcus agalactiae RepID=E7S559_STRAG	0	uniref90
STRA100724	SYT	Threonine--tRNA ligase n=10 Tax=Streptococcus agalactiae RepID=SYT_STR1	0	uniref90
STRA100733		Sensor histidine kinase VicK n=10 Tax=Streptococcus agalactiae RepID=E7S548_STRAG	0	uniref90
STRA100737		PRK02224 SMC domain protein [Natrialba magadii ATCC 43099]	4.00E-06	NCBI clusters
STRA100739		hydrolase, haloacid dehalogenase-like family	1.00E-149	kegg
STRA100742		ABC superfamily ATP binding cassette transporter, membrane protein n=11 Tax=Streptococcus agalactiae RepID=E7S539_STRAG	1.00E-168	uniref90
STRA100750	lgt	Proteoglycan diacylglycerol transferase n=9 Tax=Streptococcus agalactiae RepID=LGT_STR1	1.00E-137	uniref90
STRA100765		Phosphoglycerate mutase n=10 Tax=Streptococcus agalactiae RepID=E7S516_STRAG	1.00E-116	uniref90
STRA100774	tufA	Elongation factor Tu n=341 Tax=Streptococcus RepID=EFTU_STRP3	0	uniref90
STRA100775	TPIS	Triosephosphate isomerase n=15 Tax=Streptococcus RepID=TPIS_STR1	1.00E-136	uniref90
STRA100776		2,3-bisphosphoglycerate-dependent phosphoglycerate mutase n=2 Tax=Streptococcus RepID=F3L8D8_STRPO	1.00E-124	uniref90
STRA100779	DDL	D-alanine--D-alanine ligase n=10 Tax=Streptococcus agalactiae RepID=DDL_STR1	0	uniref90
STRA100792		Acyltransferase n=9 Tax=Streptococcus agalactiae RepID=E7S4Y8_STRAG	1.00E-132	uniref90
STRA100799		DNA-directed DNA polymerase III delta subunit n=11 Tax=Streptococcus agalactiae RepID=E7S4Y1_STRAG	0	uniref90
STRA100811	NAGB	Glucosamine-6-phosphate deaminase n=9 Tax=Streptococcus agalactiae RepID=NAGB_STR1	1.00E-131	uniref90
STRA100816	pepB	Group B oligopeptidase pepB n=9 Tax=Streptococcus agalactiae RepID=PEPB_STR3	0	uniref90
STRA100818		O-methyltransferase n=9 Tax=Streptococcus agalactiae RepID=E7S4W2_STRAG	1.00E-120	uniref90
STRA100821	SYA	Alanine--tRNA ligase n=25 Tax=Streptococcus RepID=SYA_STRPB	0	uniref90
STRA100828	nrdF-2	Ribonucleotide-pyrophosphate reductase subunit beta n=67 Tax=Streptococcus RepID=G5JRK0_STRCG	1.00E-174	uniref90
STRA100831	ptsH	Phosphocarrier protein HPr n=272 Tax=Streptococcus RepID=PTHP_STR1	2.00E-30	uniref90
STRA100832	PT1	Phosphoenolpyruvate-protein phosphotransferase n=266 Tax=Streptococcus RepID=PT1_STR1	0	uniref90
STRA100839		Biotin--[acetyl-CoA-carboxylase] ligase n=10 Tax=Streptococcus agalactiae RepID=E7S4T9_STRAG	1.00E-179	uniref90

STRA100840	METK	S-adenosylmethionine synthase n=10 Tax=Streptococcus agalactiae ReplID=METK_STR1	0	uniref90
STRA100852	MURA1	UDP-N-acetylglucosamine 1-carboxyvinyltransferase 1 n=12 Tax=Streptococcus agalactiae ReplID=MURA1_STR1	0	uniref90
STRA100859	DNLJ	DNA ligase n=10 Tax=Streptococcus agalactiae ReplID=DNLJ_STR1	0	uniref90
STRA100866	atpE	ATP synthase F0 sector subunit C n=9 Tax=Streptococcus agalactiae ReplID=E7S4R0_STRAG	4.00E-21	uniref90
STRA100868	atpF	ATP synthase subunit b n=12 Tax=Streptococcus agalactiae ReplID=ATPF_STR1	3.00E-78	uniref90
STRA100869	atpH	ATP synthase subunit delta n=10 Tax=Streptococcus agalactiae ReplID=ATPD_STR1	2.00E-97	uniref90
STRA100870	ATPA	ATP synthase subunit alpha n=109 Tax=Streptococcus ReplID=ATPA_STRP1	0	uniref90
STRA100873	atpC	ATP synthase epsilon chain n=10 Tax=Streptococcus agalactiae ReplID=ATPE_STR1	2.00E-62	uniref90
STRA100882	ADDB	ATP-dependent helicase/deoxyribonuclease subunit B n=12 Tax=Streptococcus agalactiae ReplID=ADDB_STR1	0	uniref90
STRA100894		[S] COG1624 Uncharacterized conserved protein	1.00E-112	cog
STRA100901		HAD-superfamily hydrolase n=11 Tax=Streptococcus agalactiae ReplID=E7S4M5_STRAG	1.00E-144	uniref90
STRA100920	dnaE	DNA-directed DNA polymerase III alpha subunit n=10 Tax=Streptococcus agalactiae ReplID=E7S4K4_STRAG	0	uniref90
