SpeciesPrimer: A bioinformatics pipeline dedicated to the design of qPCR primers for the quantification of bacterial species

7 Matthias Dreier[1,2], Hélène Berthoud[1], Noam Shani[1], Daniel Wechsler[1], Pilar Junier[2]

9 [1] Agroscope, Schwarzenburgstrasse 161, CH-3003 Bern, Switzerland

10 [2] Laboratory of Microbiology, University of Neuchâtel, Emile-Argand 11, CH-2000 Neuchâtel,

11 Switzerland

13 Corresponding Author:
14 Matthias Dreier[1,2]
15 Schwarzenburgstrasse 161, Bern, CH-3003, Switzerland
16 Email address: matthias.dreier@agroscope.admin.ch

Abstract

19 **Background.** Quantitative real-time PCR (qPCR) is a well-established method for detecting and

20 quantifying bacteria, and it is progressively replacing culture-based diagnostic methods in food

21 microbiology. High-throughput qPCR using microfluidics brings further advantages by

22 providing faster results, decreasing the costs per sample and reducing errors due to automatic

23 distribution of samples and reagents. In order to develop a high-throughput qPCR approach for

24 the rapid and cost-efficient quantification of microbial species in complex systems such as

25 fermented foods (for instance, cheese), the preliminary setup of qPCR assays working efficiently

26 under identical PCR conditions is required. Identification of target-specific nucleotide sequences

27 and design of specific primers are the most challenging steps in this process. To date, most

28 available tools for primer design require either laborious manual manipulation or high-

29 performance computing systems.

30 **Results.** We developed the SpeciesPrimer pipeline for automated high-throughput screening of

31 species-specific target regions and the design of dedicated primers. Using SpeciesPrimer,

32   specific primers were designed for four bacterial species of importance in cheese quality control,

33   namely *Enterococcus faecium*, *Enterococcus faecalis*, *Pediococcus acidilactici* and

34   *Pediococcus pentosaceus*. Selected primers were first evaluated *in silico* and subsequently *in*

35   *vitro* using DNA from pure cultures of a variety of strains found in dairy products. Specific

36   qPCR assays were developed and validated, satisfying the criteria of inclusivity, exclusivity and

37   amplification efficiencies.

38   **Conclusion.** In this work, we present the SpeciesPrimer pipeline, a tool to design species-

39   specific primers for the detection and quantification of bacterial species. We use SpeciesPrimer

40   to design qPCR assays for four bacterial species and describe a workflow to evaluate the

41   designed primers. SpeciesPrimer facilitates efficient primer design for species-specific

42   quantification, paving the way for a fast and accurate quantitative investigation of microbial

43   communities.

Introduction

44

45   Quantitative real-time PCR (qPCR) is a well-established method for the detection and

46   quantification of bacteria in microbiology, for instance in the context of pathogen detection in

47   clinical and veterinary diagnostics and food safety (Cremonesi et al. 2014; Curran et al. 2007;

48   Garrido-Maestu et al. 2018; Ramirez et al. 2009). Culture-based diagnostic methods are

49   progressively being replaced by qPCR due to advantages such as faster results, more specific

50   detection, and the ability to detect sub-dominant populations (Postollec et al. 2011). High-

51   throughput microfluidic qPCR brings further advantages including the fast generation of results,

52   a lower cost per sample and fewer errors due to automatic distribution of samples and reagents.

53   However, in order to work efficiently high-throughput qPCR systems use identical PCR

54   chemistry and PCR conditions for all reactions taking place on a single chip. Therefore, existing

55   qPCR assays are often not suitable and new primers have to be designed (Hermann-Bank et al.

56   2013; Ishii et al. 2013; Kleyer et al. 2017).

57   The main challenges for the successful development of any qPCR assay are the identification of

58   a specific target nucleotide sequence and the design of primers that bind exclusively to that target

59   sequence. Before microbial draft genomes became widely available, the 16S rRNA gene

60   sequence was frequently used as a target sequence. However, the regions that are targeted in the

61   16S rRNA gene do not provide sufficient resolution to differentiate between closely related

62   bacterial species (Moyaert et al. 2008; Torriani et al. 2001; Wang et al. 2007). Further,

63   housekeeping genes such as, for instance, *tuf, recA* and *pheS*, were successfully used as target

64   sequences for a variety of bacterial species in fermented foods (Falentin et al. 2010; Masco et al.

65   2007; Scheirlinck et al. 2009). Today, the steadily increasing number of prokaryotic draft

66   genomes facilitates the identification of new and unique target regions. This, in combination with

67   the increased computing power, makes it now possible to screen and compare hundreds of

68   genomes and to predict unique target sequences in a relatively short time.

69   Various commercial and open source programs facilitate the design of specific primers for a

70   target sequence, such as the standard tools Primer3 and Primer-BLAST (Untergasser et al. 2012;

71   Ye et al. 2012). Primer3 predicts suitable PCR primers for an input target sequence, while

72   Primer-BLAST combines Primer3 with a BLAST search in a selected nucleotide sequence

73   database to assess the specificity of the primers for the target sequence. Table 1 provides an

74   overview of the features of different primer design tools and pipelines. PrimerMiner (Elbrecht et

75   al. 2017) is a tool that automatically downloads sequences of marker genes for taxonomic group

76   specified by the user and creates alignments and consensus sequences as target sequences for the

77   design of degenerate primers. PrimerServer (Zhu et al. 2017) allows to design primers for

78   multiple sites across a whole genome sequence and performs a specificity check. Tools and

79   pipelines that encompass both the identification of target sequences from bacterial draft genomes

80   and the design of primer candidates include, for instance, RUCS, the find_differential_primers

81   (fdp) pipeline and TOPSI (Pritchard et al. 2012; Thomsen et al. 2017; Vijaya Satya et al. 2010).

82   RUCS is able to identify unique core sequences in a positive set of genomes (target) compared to

83   a negative set of genomes (non-target). It designs primers for the core sequences and validates

84   them with an *in silico* PCR validation method against the positive and negative reference sets.

85   Similarly, the fdp pipeline designs primers for a set of positive genomes and, further, allows to

86   extract primers specific to subclasses of the positive set and performs specificity check against a

87   negative set of genomes. TOPSI is an automated high-throughput pipeline for the design of

88   primers, primarily developed for pathogen-diagnostic assays. It identifies sequences present in

89   all input genomes and designs specific primers accordingly.

90   We aimed to design a series of primers that function with the same qPCR cycling conditions and

91   primer concentrations for later usage in a high-throughput microfluidic qPCR platform. RUCS,

92   fdp and TOPSI can be used to design species-specific primers and offer high-throughput primer

93   design. However, TOPSI could not be used because no Linux-based cluster was available. RUCS

94   and fdp were not able to design primers for all our target species. Therefore, these pipelines were

95   not suitable for our high-throughput approach.

96   This study presents a new pipeline named SpeciesPrimer developed for automated high-

97   throughput screening for species-specific target regions combined with the design of primer

98   candidates for these sequences. The process of primer design is fully automated from the

99   download of bacterial genomes to the quality control of primer candidates. The pipeline runs on

100   a standard computer with a multi-core processor and a minimum of 16 GB RAM. We have

101   applied the SpeciesPrimer pipeline to a set of four bacterial species occurring in cheese and other

102   dairy products and validated the primers *in silico* and *in vitro* by performing qPCR experiments

103   with a variety of target and non-target strains.