STRA100925	GLMS	Glucosamine--fructose-6-phosphate aminotransferase [isomerizing] n=9 Tax=Streptococcus agalactiae ReplID=GLMS_STR3	0	uniref90
STRA100930	rpsT	30S ribosomal protein S20 n=216 Tax=Streptococcus ReplID=RS20_STRPB	3.00E-28	uniref90
STRA100935		Ribose ABC superfamily ATP binding cassette transporter, ABC protein n=10 Tax=Streptococcus agalactiae ReplID=E7S4J0_STRAG	0	uniref90
STRA100936	sugar ABC transporter, permease protein, putative	Carbohydrate uptake ABC superfamily, ATP binding cassette transporter, membrane protein n=10 Tax=Streptococcus agalactiae ReplID=E7S4I9_STRAG	1.00E-158	uniref90
STRA100937	sugar ABC transporter, permease protein, putative	Carbohydrate uptake ABC superfamily, ATP binding cassette transporter, membrane protein n=10 Tax=Streptococcus agalactiae ReplID=E7S4I8_STRAG	1.00E-176	uniref90
STRA100939	LDH	L-lactate dehydrogenase n=40 Tax=Streptococcus ReplID=LDH_STRSV	1.00E-170	uniref90
STRA100945	GUAA	GMP synthase [glutamine-hydrolyzing] n=53 Tax=Streptococcus ReplID=GUAA_STRT1	0	uniref90
STRA100946	transcriptional regulator, GntR family	GntR family transcriptional regulator n=10 Tax=Streptococcus agalactiae ReplID=E7S4H9_STRAG	1.00E-133	uniref90
STRA100956		Tyrosine recombinase XerS n=10 Tax=Streptococcus agalactiae ReplID=XERS_STR1	0	uniref90
STRA100961		UPF0122 protein SUB0894 n=1 Tax=Streptococcus uberis 0140J ReplID=B9DUB5_STRU0	3.00E-45	uniref90
STRA100967	PSTB2	Phosphate import ATP-binding protein PstB 2 n=46 Tax=Streptococcus ReplID=PSTB2_STRP1	1.00E-145	uniref90
STRA100974	spxA	Regulatory protein spx n=9 Tax=Streptococcus agalactiae ReplID=SPX_STR3	5.00E-74	uniref90
STRA100975		Macrolide-efflux protein n=10 Tax=Streptococcus agalactiae ReplID=O34127_STRAG	0	uniref90
STRA100983	topA	DNA topoisomerase n=10 Tax=Streptococcus agalactiae ReplID=E7S4E1_STRAG	0	uniref90
STRA100990	RNH2	Ribonuclease HII n=10 Tax=Streptococcus agalactiae ReplID=RNH2_STR3	1.00E-128	uniref90
STRA100991		Ribosome biogenesis GTPase A n=10 Tax=Streptococcus agalactiae ReplID=E7S4D2_STRAG	1.00E-160	uniref90
STRA101015	CARB	Carbamoyl-phosphate synthase large chain n=11 Tax=Streptococcus agalactiae ReplID=CARB_STR1	0	uniref90
STRA101016	CARA	Carbamoyl-phosphate synthase small chain n=10 Tax=Streptococcus agalactiae ReplID=CARA_STR3	0	uniref90
STRA101017	PYRB	Aspartate carbamoyltransferase n=10 Tax=Streptococcus agalactiae ReplID=PYRB_STR1	1.00E-180	uniref90
STRA101018	pyrC	Dihydroorotase n=10 Tax=Streptococcus agalactiae ReplID=E7S4A9_STRAG	0	uniref90
STRA101020	PYRF	Orotidine 5'-phosphate decarboxylase n=10 Tax=Streptococcus	1.00E-135	uniref90

		agalactiae RepID=PYRF_STR1		
STRA101027	FTHS	Formate-tetrahydrofolate ligase n=9 Tax=Streptococcus agalactiae RepID=FTHS_STR1	0	uniref90
STRA101042		NA37 domain containing protein	2.00E-15	pfam
STRA101043	GLYA	Serine hydroxymethyltransferase n=10 Tax=Streptococcus agalactiae RepID=GLYA_STR1	0	uniref90
STRA101045		Protein-(Glutamine-N5) methyltransferase n=10 Tax=Streptococcus agalactiae RepID=E7S480_STRAG	1.00E-158	uniref90
STRA101048	xyIM	Probable tautomerase spyM18_1099 n=14 Tax=Streptococcus pyogenes RepID=Y1099_STRP8	5.00E-26	uniref90
STRA101049	ApbE family protein	ApbE family protein n=10 Tax=Streptococcus agalactiae RepID=Q3K116_STR1	0	uniref90
STRA101061		Pseudouridine synthase n=10 Tax=Streptococcus agalactiae RepID=E7S464_STRAG	1.00E-170	uniref90
STRA101062	PPNK	Probable inorganic polyphosphate/ATP-NAD kinase n=8 Tax=Streptococcus agalactiae RepID=PPNK_STR3	1.00E-158	uniref90
STRA101066		Cysteine desulfurase n=10 Tax=Streptococcus agalactiae RepID=E7S459_STRAG	0	uniref90
STRA101068		Redox-sensing transcriptional repressor rex n=10 Tax=Streptococcus agalactiae RepID=REX_STR3	1.00E-119	uniref90
STRA101077	potC	Spermidine/putrescine ABC superfamily ATP binding cassette transporter, permease protein n=10 Tax=Streptococcus agalactiae RepID=E7S447_STRAG	1.00E-124	uniref90
STRA101079	POTA	Spermidine/putrescine import ATP-binding protein PotA n=14 Tax=Streptococcus RepID=POTA_STR1	0	uniref90
STRA101105	pcrA	ATP-dependent DNA helicase PcrA n=10 Tax=Streptococcus agalactiae RepID=E7S415_STRAG	0	uniref90
STRA101118	PLSY	Glycerol-3-phosphate acyltransferase n=9 Tax=Streptococcus agalactiae RepID=PLSY_STR3	6.00E-89	uniref90
STRA101119	UNG	Uracil-DNA glycosylase n=11 Tax=Streptococcus agalactiae RepID=UNG_STR1	1.00E-124	uniref90
STRA101140	deoD-2	Purine nucleoside phosphorylase I n=11 Tax=Streptococcus agalactiae RepID=E7S3X5_STRAG	1.00E-154	uniref90
STRA101142	RPIA	Ribose-5-phosphate isomerase A n=9 Tax=Streptococcus agalactiae RepID=RPIA_STR1	1.00E-116	uniref90
STRA101145		Metallo-beta-lactamase superfamily protein n=11 Tax=Streptococcus agalactiae RepID=E7S3X0_STRAG	0	uniref90
STRA101148		ABC superfamily ATP binding cassette transporter, binding protein n=9 Tax=Streptococcus agalactiae RepID=E7S3W7_STRAG	1.00E-173	uniref90
STRA101152		Tetratricopeptide (TPR) domain protein n=10 Tax=Streptococcus agalactiae RepID=E7S3W3_STRAG	0	uniref90
STRA101162		SAM-dependent methyltransferase n=11 Tax=Streptococcus agalactiae RepID=E7S3V2_STRAG	1.00E-128	uniref90
STRA101167		Short-chain dehydrogenase/reductase family oxidoreductase n=9 Tax=Streptococcus agalactiae RepID=E7S3U7_STRAG	1.00E-144	uniref90
STRA101169		PRK05270 galactose-1-phosphate uridylyltransferase [Enterococcus faecalis V583]	5.00E-07	NCBI clusters
STRA101170	GTP-binding protein	GTPase HflX n=9 Tax=Streptococcus agalactiae RepID=E7S3U4_STRAG	0	uniref90
STRA101171	MIAA	tRNA dimethylallyltransferase n=10 Tax=Streptococcus agalactiae RepID=MIAA_STR1	1.00E-170	uniref90
STRA101203	rplJ	50S ribosomal protein L10 n=13 Tax=Streptococcus RepID=RL10_STR1	4.00E-86	uniref90
STRA101211	similar to ATP- dependent Clp protease subunit X	ATP-dependent Clp protease ATP-binding subunit ClpX n=9 Tax=Streptococcus RepID=CLPX_STR5	0	uniref90
STRA101213	folA	Dihydrofolate reductase n=10 Tax=Streptococcus agalactiae RepID=E7S3R5_STRAG	3.00E-93	uniref90
STRA101215		Hydroxymethylglutaryl-CoA synthase n=10 Tax=Streptococcus agalactiae RepID=E7S3R3_STRAG	0	uniref90
STRA101218		Hemolysin III n=10 Tax=Streptococcus agalactiae RepID=E7S3Q9_STRAG	1.00E-107	uniref90
STRA101224		Diphosphomevalonate decarboxylase n=10 Tax=Streptococcus agalactiae RepID=E7S3Q3_STRAG	1.00E-167	uniref90
STRA101225		Mevalonate kinase n=10 Tax=Streptococcus agalactiae RepID=E7S3Q2_STRAG	1.00E-143	uniref90
STRA101239		ABC superfamily ATP binding cassette transporter, ABC protein n=10 Tax=Streptococcus agalactiae RepID=E7S3N6_STRAG	0	uniref90

STRA101240	CCA	CCA-adding enzyme n=10 Tax=Streptococcus agalactiae ReplID=CCA_STR A3	0	uniref90
STRA101252	FENR	Ferredoxin--NADP reductase n=9 Tax=Streptococcus agalactiae ReplID=FENR_STR A1	0	uniref90
STRA101257	rpsP	30S ribosomal protein S16 n=22 Tax=Streptococcus ReplID=RS16_STR A1	2.00E-46	uniref90
STRA101263	PYRR	Bifunctional protein pyrR n=9 Tax=Streptococcus agalactiae ReplID=PYRR_STR A1	2.00E-84	uniref90
STRA101266		LysR family transcriptional regulator n=10 Tax=Streptococcus agalactiae ReplID=E7S3K7_STR AG	1.00E-175	uniref90
STRA101271	THII	Probable tRNA sulfurtransferase n=10 Tax=Streptococcus agalactiae ReplID=THII_STR A3	0	uniref90
STRA101272		Cysteine desulfurase n=10 Tax=Streptococcus agalactiae ReplID=E7S3K1_STR AG	0	uniref90