Description

104

Overview

105

106   The SpeciesPrimer pipeline consists of three main parts (Table 2). First, genome assemblies are

107   downloaded, annotated and then subjected to quality control. Second, a pan-genome analysis is

108   performed to identify single copy core genes. Conserved sequences of these core genes are then

109   extracted and the specificity for the target species is assessed. Finally, primers are designed for

110   these species-specific conserved core gene sequences and subsequently evaluated in a primer

111   quality control step. An overview of the features of the tools used for SpeciesPrimer can be

112   found in Table S1.

Part 1: Input genome assemblies

113

114   The minimal command line input for the pipeline is the species name. Further, a list of non-target

115   species names can be specified (e.g., species found in the investigated ecosystem but that should

116   not be detected in the specific qPCR assay). For downloading genome assemblies from the

117   National Center for Biotechnology Information (NCBI) automatically, a valid e-mail address is

118   required for accessing the NCBI E-utilities services (Sayers 2009). The pipeline works with a

119   pre-formatted NCBI BLAST database (nt), containing partially non-redundant nucleotide

120   sequences. A local copy of the nt database is required. It can be downloaded from NCBI using

121   the update_blastdb.pl script from the BLAST+ package (Altschul et al. 1990), via FTP from the

122   NCBI FTP server or with the pipeline script (getblastdb.py). The nt database, which consists of

123   sequences from GenBank, EMBL (European Molecular Biology Laboratory) and DDBJ (DNA

124 Data Bank of Japan), was selected because it has a large coverage of diverse sequences, but it is
125 not as large as for example the refseq_genomic database (Tao et al. 2011). The evaluation of the
126 specificity of the target sequence for the target species does not rely on small differences in the
127 nucleotide sequence, but on the overall similarity. Therefore, even with one genome sequence
128 per non-target species we would expect to find similarities in the core genes of the non-target
129 species. Each additional genome of this species in a database would then allow finding more
130 potential sequence similarities in shell genes, cloud genes and strain-specific genes. On the one
131 hand, a more extensive database could better predict the specificity of a sequence, but on the
132 other it would increase the size of the database and the time required for the BLAST search.
133 The user-provided species name is used to search for genome assemblies in the NCBI database.
134 The Biopython Entrez module (Cock et al. 2009) searches the NCBI taxonomic identity (taxid)
135 for the target species in the taxonomy database and downloads the genome assembly summary
136 report. Afterwards, SpeciesPrimer downloads the genome assemblies in FASTA format from the
137 NCBI RefSeq FTP server using the links specified in the summary report. Finally, the
138 downloaded genome assemblies are annotated with Prokka (Seemann 2014).
139 The quality of the genome assemblies is a crucial factor for the pan-genome analysis. Genome
140 assemblies deposited with the wrong taxonomic label or low-quality assemblies drastically
141 reduce the number of identified core genes and of conserved sequences for primer design. The
142 initial quality control step is intended to remove such assemblies from the subsequent analysis.
143 For the verification of the taxonomic classification, the user can choose one or several genes
144 from five conserved housekeeping genes (16S rRNA, *tuf, recA, dnaK* and *pheS*). Genome
145 assemblies without an annotation for the specified conserved housekeeping genes or genome
146 assemblies consisting of more than 500 contigs are removed from the downstream pan-genome
147 analysis. The sequences of the specified conserved housekeeping genes are blasted against the
148 local nt database. Genome assemblies pass the quality control if the best BLAST hit for all
149 sequences is a sequence arising from the target species.

150 **Part 2: Identification of target sequences for primer design**
151 A pan-genome analysis is performed using Roary (Page et al. 2015) to identify the core genes of
152 the target species. Based on the results of the pan-genome analysis, single copy core genes are
153 identified. The gene_presence_absence.csv produced by Roary reports the presence (or absence)
154 of every annotated gene for every input genome assembly. Single copy core genes are the genes

155 for which the number of assemblies harboring the sequence and the number of total identified

156 sequences equals the number of total input assemblies. An sqlite3 database containing all

157 annotated sequences of all assemblies is compiled using the DBGenerator.py script from the

158 Microbial Genomics Lab GitHub repository (https://github.com/microgenomics/tutorials). This

159 database is queried for single copy core genes and the nucleotide sequences are saved in multi-

160 FASTA format. Each multi-FASTA file contains the sequences of one single copy core gene

161 from each input genome assembly. These sequences are aligned using the probabilistic multiple

162 alignment program Prank (Löytynoja 2014). A consensus sequence with ambiguous bases is then

163 created using the consambig function from the EMBOSS package (Rice et al. 2000). The

164 alignments and extraction of the consensus sequences are performed in parallel for several core

165 genes using GNU parallel (Tange 2011). Continuous consensus sequences longer than the

166 minimal PCR product length, harboring less than two ambiguous bases in the range of 20 bases

167 are used for the subsequent steps of the pipeline.

168 These conserved consensus sequences are used for a BLAST search against the local nt database

169 using the discontiguous BLAST algorithm and an e-value cutoff of 500. For all hits in the

170 BLAST results, the species name is extracted from the sequence description and compared with

171 the names in the species list (non-target species). If any species name in the species list matches

172 a hit in the BLAST results the corresponding query sequence is discarded, otherwise the

173 sequence is classified as specific for the target and considered for primer design.

174 **Part 3: Primer design**

175 Primer3 is used to design primers for the unique single copy core gene sequences. As pipeline

176 default the optimum primer melting temperature is set to 60 °C and the maximal primer length is

177 set to 26 bases, all other settings are the default settings of the primer3web version

178 (http://primer3.ut.ee, accessed November 29, 2018). The minimal and maximal amplicon size of

179 the PCR product can be specified individually for every target species through the command line

180 options. The other parameters for primer3 cannot be changed individually, but the general

181 primer3 settings can be changed by modifying the primer3 settings file.

182 The primer quality control consists of three parts, an *in silico* PCR to evaluate the specificity of

183 the primer for the template, an estimation of secondary structures of the amplicon sequence and

184 the estimation of the potential to form primer dimers. The specificity check (Figure 1) for each

185 primer pair is performed with MFEprimer-2.0 (Qu et al. 2012). For the evaluation of the

186 specificity, three indexed databases are generated: the target template database, the non-target
187 sequence database and the target genome database. The target template database consists of the
188 unique conserved core gene sequences used as template for primer design. The non-target
189 sequence database is compiled from sequences of non-target species that show similarities to the
190 primer sequences. To identify these sequences, a BLAST search with all primers against the
191 local nt database is performed. BLAST hits with a species name in the description matching a
192 name in the user-specified non-target species list are selected. These selected sequences and
193 4000 base pairs up- and downstream are extracted from the nt database using the blastdbcmd
194 tool. The target genome database is composed of maximal 10 of the input genome assemblies. If
195 the assembly summary report from the automatic download of genome assemblies from NCBI is
196 available, the genome assemblies as complete as possible are preferred (assembly status:
197 complete > chromosome > scaffold > contig). The target sequence database is used to evaluate
198 the maximum primer pair coverage (PPC, maximum value = 100), a value used by MFEprimer-
199 2.0 to score the ability of the primer pair to bind to a DNA template. All primer pairs with a PPC
200 value lower than the specified threshold (mfethreshold, default = 90) for their template are
201 excluded. Next, MFEprimer-2.0 is used to score the binding of the primer pairs to the sequences
202 of the non-target sequence and the target genome database. The difference of the PPC for the
203 DNA template and the specified threshold (Δthreshold = PPC – mfethreshold) is used as a
204 threshold for the maximum PPC value a primer pair is allowed to have for a non-target sequence.
205 Strong secondary structures at the 5'- or the 3'- end of the PCR product could impair efficient
206 primer binding. Therefore, the PCR products of the primer pairs are submitted to mfold (Zuker et
207 al. 1999) to exclude PCR products with strong secondary structures at the annealing temperature
208 of 60 °C. Moreover, as primer dimers can yield unspecific signals during the qPCR run, the 3'-
209 ends of the primer pairs are checked for their potential to form homo- or hetero-dimers using a
210 Perl script (MPprimer_dimer_check.pl) from MPprimer (Shen et al. 2010).
211 The pipeline output is a list containing the primer name, primer pair coverage (MFEprimer) and
212 penalty values, primer and template sequences and melting temperatures (Primer3). Further, a
213 report of the genome assembly quality control, a file containing the pipeline run statistics, the
214 core gene alignment and the phylogeny in newick format can be found in the output directory.