STRA101275		PRK13676 multi-domain protein	5.00E-37	CDD
STRA101279		Methyltransferase n=11 Tax=Streptococcus agalactiae ReplID=E7S3J4_STR AG	0	uniref90
STRA101282	rpml	50S ribosomal protein L35 n=3 Tax=Streptococcus ReplID=E3CLI6_STR DO	5.00E-25	uniref90
STRA101285		Unassigned protein		
STRA101289		MOP superfamily multidrug/oligosaccharidyl-lipid/polysaccharide flippase transporter n=10 Tax=Streptococcus agalactiae ReplID=E7S3I4_STR AG	0	uniref90
STRA101297	pflA	Pyruvate formate-lyase activating enzyme n=9 Tax=Streptococcus agalactiae ReplID=E7S3H6_STR AG	1.00E-142	uniref90
STRA101298	CBS domain protein	CBS domain protein n=10 Tax=Streptococcus agalactiae ReplID=E7S3H5_STR AG	0	uniref90
STRA101302		Membrane protein, putative n=10 Tax=Streptococcus agalactiae ReplID=Q3D2E1_STR AG	1.00E-89	uniref90
STRA101326		predicted metal-sulfur cluster biosynthetic enzyme	4.00E-48	kegg
STRA101328	dnaG	DNA primase n=10 Tax=Streptococcus agalactiae ReplID=E7S3E5_STR AG	0	uniref90
STRA101363		ABC superfamily ATP binding cassette transporter, ABC protein n=10 Tax=Streptococcus ReplID=E0PKX3_STR GY	1.00E-127	uniref90
STRA101377	vacB	Ribonuclease R n=10 Tax=Streptococcus agalactiae ReplID=E7S389_STR AG	0	uniref90
STRA101382	COAE	Dephospho-CoA kinase n=9 Tax=Streptococcus agalactiae ReplID=COAE_STR A1	1.00E-109	uniref90
STRA101383	FPG	Formamidopyrimidine-DNA glycosylase n=9 Tax=Streptococcus agalactiae ReplID=FPG_STR A1	1.00E-158	uniref90
STRA101391		PRK00016 putative metalloprotease [Streptococcus agalactiae NEM316]	4.00E-89	NCBI clusters
STRA101397		PRK13672 multi-domain protein	7.00E-28	CDD
STRA101399		S1 RNA binding domain protein n=9 Tax=Streptococcus agalactiae ReplID=E7S364_STR AG	1.00E-156	uniref90
STRA101400	frr	Ribosome-recycling factor n=11 Tax=Streptococcus ReplID=RRF_STR A1	1.00E-101	uniref90
STRA101408	rplK	50S ribosomal protein L11 n=262 Tax=Streptococcus ReplID=RL11_STR P1	4.00E-72	uniref90
STRA101415	similar to DNA translocase	DNA translocase FtsK n=11 Tax=Streptococcus agalactiae ReplID=FTSK_STR A3	0	uniref90
STRA101421	pfs	5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase n=10 Tax=Streptococcus agalactiae ReplID=E7S342_STR AG	1.00E-111	uniref90
STRA101423		ADP-ribose diphosphatase n=10 Tax=Streptococcus agalactiae ReplID=E7S340_STR AG	2.00E-84	uniref90
STRA101424	GLMU	Bifunctional protein GlmU n=10 Tax=Streptococcus agalactiae ReplID=GLMU_STR A1	0	uniref90
STRA101438	metG	Methionine--tRNA ligase n=10 Tax=Streptococcus agalactiae ReplID=E7S324_STR AG	0	uniref90
STRA101444		ArsC family protein n=11 Tax=Streptococcus agalactiae ReplID=E7S317_STR AG	2.00E-50	uniref90
STRA101450		Initiation-control protein yabA n=10 Tax=Streptococcus agalactiae ReplID=YABA_STR A1	2.00E-47	uniref90
STRA101468	rIuB	Pseudouridine synthase n=20 Tax=Streptococcus ReplID=E0PLU3_STR GY	1.00E-124	uniref90
STRA101469	segregation and condensation protein B	Segregation and condensation protein B n=9 Tax=Streptococcus agalactiae ReplID=SCP B_STR A1	2.00E-97	uniref90

STRA101470	segregation and condensation protein A	Segregation and condensation protein A n=9 Tax=Streptococcus agalactiae RepID=SCPA_STR1	1.00E-131	uniref90
STRA101473		Phosphoesterase n=10 Tax=Streptococcus agalactiae RepID=E7S2Y6_STRAG	1.00E-97	uniref90
STRA101475	MURI	Glutamate racemase n=9 Tax=Streptococcus agalactiae RepID=MURI_STR3	1.00E-152	uniref90
STRA101476		UPF0154 protein SAG1601 n=8 Tax=Streptococcus agalactiae RepID=Y1601_STR5	2.00E-37	uniref90
STRA101479		Membrane protein n=10 Tax=Streptococcus agalactiae RepID=E7S2Y0_STRAG	5.00E-96	uniref90
STRA101490	MURC	UDP-N-acetylmuramate--L-alanine ligase n=12 Tax=Streptococcus agalactiae RepID=MURC_STR1	0	uniref90
STRA101495	dnal	Primosomal protein Dnal n=10 Tax=Streptococcus agalactiae RepID=E7S2W5_STRAG	1.00E-176	uniref90
STRA101500		Nucleic acid-binding protein n=5 Tax=Streptococcus agalactiae RepID=E7S2W0_STRAG	1.00E-71	uniref90
STRA101503		Ribosomal RNA small subunit methyltransferase G n=10 Tax=Streptococcus agalactiae RepID=RSMG_STR1	1.00E-120	uniref90
STRA101505		Trk family potassium (K+) transporter, NAD+ binding protein n=9 Tax=Streptococcus agalactiae RepID=E7S2V5_STRAG	1.00E-125	uniref90
STRA101511	ABC transporter, permease protein	D-methionine ABC superfamily ATP binding cassette transporter, membrane protein n=10 Tax=Streptococcus agalactiae RepID=E7S2U8_STRAG	1.00E-81	uniref90
STRA101512	ABC transporter, ATP-binding protein	Methionine import ATP-binding protein MetN n=10 Tax=Streptococcus agalactiae RepID=METN_STR1	0	uniref90
STRA101530		lojap protein family protein n=10 Tax=Streptococcus agalactiae RepID=E7S2S9_STRAG	5.00E-61	uniref90
STRA101534	putative RNA-binding protein containing KH domain	RNA-binding protein n=12 Tax=Streptococcus RepID=E7S2S5_STRAG	5.00E-53	uniref90
STRA101538	gatB	Aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B n=17 Tax=Streptococcus RepID=GATB_STR1	0	uniref90
STRA101539	gatA	Glutamyl-tRNA(Gln) amidotransferase subunit A n=10 Tax=Streptococcus agalactiae RepID=GATA_STR1	0	uniref90
STRA101548		Universal stress family protein n=9 Tax=Streptococcus agalactiae RepID=E7S2R0_STRAG	2.00E-81	uniref90
STRA101551	aroE	Shikimate dehydrogenase n=9 Tax=Streptococcus agalactiae RepID=E7S2Q7_STRAG	1.00E-161	uniref90
STRA101553	recG	ATP-dependent DNA helicase n=23 Tax=Streptococcus RepID=A2RCU6_STRPG	0	uniref90
STRA101555	ALR	Alanine racemase n=9 Tax=Streptococcus agalactiae RepID=ALR_STR1	0	uniref90
STRA101556	ACPS	Holo-[acyl-carrier-protein] synthase n=10 Tax=Streptococcus agalactiae RepID=ACPS_STR1	2.00E-63	uniref90
STRA101561		PTS family beta-glucosides porter, IIABC component n=10 Tax=Streptococcus agalactiae RepID=E7S2P7_STRAG	0	uniref90
STRA101562	scrB	Sucrose-6-phosphate hydrolase n=10 Tax=Streptococcus agalactiae RepID=E7S2P6_STRAG	0	uniref90
STRA101565		[S] COG1302 Uncharacterized protein conserved in bacteria	8.00E-54	cog
STRA101576		Putative membrane protein n=1 Tax=Streptococcus uberis 0140J RepID=B9DVK1_STRU0	5.00E-74	uniref90
STRA101579	ssb-3	Single-stranded DNA-binding protein 1 n=9 Tax=Streptococcus agalactiae RepID=SSB1_STR3	3.00E-73	uniref90
STRA101580	rpsF	30S ribosomal protein S6 n=3 Tax=Streptococcus RepID=E8KT13_STRVE	2.00E-44	uniref90
STRA101581	mutY	A/G-specific adenine glycosylase n=26 Tax=Streptococcus RepID=B5XI42_STRPZ	0	uniref90
STRA101585		CvpA family protein n=10 Tax=Streptococcus agalactiae RepID=E7S2M1_STRAG	4.00E-80	uniref90
STRA101587	RNH3	Ribonuclease HIII n=10 Tax=Streptococcus agalactiae RepID=RNH3_STR1	1.00E-170	uniref90
STRA101588		Signal peptidase I n=10 Tax=Streptococcus agalactiae RepID=E7S2L8_STRAG	2.00E-99	uniref90

STRA101591	DPO4	DNA polymerase IV n=10 Tax=Streptococcus agalactiae ReplID=DPO4_STR1	0	uniref90
STRA101633		Probable tRNA threonylcarbamoyladenine biosynthesis protein Gcp n=9 Tax=Streptococcus agalactiae ReplID=GCP_STR1	1.00E-180	uniref90
STRA101636		PRK13667 multi-domain protein	3.00E-28	CDD
STRA101639	glnA	PRK09469 glutamine synthetase [Cronobacter turicensis z3032]	7.00E-77	NCBI clusters
STRA101642	PGK	Phosphoglycerate kinase n=265 Tax=Streptococcus ReplID=PGK_STRGC	0	uniref90
STRA101645	fusA	Elongation factor G n=265 Tax=Streptococcus ReplID=EFG_STRP2	0	uniref90
STRA101647	rpsL	30S ribosomal protein S12 n=71 Tax=Streptococcus ReplID=RS12_STRP1	8.00E-71	uniref90