Materials & Methods

Primer design

SpeciesPrimer pipeline runs were performed on a virtual machine (Oracle VM VirtualBox 5.2.8)
with Ubuntu 16.04 (64-bit) and Docker installed, using 22 of 24 logical processors from two
Intel Xeon E5-2643 CPUs, 32 GB of RAM, a solid-state drive and a LAN Internet connection.
To show the performance of the SpeciesPrimer pipeline on consumer hardware the runs were
repeated on a laptop with an Ubuntu 16.04 (64-bit) operating system, an i7-3610QM CPU (8
logical processors), 8 GB RAM, a solid-state drive and a wireless LAN Internet connection. The
used Docker image is available at https://hub.docker.com/r/biologger/speciesprimer.
The species list consisted of 259 species and subspecies names detected in dairy products,
namely from species names collected from data of 16S rRNA meta-genome sequencing studies
in milk and cheese varieties (Marco Meola Agroscope, pers. comm.) and dairy-related bacteria
from the list of bacterial species and subspecies with technological beneficial use in food
products (Almeida et al. 2014).
The SpeciesPrimer pipeline was run with the input genome assemblies, parameters and the
species list specified in the supplemental information (Dataset S1). Genome assemblies from the
Agroscope Culture Collection were included for the Pediococci.

In silico validation

For the *in silico* validation, PCR products for the designed primer pairs were used for an online
BLAST search against the RefSeq Genomes Database (refseq_genomes) limited to bacterial
genomes. The search was performed by qblast (Biopython), using blastn, the maximum hitlist
size was set to 5000 and the expect threshold (e-value) was set to 500.
Primer pairs were tested for specificity using online Primer-BLAST. The primers were blasted
against the nucleotide collection BLAST database (nr) limited to sequences from bacteria. The nr
(non-redundant nucleotide) database was chosen to get the broadest coverage for the BLAST
search. Default settings were used, except for the primer specificity stringency that was set to
ignore targets that have nine or more mismatches to the primer.

242   *In vitro* **validation**

243   The inclusivity of the primer pairs was assayed by performing qPCR with 2 ng DNA of 21 to 25

244   strains of the target species in technical duplicates. The PCR efficiency was examined by ten-

245   fold dilution series of the type strain DNA in a range from 10^6 to 10^1 genome copies per reaction.

246   DNA concentration for the corresponding number of genome copies was estimated by taking the

247   genome size of the type strain (https://www.ncbi.nlm.nih.gov/genome) and an average weight of

248     $1.096 \cdot 10^{-21}$ g per base pair.

249   The exclusivity of the primer pairs was assayed by performing qPCR in technical duplicates with

250   2 ng DNA from various bacterial species found in dairy products. Because the number of

251   samples per run was limited, four separate runs were required to measure all non-target strains.

252   In each run three strains of the target species (positive control) and a no template control were

253   included.

254   **Bacterial strains**

255   Strains stored within the Agroscope Culture Collection at -80 °C in sterile reconstituted skim

256   milk powder (10 %, w/v), were reactivated and cultivated according to the conditions specified

257   in Dataset S2.

258   **DNA extraction**

259   Unless otherwise noted, all reagents were purchased from Merck, Darmstadt, Germany.

260   Bacterial pellets harvested from 1 ml culture by centrifugation (10000 x g, 5 min, room

261   temperature) were used for DNA extraction. For a pre-lysis treatment, the bacterial cells were

262   incubated in 1 ml of 50 mM sodium hydroxide for 15 min at room temperature. Afterwards cells

263   were collected by centrifugation (10000 x g, 5 min, room temperature) and then treated with

264   lysozyme (2.5 mg/ml dissolved in 100 mM Tris(hydroxymethyl)aminomethane, 10 mM

265   ethylendiaminetetraacetic acid (EDTA; Calbiochem, San Diego, USA), 25 % (w/v) sucrose, pH

266   8.0) for 1 hour at 37 °C. After the pre-lysis treatment, the bacterial cells were collected by

267   centrifugation (10000 x g, 5 min, room temperature). Cell lysis and genomic DNA extraction

268   was performed using the EZ1 DNA Tissue kit and a BioRobot® EZ1 workstation (Qiagen,

269   Hilden, Germany) according to the manufacturer's instructions and eluted in a volume of 100 µl.

270   The DNA concentration was measured using a NanoDrop® ND-1000 Spectrophotometer

271   (NanoDrop Technologies, Thermo Fisher Scientific, Waltham, MA, USA).

272   **Quantitative real-time PCR**

273   The qPCR assays were performed in a total reaction mix volume of 12 µl containing 6 µl 2x

274   SsoFast™ EvaGreen® Supermix with low ROX (Biorad, Cressier, Switzerland), 500 nM of

275   forward and reverse primers, respectively, and 2 µl of DNA. Each sample was measured in

276   technical duplicates. The qPCR cycling conditions were an initial denaturation at 95 °C for 1

277   minute followed by 35 cycles of 95 °C for 5 seconds and 60 °C for 1 minute. For the melting

278   curve analysis, a gradient from 60 – 95 °C with 1 °C steps per 3 seconds was performed. All

279   qPCR assays were run on a Corbett Rotor-Gene 3000 (Qiagen). The analysis was performed

280   using Rotor-Gene 6000 Software 1.7 with dynamic tube normalization and a threshold of 0.05

281   for quantification cycle (Cq) value calculation, the five first cycles were ignored for the

282   determination of the Cq values. The peak calling threshold for the melt curve analysis was set to

283   -2 dF/dT and a temperature threshold was set 2 °C lower than the positive control peak.

284   **Phylogeny and average nucleotide identity calculations**

285   The phylogeny was created with Roary and FastTree 2 during the pipeline runs and iTOL

286   (Letunic & Bork 2019) was used to visualize the tree. Average nucleotide identity (ANI)

287   calculations were performed with pyani v0.2.9 (Pritchard et al. 2016) using the ANIm method.

288   The heatmap was created from the ANIm_percentage_identity.tab output file using the

289   clustermap function of the python seaborn module and modified color bar settings from pyani.

290   For the color bars on top and on the left of the heatmap, the assemblies were assigned to the

291   same color as in the phylogeny tree. Row and column names (genome assembly accessions) can

292   be found in Dataset S3.

293   **Comparison of primer design pipelines**

294   The positive genome sets for RUCS and fdp were the same genome assemblies used for the

295   SpeciesPrimer pipeline. SpeciesPrimer uses by default the NCBI nt database and the species list

296   for the specificity checks, whereas RUCS and fdp require a negative set of genomes. Therefore, a

297   set of (representative) genome assemblies from NCBI was downloaded for the species from the

298   species list. From these assemblies a BLAST database was prepared for SpeciesPrimer. The

299   same genome assemblies, excluding the assembly of the target species, were used as a negative

300   set for RUCS and to make a BLAST database for fdp. For both tools, the minimal and maximal

301   PCR product size was set to 70 and 200, respectively. The tab separated config file for fdp was

302   created using the assembly accession as name, the species as class and providing the absolute

303   path of the genome assembly files. The script was started with the blastdb option to provide the

304   path to the previously prepared BLAST database with the non-target genome sequences. For

305   RUCS the entry point full was selected and the annotation of the target sequences was omitted.

306   SpeciesPrimer was configured to run with the custom BLAST database, without a species list

307   and the download and annotation step for the genome assemblies was omitted to provide

308   comparable running conditions. The accessions of the input genome assemblies and the

309   commands used can be found in Dataset S4. Primers used for a specificity check using Primer-

310   BLAST (nr database limited to sequences from bacteria) were the two primer pairs with the best

311   score in the results_best.tsv files (RUCS), the two best ranked primer pairs for SpeciesPrimer

312   and the primers reported in the universal_primers.eprimer3 files (fdp).