STRA101650		RmuC domain-containing protein n=10 Tax=Streptococcus agalactiae ReplID=F8Y2U9_STRAG	0	uniref90
STRA101664	DLTC	D-alanine--poly(phosphoribitol) ligase subunit 2 n=10 Tax=Streptococcus agalactiae ReplID=DLTC_STR1	4.00E-37	uniref90
STRA101699		tRNA-dihydrouridine synthase n=10 Tax=Streptococcus agalactiae ReplID=E7S2B5_STRAG	0	uniref90
STRA101703	ctsR	Transcriptional regulator CtsR n=10 Tax=Streptococcus agalactiae ReplID=E7S2B1_STRAG	4.00E-83	uniref90
STRA101706	rpsB	30S ribosomal protein S2 n=132 Tax=Streptococcus ReplID=RS2_STRSV	1.00E-105	uniref90
STRA101712	pepO	Peptidase family M13 n=11 Tax=Streptococcus ReplID=G5KGT4_9STRE	0	uniref90
STRA101725	cdsA	Phosphatidate cytidyltransferase n=10 Tax=Streptococcus agalactiae ReplID=E7S224_STRAG	1.00E-143	uniref90
STRA101752		GTP diphosphokinase n=10 Tax=Streptococcus agalactiae ReplID=E7S1W1_STRAG	0	uniref90
STRA101780		RsmE family RNA methyltransferase n=11 Tax=Streptococcus agalactiae ReplID=E7S1T2_STRAG	1.00E-126	uniref90
STRA101845	nusG	Transcription antitermination protein nusG n=66 Tax=Streptococcus ReplID=NUSG_STRP1	7.00E-96	uniref90
STRA101850	pbp2A	Penicillin-binding protein 2A n=19 Tax=Streptococcus agalactiae ReplID=B2NHG4_STRAG	0	uniref90
STRA101851		Pseudouridine synthase n=9 Tax=Streptococcus agalactiae ReplID=Q3D7C8_STRAG	1.00E-173	uniref90
STRA101867		GNAT family acetyltransferase n=9 Tax=Streptococcus agalactiae ReplID=E7S1P1_STRAG	9.00E-84	uniref90
STRA101870	nrdD	Anaerobic ribonucleoside-triphosphate reductase large subunit n=9 Tax=Streptococcus agalactiae ReplID=E7S1P5_STRAG	0	uniref90
STRA101875		[S] COG4472 Uncharacterized protein conserved in bacteria	2.00E-41	cog
STRA101876	spxA	Regulatory protein Spx n=10 Tax=Streptococcus ReplID=B9DW64_STRU0	2.00E-63	uniref90
STRA101877	recA	Protein RecA n=2 Tax=Streptococcus agalactiae ReplID=RECA_STRAG	0	uniref90
STRA101880	RUVA	Holliday junction ATP-dependent DNA helicase RuvA n=10 Tax=Streptococcus agalactiae ReplID=RUVA_STR1	1.00E-101	uniref90
STRA101885	hexA	DNA mismatch repair protein MutS n=10 Tax=Streptococcus agalactiae ReplID=MUTS_STR1	0	uniref90
STRA101890		[S] COG1284 Uncharacterized conserved protein	1.00E-118	cog
STRA101919		TetR family transcriptional regulator n=9 Tax=Streptococcus agalactiae ReplID=E7S1D2_STRAG	1.00E-100	uniref90
STRA101943	rplI	50S ribosomal protein L9 n=11 Tax=Streptococcus agalactiae ReplID=RL9_STR1	1.00E-59	uniref90
STRA101947	trmU	tRNA-specific 2-thiouridylase mnmA n=10 Tax=Streptococcus agalactiae ReplID=MNMA_STR1	0	uniref90
STRA101955	pgsA	CDP-diaclyglycerol-glycerol-3-phosphate 3-phosphatidyltransferase n=10 Tax=Streptococcus agalactiae ReplID=E7S194_STRAG	7.00E-97	uniref90
STRA101956		M16C subfamily protease n=11 Tax=Streptococcus agalactiae ReplID=E7S193_STRAG	0	uniref90
STRA101957		peptidase M16 inactive domain protein	0	kegg
STRA101977	htrA	Serine protease HtrA n=10 Tax=Streptococcus agalactiae ReplID=E7S172_STRAG	0	uniref90

**Table T6.** Annotation of *Streptococcus pyogenes*-specific or expanded protein clusters identified when compared to the *S. equi* proteomes from the OMA December2012 release. For each cluster, the protein from *S. pyogenes* M1 GAS was used as the group's representative for querying NCBI's nr protein database and annotated by taking the first hit produced by BLASTP. The first 42 proteins are found in all 11 *S. pyogenes* proteomes. An additional 13 proteins (43-55) were found when their presence was requested in only 90% of the *S. pyogenes* strains. The final three protein families have been selectively expanded in the *S. pyogenes* lineage.