Results

Primer design

315   The SpeciesPrimer pipeline runs were completed in two to eight hours, excluding the time

316   required for downloading and annotation of the genome assemblies. Depending on the number of

317   genome assemblies, downloading and annotation of the genome assemblies took from 24

318   minutes (27) to 12 hours 27 minutes (575). The average time for downloading and annotation of

319   single assemblies was two seconds and one minute six seconds, respectively. On the consumer

320   laptop using a wireless LAN Internet connection the time required for the downloads has

321   doubled, while the annotation took 1.8 times longer. The pipeline runs took in total three times

322   longer, they were completed in seven to 29 hours. The analysis of the *Enterococcus faecalis*,

323   *Enterococcus faecium*, *Pediococcus acidilactici* and *Pediococcus pentosaceus* assemblies

324   resulted in 15, 2, 2 and 160 identified primer pair candidates, respectively (Table 3). The primer

325   pair candidates for *E. faecalis* and *P. pentosaceus* were filtered for the highest primer pair

326   coverage score (*E. faecalis*: 2*; P. pentosaceus*: 29); for *P. pentosaceus*, only the two primer pairs

327   with the lowest primer pair penalty values were selected.

328   The phylogeny tree from the concatenated core genes of *E. faecium* shows the phylogenetic

329   distance of two distinct groups of sequences, a main cluster with 531 sequences and a subcluster

330   with 44 sequences (Figure 2). The tree made with the concatenated core gene sequences of *P.*

331   *acidilactici* shows the phylogenetic distance of one sequence from all other sequences (Figure 3).

332   From this observation, the existence of different taxonomic units was suspected. Calculation of

333   the average nucleotide identity (ANI) has been proposed as a valuable tool to determine species

334   boundaries (Richter & Rossello-Mora 2009). Therefore, we performed ANI calculations for the

335   genome assemblies and displayed the results in a clustered heatmap (Figure 4). All genome

336   assemblies show an alignment coverage of at least 60 % to each other (Dataset S3), indicating

337   they are correctly assigned at the genus level. The clustering of the *E. faecium* genome

338   assemblies in Figure 4 A shows two distinct clusters corresponding to the clusters in the

339   phylogenetic tree (Figure 2). The assemblies of the two clusters have ANIm values at the border

340   of the species threshold cutoff as depicted by the white to light purple colored cells. Clustering of

341   the *P. acidilactici* genome assemblies in Figure 4 B shows three distinct clusters corresponding

342   to the clusters in the phylogenetic tree (Figure 3). The purple cells indicate that the assemblies of

343   two larger clusters belong to the same species, while the assembly with the orange color bar has

344   ANIm percentage identity values below the proposed species threshold cutoffs (95 - 96 %) (Kim

345   et al. 2014; Richter & Rossello-Mora 2009) as indicated by the blue cells. *P. acidilactici* strain

346   FAM 18987 should therefore probably be assigned to a new species or subspecies. However, for

347   certain species lower boundary cutoffs might be reasonable (Ciufo et al. 2018). According to the

348   current taxonomic classification, we proceeded with the assumption that these genome

349   assemblies reflected the actual diversity of strains and thus included the assemblies for the

350   primer design.

351   Two test cases were generated to exemplify the influence of the input genome assemblies on the

352   pipeline results. Firstly, a single genome assembly with a wrong taxonomic label was used as

353   input in addition to the correctly labelled genome assemblies. Introducing a genome assembly

354   with a wrong taxonomic label (GCF_000415325.2, *E. faecalis*) into the pool of *E. faecium*

355   genome assemblies resulted in a decrease of identified core genes (from 1131 to 43) and

356   provided no species-specific sequence. Secondly, the genome assembly of the *P. acidilactici*

357   strain (FAM 18987) that was distinct from the other assemblies in the phylogenetic tree and had

358   ANI values below the species threshold cutoff was excluded from the pipeline run. This resulted

359   in an increased number of identified core genes (from 921 to 1238), of species-specific

360   sequences (from 54 to 516) and of reported primer pairs (from 2 to 53).

361   ***In silico* validation**

362   Two parameters were selected as criteria for the primer validation using web-based BLAST.

363   First, the BLAST hits for the predicted PCR product sequence should only match the target

364   species. If sequences of other bacterial species matched to parts of the sequence, the

365   corresponding primer pairs were discarded, unless more than three mismatches were found in

366   each primer-binding region for the forward and reverse primers. Second, the primer binding sites

367   in the target sequences were not allowed to have mismatches in the 3'-end region. The criterion

368   for the primer validation by Primer-BLAST was that no predicted PCR products for other

369   bacterial species had been reported by Primer-BLAST. Primer pairs exclusively binding to the

370   target sequence of the target species were classified as specific. The results of the *in silico*

371   validation are summarized in Table 4. With the exception of Ec_faeca_g3060_1_P0 and

372   Ec_faeci_cysS_3_P1, all primer pairs showed a perfect match to their target sequences. For

373   primer pair Ec_faeca_g3060_1_P0, the first three nucleotides of one sequence out of 690 are

374   missing in the forward primer-binding region. For Ec_faeci_cysS_3_P1, only one sequence out

375   of 1058 aligned sequences showed a single nucleotide transition in the reverse primer-binding

376   region (Dataset S5, page 2-3).

377   ***In vitro* validation**

378   The specificity of the qPCR assays was assessed with 21 to 25 strains of the target species

379   (inclusivity) and 120 non-target bacterial strains found in dairy products (exclusivity). The qPCR

380   assay performance was assessed by 10-fold dilution series of type strain DNA from 10^6 to 10^1

381   copies per reaction. The results of the qPCR runs were interpreted as positive if both qPCR

382   reactions (duplicates) reached the fluorescent threshold before quantification cycle 35 and the

383   peak of the melting curve analysis was above the peak calling threshold (-2 dF/dT). A summary

384   of the results is shown in Table 5. The primer sequences can be found in Table S2. The

385   inclusivity of the qPCR assays was 100 % for the assays Ec_faeca_acuI, Ec_faeca_g3060,

386   Ec_faeci_cysS, Pd_acidi_asnS, Pd_acidi_g1164, Pd_pento_nagK and Pd_pento_g4364. Only

387   one qPCR assay, Ec_faeca_purD was negative for one of the tested target strains.

388   Out of the 120 non-target strains analyzed to determine the exclusivity of the qPCR assays

389   (Figure 5), all strains were negative for Ec_faeca_acuI and Pd_acidi_asnS. The assay

390   Pd_pento_nagK targeting *P. pentosaceus* was positive for two out of three tested *Leuconostoc*

391 *lactis* strains, the fluorescence signal reached the threshold after Cq 26, and the melting curve

392 analysis showed a peak at 85 °C, while the positive control samples for this assay displayed a

393 peak at 83.5 °C. Nine out of the 120 non-target strains were positive for the Ec_faeca_g3060

394 qPCR assay, for these samples the fluorescence signals reached the threshold after Cq 26 and had

395 a melting curve peak at a higher temperature than the target PCR product. The assays

396 Pd_acidi_g1164 and Pd_pento_g4364 were positive for five and eight non-target strains,

397 respectively. Notably, all three tested *Lactobacillus paracasei* strains were positive for the

398 Pd_acidi_g1164 assay, the fluorescence signal reached the threshold around Cq 21 and 22 and

399 they showed a distinct melting curve peak at 86 °C.