Annot. No.	OMA ID	gene	Subject title	E-score	Source DB
1	STRP100064		Unassigned protein		
2	STRP100068		putative histidine triad (HIT) protein	2.00E-69	kegg
3	STRP100083		Unassigned protein		
4	STRP100100	atoE	Short-chain fatty acids transporter n=16 Tax=Streptococcus pyogenes RepID=Q1J8T9_STRPF	0	uniref90
5	STRP100101		Acetyl-CoA acetyltransferase 1 n=15 Tax=Streptococcus pyogenes RepID=A2RC85_STRPG	0	uniref90
6	STRP100102		Acetate CoA-transferase alpha subunit n=14 Tax=Streptococcus pyogenes RepID=Q1J8T6_STRPF	1.00E-116	uniref90
7	STRP100103		Acetate CoA-transferase beta subunit n=14 Tax=Streptococcus pyogenes RepID=Q5XE60_STRP6	1.00E-118	uniref90
8	STRP100108		V-type sodium ATP synthase subunit I n=15 Tax=Streptococcus pyogenes RepID=Q491H3_STRP1	0	uniref90
9	STRP100109	ntpK	V-type sodium ATP synthase subunit K n=13 Tax=Streptococcus pyogenes RepID=Q1J8S9_STRPF	9.00E-74	uniref90
10	STRP100110	ntpE	V-type ATP synthase, subunit E n=13 Tax=Streptococcus pyogenes RepID=E0PYB4_STRPY	4.00E-97	uniref90
11	STRP100111	ntpC	V-type sodium ATP synthase subunit C n=14 Tax=Streptococcus pyogenes RepID=A2RC95_STRPG	0	uniref90
12	STRP100112	VATA	V-type ATP synthase alpha chain n=15 Tax=Streptococcus pyogenes	0	uniref90



			ReplID=VATA_STRP1		
13	STRP100113		V-type ATP synthase beta chain n=50 Tax=Streptococcus ReplID=VATB_STRP1	0	uniref90
14	STRP100114	ntpD	V-type ATP synthase subunit D n=15 Tax=Streptococcus pyogenes ReplID=VATD_STRPZ	1.00E-104	uniref90
15	STRP100125		Unassigned protein		
16	STRP100148		Membrane protein n=130 Tax=Streptococcus ReplID=B2DQJ9_STRPN	0	uniref90
17	STRP100149	nadC	Nicotinate-nucleotide diphosphorylase n=131 Tax=Streptococcus ReplID=E0PVG0_STRPY	1.00E-155	uniref90
18	STRP100343		Unassigned protein	8.00E-41	uniref90
19	STRP100348		Unassigned protein	1.00E-173	uniref90
20	STRP100349		Unassigned protein	5.00E-60	uniref90
21	STRP100350		Unassigned protein		
22	STRP100360		Glutaminy-peptide cyclotransferase n=16 Tax=Streptococcus pyogenes ReplID=Q1JI41_STRPD	1.00E-137	uniref90
23	STRP100362		Permease n=14 Tax=Streptococcus pyogenes ReplID=E0PXE1_STRPY	1.00E-144	uniref90
24	STRP100363		Permease n=15 Tax=Streptococcus pyogenes ReplID=E0PXE0_STRPY	8.00E-75	uniref90
25	STRP100384		Shikimate dehydrogenase n=15 Tax=Streptococcus pyogenes ReplID=E0PXB1_STRPY	1.00E-178	uniref90
26	STRP100385		Xylose isomerase-like TIM barrel family protein n=16 Tax=Streptococcus pyogenes ReplID=G4R3F0_STRPY	1.00E-127	uniref90
27	STRP100386		Putative F420-0:gamma-glutamyl ligase n=1 Tax=Streptococcus porcinus str. Jelinkova 176 ReplID=F3L7A2_STRPO	3.00E-91	uniref90

28	STRP100387		MviM domain containing protein	3.00E-05	CDD
29	STRP100388		Methionine adenosyltransferase n=2 Tax=Streptococcus pseudoporcinus ReplID=E4L1P8_9STRE	0	uniref90
30	STRP100389		HprK_C domain containing protein	3.00E-06	CDD
31	STRP100390		Glycosyl transferase 2 family protein n=13 Tax=Streptococcus pyogenes ReplID=G4R3F5_STRPY	1.00E-177	uniref90