400 The calculated efficiency of the qPCR assays was between 92 and 100 %. The linear regression

401 equations ($Cq = slope * \log(copies) + intercept$) had slopes between -3.329 and -3.523 and

402 correlation coefficients of 0.990 or above. Dataset S6 contains the qPCR raw data and Dataset S7

403 a summary of the qPCR data.

404 **Comparison of primer design pipelines**

405 The running times and the number of reported primer pairs of RUCS, the fdp pipeline, and

406 SpeciesPrimer were compared. The download and annotation times were not considered since

407 RUCS and fdp do not include this feature. RUCS and fdp were both able to design primer pairs

408 for all four bacterial targets. The runs with RUCS were completed in two hours and 11 minutes

409 to five hours and 20 minutes and between 107 and 629 primer candidates were reported. The

410 specificity check using online Primer-BLAST showed that the best-ranked primer pair for each

411 of the targets was specific and perfectly matched to the primer binding region for all targets in

412 the nr database. The fdp runs were completed in 2 minutes to 17 hours 44 minutes. Three primer

413 pairs were reported for *E. faecium* and *P. acidilactici* and six primer pairs were reported for

414 *E. faecalis* and *P. pentosaceus*. Primer-BLAST results indicate that the best primer pairs for all

415 target species are specific. The best primer pair for *P. acidilactici* showed a two-nucleotide

416 mismatch in the primer binding region of one target sequence. A one-nucleotide mismatch in the

417 primer binding region of one target sequence was also observed in the primer pair for *E. faecalis*

418 (Dataset S4, Primer-BLAST summary). The results of the SpeciesPrimer runs differ from the

419 runs with the nt BLAST database presented in detail above. For the Enterococci the best reported

420 primer pairs remain the same, while different primer pairs were ranked best for the Pediococci.

421   In summary, all pipelines were able to design species-specific primers for all of the four target

422   species using the given input sequences. The results of the comparison are summarized in Table

423   6.

Discussion

425   After setup of the SpeciesPrimer Docker container, the download of the local BLAST database

426   and the selection of the SpeciesPrimer run settings, no further manual handling was required to

427   get primer pair candidates for all four bacterial species after a total time of 44 hours and 30

428   minutes (high end desktop). The number of input genomes and subsequently the number of

429   retrieved primer pairs for the specificity check have the highest impact on speed. During the

430   specificity check, blasting the primer sequences optimized for short sequences (blastn-short) and

431   the subsequent compilation and indexing of the non-target sequence database are the most time

432   consuming steps.

433   The results of the SpeciesPrimer pipeline for the four target species ranged from two to 160

434   identified primer pair candidates. Several factors can influence the number of identified primer

435   pairs, such as the quality of the input genome assemblies, assemblies with wrong taxonomic

436   labels and the genetic diversity within the species. A low-quality assembly with missing

437   sequences or contaminations can decrease the number of identified core genes. The initial quality

438   control helps to minimize the risk that such assemblies are included in the pipeline runs.

439   However, also an increased sequence diversity, either due to sequencing errors, assembly errors

440   or real diversity, limits the number and the length of identified conserved sequences.

441   Subsequently this affects the yield of reported primer pairs, since the pipeline selects highly

442   conserved sequences for primer design. The two test cases designed to exemplify the influence

443   of the input genome assemblies on the pipeline results illustrate that the SpeciesPrimer pipeline

444   performs best on closely related (same species) genome assemblies with a good overall quality.

445   The specificity of the designed primers was evaluated *in silico* by BLAST with a more extensive

446   database (RefSeq Genome) than the one used for the specificity check during primer design. The

447   validation showed that the specificity of the tested amplicons was high and no other species than

448   the target species had an identical sequence. Most target sequences in the database showed a

449   perfect match for the primers in the primer-binding region. For all tested primer pairs, only the

450   expected PCR products for the target species and no amplicons for other sequences were

451   predicted by Primer-BLAST. The results of Primer-BLAST indicate that the reported primer

452   pairs were very specific, even though the species list used for the specificity evaluation during

453   primer design covered only 259 non-target species.

454   In this work, 21 to 25 target strains for each target species and 120 non-target strains have been

455   tested to assess inclusivity and exclusivity of the qPCR assays, respectively. The *in vitro*

456   validation of primer pairs has shown that the *in silico* validation is not always able to predict

457   non-target PCR products. The fluorescence signals occurring at late quantification cycles (Cq >

458   30) are probably due to PCR products with suboptimal primer binding. Testing the qPCR assays

459   in mixtures and communities could be interesting to assess if these PCR products also

460   accumulate in presence of target DNA. The specificity could be sufficient in mixtures due to

461   competition for the primers and the difference in primer binding and amplification efficiency.

462   For many research applications, qPCR assays with a low signal in negative samples are

463   acceptable, assuming that low-level signals can be distinguished from low concentrations of

464   target species DNA by the melting curve analysis (Ririe et al. 1997). Furthermore, for many

465   applications, the annealing temperature can be optimized by empirical determination of a

466   suitable annealing temperature and the primer concentration can be adjusted to improve the

467   specificity of the assay (Bio-Rad Laboratories 2019). We did not try to optimize our assays with

468   these measures, because the aim was to design primers for high-throughput qPCR, requiring the

469   exact same PCR conditions. For the tested qPCR conditions, the most specific qPCR assays were

470   Ec_faeca_acuI (*E. faecalis*), Ec_faeci_cysS (*E. faecium*), Pd_acidi_asnS (*P. acidilactici*) and

471   Pd_pento_nagK (*P. pentosaceus*). Further work will be necessary in order to make these qPCR

472   assays fully operational for the quantification of bacteria in a complex system such as food. For

473   instance, suitable qPCR standards should be designed and validated, so that the limit of detection

474   of each assay can be determined (Forootan et al. 2017).

475   Primer-BLAST, fdp and RUCS allow designing primers for different applications, but demand

476   experience and manual manipulations. Primer-BLAST designs primers and performs specificity

477   checks, but requires a target sequence provided by the user. In the case of RUCS, manual

478   manipulation and some experience is needed to prepare the positive and negative reference sets.

479   The same applies to fdp and the results from the comparison indicate that fdp has its strength in

480   the identification of strain-specific primer pairs and for subsets of the positive set as implied in

481   the name. The observed mismatches in the primer-binding region are probably due to the

482   alignment-free approach the pipeline uses. This seems to drastically increase the speed, but it is

483 not taking into account the conservation of the target sequence and therefore mismatches, e.g.

484 due to single nucleotide polymorphisms (SNPs), can be found in the primer binding region. For a

485 large number of input assemblies, e.g. for *E. faecalis* (575), fdp requires distinctively more time

486 to run, which is a known issue caused by the cross-validation prediction step using PrimerSearch

487 (Pritchard et al. 2012).

488 Compared to primer-BLAST and RUCS, the task SpeciesPrimer performs is really specialized,

489 namely to design primers for species-specific sequences. In contrast, SpeciesPrimer requires no

490 previous knowledge about the input genome assemblies and no manual manipulation of

491 sequences has to be performed. The ability of SpeciesPrimer to run on standard computers with

492 good performance instead of specialized high-performance computers should allow primer

493 design for the wider range of scientists. Docker containers simplify the installation procedure and

494 should allow non-bioinformaticians to setup and use the SpeciesPrimer pipeline.

Conclusions

496 In this work, we presented the SpeciesPrimer pipeline, which is a fully automated pipeline from

497 the download of bacterial genomes, the identification of conserved species-specific core genes to

498 primer design and subsequent quality control of primer candidates. Primers for four bacterial

499 species were designed and validated and have shown to perform adequately under the same

500 qPCR conditions.

501 A standard computer with good performance, good quality genome assemblies, a local copy of

502 the nt BLAST database and a list of non-target bacterial species are the only requirements for

503 primer design with SpeciesPrimer. A complete image of a Linux OS with all dependencies and

504 the pipeline scripts is available from Dockerhub. To simplify primer design for users not familiar

505 with command line tools, a graphic user interface is provided in the latest version of

506 SpeciesPrimer. SpeciesPrimer facilitates efficient primer design for species-specific

507 quantification, paving the way for a fast and accurate quantitative investigation of microbial

508 communities.

Acknowledgements

510 We would like to thank Marco Meola and Remo Schmidt for critically reading the manuscript

511 and many helpful comments and Daniel Marzohl, Nadine Sidler, Elvira Wagner and Kotchanoot

512 Srikham for their valuable technical help.

References

Almeida M, Hébert A, Abraham A-L, Rasmussen S, Monnet C, Pons N, Delbès C, Loux V, Batto J-M, Leonard P, Kennedy S, Ehrlich SD, Pop M, Montel M-C, Irlinger F, and Renault P. 2014. Construction of a dairy microbial genome catalog opens new perspectives for the metagenomic analysis of dairy fermented products. *BMC Genomics* 15:1101. 10.1186/1471-2164-15-1101

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403-410. 10.1016/s0022-2836(05)80360-2

Ciufo S, Kannan S, Sharma S, Badretdin A, Clark K, Turner S, Brover S, Schoch CL, Kimchi A, and DiCuccio M. 2018. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int J Syst Evol Microbiol* 68:2386-2392. 10.1099/ijsem.0.002809

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, and de Hoon MJ. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422-1423. 10.1093/bioinformatics/btp163

Cremonesi P, Pisani LF, Lecchi C, Ceciliani F, Martino P, Bonastre AS, Karus A, Balzaretti C, and Castiglioni B. 2014. Development of 23 individual TaqMan(R) real-time PCR assays for identifying common foodborne pathogens using a single set of amplification conditions. *Food Microbiol* 43:35-40. 10.1016/j.fm.2014.04.007

Curran T, Coyle PV, McManus TE, Kidney J, and Coulter WA. 2007. Evaluation of real-time PCR for the detection and quantification of bacteria in chronic obstructive pulmonary disease. *FEMS Immunol Med Microbiol* 50:112-118. 10.1111/j.1574-695X.2007.00241.x

Elbrecht V, Leese F, and Bunce M. 2017. PrimerMiner: an R package for development and in silico validation of DNA metabarcoding primers. *Methods in Ecology and Evolution* 8:622-626. 10.1111/2041-210x.12687

Falentin H, Postollec F, Parayre S, Henaff N, Le Bivic P, Richoux R, Thierry A, and Sohier D. 2010. Specific metabolic activity of ripening bacteria quantified by real-time reverse transcription PCR throughout Emmental cheese manufacture. *Int J Food Microbiol* 144:10-19. 10.1016/j.ijfoodmicro.2010.06.003

Forootan A, Sjoback R, Bjorkman J, Sjogreen B, Linz L, and Kubista M. 2017. Methods to determine limit of detection and limit of quantification in quantitative real-time PCR (qPCR). *Biomol Detect Quantif* 12:1-6. 10.1016/j.bdq.2017.04.001

Garrido-Maestu A, Azinheiro S, Carvalho J, and Prado M. 2018. Rapid and sensitive detection of viable Listeria monocytogenes in food products by a filtration-based protocol and qPCR. *Food Microbiol* 73:254-263. 10.1016/j.fm.2018.02.004

Hermann-Bank ML, Skovgaard K, Stockmarr A, Larsen N, and Mølbak L. 2013. The Gut Microbiotassay: a high-throughput qPCR approach combinable with next generation sequencing to study gut microbial diversity. *BMC Genomics* 14:788. 10.1186/1471-2164-14-788

Ishii S, Segawa T, and Okabe S. 2013. Simultaneous Quantification of Multiple Food- and Waterborne Pathogens by Use of Microfluidic Quantitative PCR. *Appl Environ Microbiol* 79:2891-2898. 10.1128/aem.00205-13

Kim M, Oh HS, Park SC, and Chun J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* 64:346-351. 10.1099/ijs.0.059774-0

559   Kleyer H, Tecon R, and Or D. 2017. Resolving Species Level Changes in a Representative Soil
560   Bacterial Community Using Microfluidic Quantitative PCR. *Frontiers in Microbiology* 8.
561   10.3389/fmicb.2017.02017
562   Masco L, Vanhoutte T, Temmerman R, Swings J, and Huys G. 2007. Evaluation of real-time
563   PCR targeting the 16S rRNA and recA genes for the enumeration of bifidobacteria in
564   probiotic products. *International Journal of Food Microbiology* 113:351-357.
565   http://dx.doi.org/10.1016/j.ijfoodmicro.2006.07.021
566   Moyaert H, Pasmans F, Ducatelle R, Haesebrouck F, and Baele M. 2008. Evaluation of 16S
567   rRNA Gene-Based PCR Assays for Genus-Level Identification of Helicobacter Species. *J
568   Clin Microbiol* 46:1867-1869. 10.1128/jcm.00139-08
569   Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane
570   JA, and Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis.
571   *Bioinformatics* 31:3691-3693. 10.1093/bioinformatics/btv421
572   Postollec F, Falentin H, Pavan S, Combrisson J, and Sohier D. 2011. Recent advances in
573   quantitative PCR (qPCR) applications in food microbiology. *Food Microbiol* 28:848-861.
574   http://dx.doi.org/10.1016/j.fm.2011.02.008
575   Pritchard L, Glover RH, Humphris S, Elphinstone JG, and Toth IK. 2016. Genomics and
576   taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens.
577   *Analytical Methods* 8:12-24. 10.1039/c5ay02550h
578   Pritchard L, Holden NJ, Bielaszewska M, Karch H, and Toth IK. 2012. Alignment-free design of
579   highly discriminatory diagnostic primer sets for Escherichia coli O104:H4 outbreak
580   strains. *PLoS One* 7:e34498. 10.1371/journal.pone.0034498
581   Qu W, Zhou Y, Zhang Y, Lu Y, Wang X, Zhao D, Yang Y, and Zhang C. 2012. MFEprimer-2.0:
582   a fast thermodynamics-based program for checking PCR primer specificity. *Nucleic
583   Acids Res* 40:W205-208. 10.1093/nar/gks552
584   Ramirez M, Castro C, Palomares JC, Torres MJ, Aller AI, Ruiz M, Aznar J, and Martin-
585   Mazuelos E. 2009. Molecular detection and identification of Aspergillus spp. from
586   clinical samples using real-time PCR. *Mycoses* 52:129-134. 10.1111/j.1439-
587   0507.2008.01548.x
588   Rice P, Longden I, and Bleasby A. 2000. EMBOSS: The European Molecular Biology Open
589   Software Suite. *Trends in Genetics* 16:276-277. https://doi.org/10.1016/S0168-
590   9525(00)02024-2
591   Richter M, and Rossello-Mora R. 2009. Shifting the genomic gold standard for the prokaryotic
592   species definition. *Proc Natl Acad Sci U S A* 106:19126-19131.
593   10.1073/pnas.0906412106
594   Ririe KM, Rasmussen RP, and Wittwer CT. 1997. Product Differentiation by Analysis of DNA
595   Melting Curves during the Polymerase Chain Reaction. *Analytical Biochemistry* 245:154-
596   160. https://doi.org/10.1006/abio.1996.9916
597   Scheirlinck I, Van der Meulen R, De Vuyst L, Vandamme P, and Huys G. 2009. Molecular
598   source tracking of predominant lactic acid bacteria in traditional Belgian sourdoughs and
599   their production environments. *J Appl Microbiol* 106:1081-1092. 10.1111/j.1365-
600   2672.2008.04094.x
601   Shen Z, Qu W, Wang W, Lu Y, Wu Y, Li Z, Hang X, Wang X, Zhao D, and Zhang C. 2010.
602   MPprimer: a program for reliable multiplex PCR primer design. *BMC Bioinformatics*
603   11:143. 10.1186/1471-2105-11-143