32	STRP100396		Unassigned protein		
33	STRP100397		Unassigned protein	0	uniref90
34	STRP100399		Plasmid stabilization system toxin protein n=17 Tax=Streptococcus ReplID=Q1JCY0_STRPB	5.00E-51	uniref90
35	STRP100446		Unassigned protein		
36	STRP100565		DUF1722 multi-domain protein	7.00E-30	CDD
37	STRP100627	czcD	Cation efflux family protein n=13 Tax=Streptococcus pyogenes ReplID=A2RF51_STRPG	1.00E-145	uniref90
38	STRP100707		Phage protein n=5 Tax=Streptococcus ReplID=E7S238_STRAG	8.00E-12	uniref90
39	STRP100775	hylA	Hyaluronate lyase n=31 Tax=cellular organisms ReplID=Q1J6X4_STRPF	0	uniref90
40	STRP100836		Sensor histidine kinase n=17 Tax=Streptococcus ReplID=E0PW29_STRPY	0	uniref90
41	STRP100974	cfa	CAMP factor n=16 Tax=Streptococcus pyogenes ReplID=A2RE73_STRPG	1.00E-145	uniref90
42	STRP101070		Unassigned protein		
43	STRP101307	LACA1	Galactose-6-phosphate isomerase subunit lacA 1 n=20 Tax=Streptococcus ReplID=LACA1_STRP3	5.00E-75	uniref90
44	STRP101359		Unassigned protein	4.00E-31	uniref90
45	STRP101369		ABC transporter permease protein n=15 Tax=Streptococcus pyogenes ReplID=Q1JA65_STRPB	1.00E-111	uniref90

46	STRP101408		Phosphatidylglycerophosphatase B n=24 Tax=Streptococcus RepID=Q1J517_STRPF	1.00E-77	uniref90
47	STRP101432	nupC	Nucleoside permease nupC n=24 Tax=Streptococcus RepID=Q1J4Z2_STRPF	0	uniref90
48	STRP101434		GntR family regulatory protein n=14 Tax=Streptococcus pyogenes RepID=A2RCN1_STRPG	1.00E-137	uniref90
49	STRP101461		Glycerate kinase n=16 Tax=Streptococcus pyogenes RepID=E0PU32_STRPY	0	uniref90
50	STRP101463	hsdS	Type I restriction-modification system specificity subunit n=3 Tax=Streptococcus pyogenes RepID=Q1J9W3_STRPB	0	uniref90
51	STRP101469	salA	SalA2 lantibiotic salivaricin A2 n=5 Tax=Streptococcus salivarius RepID=G2GW05_STRSL	5.00E-19	uniref90
52	STRP101499		Unassigned protein	6.00E-18	uniref90
53	STRP101500		Probable transcriptional regulator n=12 Tax=Streptococcus pyogenes RepID=E0PTY5_STRPY	2.00E-45	uniref90
54	STRP101555	PRSA2	Foldase protein prsA 2 n=14 Tax=Streptococcus pyogenes RepID=PRSA2_STRP6	1.00E-133	uniref90
55	STRP101556	SPEB	Streptopain n=23 Tax=Streptococcus pyogenes RepID=SPEB_STRP3	0	uniref90
56	STRP101557		Unassigned protein	3.00E-24	uniref90
57	STRP101559	ropB	Transcriptional regulator n=49 Tax=Streptococcus pyogenes RepID=Q1J9I9_STRPB	1.00E-159	uniref90
58	STRP101640	uviB	Bacteriocin UviB n=24 Tax=Streptococcus RepID=E0PTK3_STRPY	1.00E-51	uniref90

**Table T7.** Estimated core-genome sizes and memory requirements of algorithm BDBH in default mode and in minimum BLAST mode (-b parameter), enforcing 75% alignment coverage.

test set	number of sequences	BDBH		BDBHmin	
		core clusters	RAM(MB)	core clusters	RAM(MB)
Escherichia coli101	497131	891	N.D.	890	4207
Escherichia coli20	93612	1935	3429	1934	1058
Vibrionales38	165536	559	7699	559	1450
Gammaproteobacteria13	42011	250	636	249	369