604   Tange O. 2011. GNU Parallel - The Command-Line Power Tool. *login: The USENIX Magazine*
605   36:42-47.
606   Tao T, Madden T, Christiam C, and Szilagyi L. 2011. BLAST FTP Site. *Available at*
607   *https://www.ncbi.nlm.nih.gov/books/NBK62345/* (accessed September 24 2019).
608   Thomsen MCF, Hasman H, Westh H, Kaya H, and Lund O. 2017. RUCS: rapid identification of
609   PCR primers for unique core sequences. *Bioinformatics* 33:3917-3921.
610   10.1093/bioinformatics/btx526
611   Torriani S, Felis GE, and Dellaglio F. 2001. Differentiation of Lactobacillus plantarum, L.
612   pentosus, and L. paraplantarum by recA gene sequence analysis and multiplex PCR assay
613   with recA gene-derived primers. *Appl Environ Microbiol* 67:3450-3454.
614   10.1128/aem.67.8.3450-3454.2001
615   Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, and Rozen SG. 2012.
616   Primer3--new capabilities and interfaces. *Nucleic Acids Res* 40:e115. 10.1093/nar/gks596
617   Vijaya Satya R, Kumar K, Zavaljevski N, and Reifman J. 2010. A high-throughput pipeline for
618   the design of real-time PCR signatures. *BMC Bioinformatics* 11:340. 10.1186/1471-2105-
619   11-340
620   Wang LT, Lee FL, Tai CJ, and Kasai H. 2007. Comparison of gyrB gene sequences, 16S rRNA
621   gene sequences and DNA-DNA hybridization in the Bacillus subtilis group. *Int J Syst*
622   *Evol Microbiol* 57:1846-1850. 10.1099/ijs.0.64685-0
623   Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, and Madden TL. 2012. Primer-BLAST:
624   a tool to design target-specific primers for polymerase chain reaction. *BMC*
625   *Bioinformatics* 13:134. 10.1186/1471-2105-13-134
626   Zhu T, Liang C, Meng Z, Li Y, Wu Y, Guo S, and Zhang R. 2017. PrimerServer: a high-
627   throughput primer design and specificity-checking platform. *bioRxiv*:181941.
628   10.1101/181941
629   Zuker M, Mathews DH, and Turner DH. 1999. Algorithms and Thermodynamics for RNA
630   Secondary Structure Prediction: A Practical Guide. In: Barciszewski J, and Clark BFC,
631   eds. *RNA Biochemistry and Biotechnology*. Dordrecht: Springer Netherlands, 11-43.
632
633

Overview of the features of different primer design tools and pipelines.

Tool	RUCS	fdp	TOPSI	Species-Primer	Primer-Miner	Primer-Server	Primer-BLAST
Reference	(Thomsen et al. 2017)	(Pritchard et al. 2012)	(Vijaya Satya et al. 2010)	(this study)	(Elbrecht et al. 2017)	(Zhu et al. 2017)	(Ye et al. 2012)
Primer specificity	Bacterial strains / species			Bacterial species	Taxonomic groups	Input sequence	
Inputs							
Taxonomic group(s)	-	-	-	Species	Order, Family	-	-
Target gene(s)	-	-	-	-	x	-	x
Genome assemblies	x	x	x	x	-	-	-
Target sequences	-	-	-	-	x	x	x
Primer sequences	x	x	-	-	x	x	x
Automatic download of input sequences	-	-	-	x	x	-	-
Identification of target sequences	x	x	x	x	-	-	-
Identification of conserved regions	x	-	x	x	x	-	-
Primer design	x	x	x	x	-	x	x
Specificity check							
Target sequences	Input sequences	Input sequences	BLAST DB	BLAST DB	-	-	BLAST DB
Primer	Input sequences	BLAST DB	BLAST DB	BLAST DB	Alignment	BLAST DB	BLAST DB
Primer quality control	x	x	x	x	-	x	
Primer3 cutoffs	x	x	x	x	-	x	x
Primer dimer	-	-	-	x	-	-	-
Hairpin	-	-	-	-	-	-	-
Amplicon secondary structures	-	-	-	x	-	-	-
High-throughput primer design	x	x	x	x	-	x	-
Batch processing	-	-	-	Full runs	Download	-	-
Works on standard computers	x	x	-	x	x	x	-
Graphic user interface	x	-	-	x	-	x	x
Web service	x	-	*	-	-	x	x

1 fdp: find_differential_primers
2 x: Feature supported
3 - : Feature not supported
4 *: Access has to be requested
5 QC: Quality control
6 CDS: Coding sequences

Overview of the SpeciesPrimer pipeline workflow and the used software.

Pipeline workflow	Tools (Version*)	Reference
Input genome assemblies		
- download	NCBI Entrez (Biopython 1.73)	(Cock et al. 2009; Sayers 2009)
- annotation	Prokka (1.13.7)	(Seemann 2014)
- quality control	BLAST+ (2.9.0+)	(Altschul et al. 1990)
Core gene sequences		
- identification	Roary (3.12.0)	(Page et al. 2015)
- phylogeny	FastTree 2 (2.1.11)	(Price et al. 2010)
- selection of conserved sequences	Prank (.150803) consambig (EMBOSS 6.6.0.0) GNU parallel (20161222)	(Löytynoja 2014) (Rice et al. 2000) (Tange 2011)
- evaluation of specificity	BLAST+	(Altschul et al. 1990)
Primer		
- design	Primer3 (2.4.0)	(Untergasser et al. 2012)
- quality control	BLAST+, MFEprimer (2.0), MPprimer (1.5), Mfold (3.6)	(Altschul et al. 1990) (Qu et al. 2012) (Shen et al. 2010) (Zuker et al. 1999)

1   * Docker image June 13, 2019

Pipeline input and run statistics.

Two different computers were used to run the SpeciesPrimer pipeline depicted as high end desktop and consumer laptop. The high end desktop is running Ubuntu 16.04 in a virtual machine with two Xeon E5-2643 CPU's (22 logical processors), 32 GB RAM and a solid-state drive. The download of the genome assemblies was performed using a LAN connection. The consumer laptop is running on Ubuntu 16.04 with an i7-3610QM CPU (8 logical processors), 8 GB RAM and a solid-state drive. The download of the genome assemblies was performed using a wireless LAN connection.

Species	*E. faecalis*	*E. faecium*	*P. acidilactici*	*P. pentosaceus*
Pipeline input				
NCBI genomes	390	575	9	14
ACC genomes	0	0	118	13
Total genome assemblies	**390**	**575**	**127**	**27**
Download and annotation (h:min)				
High end desktop	9:04	12:27	1:55	0:24
Consumer laptop	15:56	22:18	3:10	0:42
Pipeline statistics				
Running time (h:min)				
High end desktop	6:11	8:05	1:55	4:25
Consumer laptop	19:52	28:56	6:59	6:47
Core genes	1375	1131	921	1341
Single copy core genes	632	563	641	889
Conserved sequences	1128	624	566	2782
Species-specific sequences	329	36	54	672
Potential primer pairs	89	4	7	632
Primer pairs after QC	**15**	**2**	**2**	**160**

1   QC: primer quality control, ACC: Agroscope Culture Collection

Summary of the *in silico* validation of the selected primer pairs.

Primer pair coverage (PPC) is a value used by MFEprimer-2.0 to score the ability of the primer pair to bind to a DNA template. The number of perfect matches of the primers to the primer binding region and the total number of target sequences are indicated in brackets.

Target species	Primer pair	PPC	BLAST (perfect/total)	Primer-BLAST (perfect/total)
E. faecalis	Ec_faeca_acuI_1_P0	100	specific (694/694)	specific (55/55)
	Ec_faeca_g3060_1_P0	100	specific (689/690)	specific (55/55)
E. faecium	Ec_faeci_cysS_3_P1	96.7	specific (1057/1058)	specific (148/148)
	Ec_faeci_purD_2_P0	93.3	specific(1083/1083)	specific (148/148)
P. acidilactici	Pd_acidi_asnS_2_P0	90.1	specific (19/19)	specific (11/11)
	Pd_acidi_g1164_1_P0	93.3	specific (23/23)	specific (11/11)
P. pentosaceus	Pd_pento_nagK_1_P0	100	specific (15/15)	specific (9/9)
	Pd_pento_g4364_1_P0	100	specific (15/15)	specific (9/9)

Summarized results of the *in vitro* validation of the selected qPCR assays.

Inclusivity: Number of positive DNA samples / total number of target species DNA samples.

Exclusivity: Number of DNA samples showing a fluorescence signal below quantification cycle 35 and a melting curve peak above the threshold / total number of non-target DNA samples.

Calculated efficiency, slope, intercept and correlation coefficient (R^2) of the linear regression equation.

Species	E. faecalis		E. faecium		P. acidilactici		P. pentosaceus	
Target gene	*acuI*	g3060	*cysS*	*purD*	*asnS*	g1164	*nagK*	g4364
Inclusivity	22/22	22/22	25/25	24/25	21/21	21/21	25/25	25/25
Exclusivity	0/120	9/120	0/120	0/120	0/120	5/120	2/120	8/120
Efficiency	98 %	97 %	92 %	97 %	99 %	100 %	94 %	92 %
Slope	-3.382	-3.387	-3.539	-3.396	-3.356	-3.329	-3.470	-3.523
Intercept	32.107	32.694	32.006	31.051	30.835	30.282	32.286	33.211
R²	0.998	0.997	0.990	0.996	0.997	0.995	0.996	0.997

Comparison of different primer design pipelines.

The hardware used to run the pipelines was a high end desktop running Ubuntu 16.04 in a virtual machine with two Xeon E5-2643 CPU's (22 logical processors), 32 GB RAM and a solid-state drive. For the Primer-BLAST results the number of perfect matches of the primers to the primer binding region and the total number of target sequences are indicated in brackets.

Target species	*Enterococcus faecalis*	*Enterococcus faecium*	*Pediococcus acidilactici*	*Pediococcus pentosaceus*
Running time (h:min)				
RUCS	5:17	5:20	2:11	3:16
fdp	8:32	17:44	0:36	0:02
SpeciesPrimer	5:23	6:53	1:35	1:11
Primer pairs				
RUCS	199	107	123	629
fdp	6	3	3	6
SpeciesPrimer	11	3	2	36
Primer-BLAST (perfect / total) Best primer pair				
RUCS	specific (59/59)	specific (153/153)	specific (11/11)	specific (9/9)
fdp	specific (58/59)	specific (153/153)	specific (10/11)	specific (9/9)
SpeciesPrimer	specific (59/59)	specific (153/153)	specific (11/11)	specific (9/9)

1   fdp: find_differential_primers

Figure 1

Schematic workflow of the database creation and the specificity check using
MFEprimer-2.0.
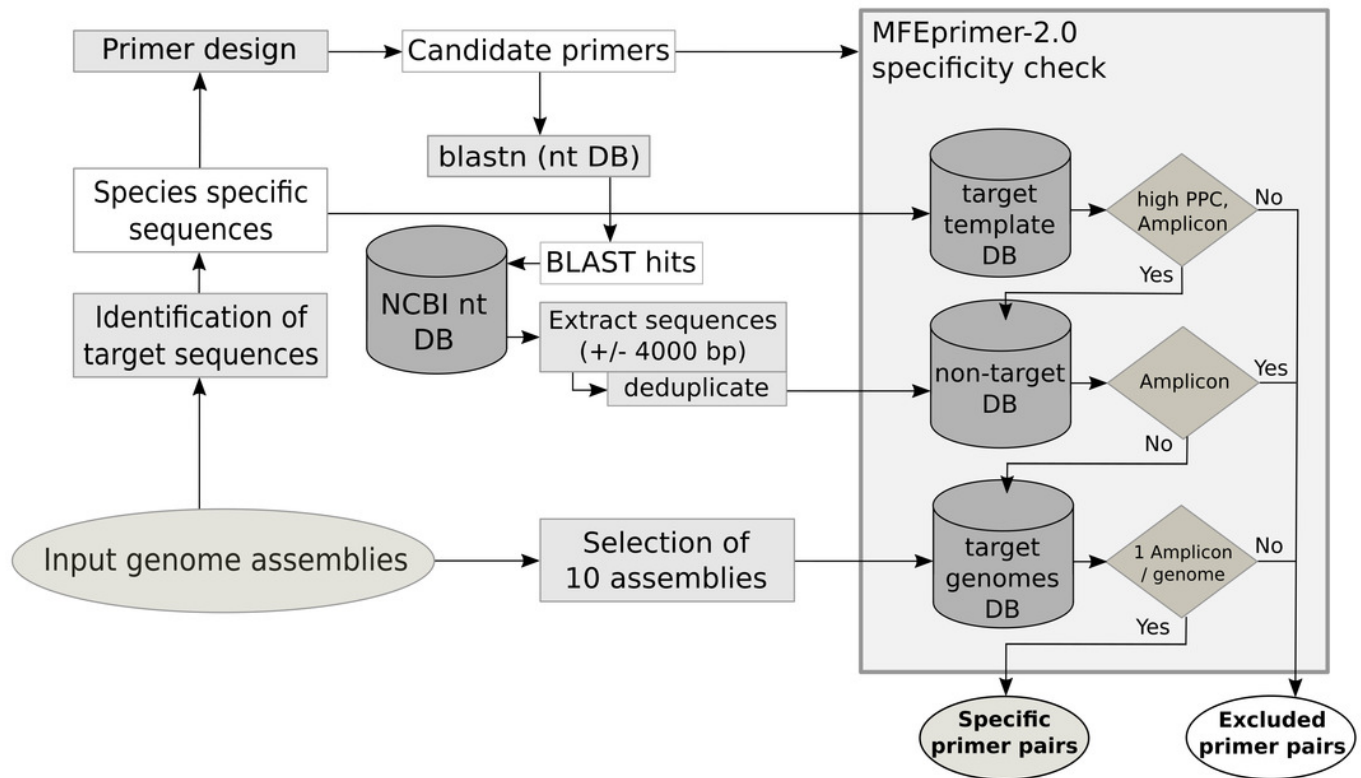
Figure 2

Phylogenetic tree based on the alignment of concatenated core genes of 575 *Enterococcus faecium* genome assemblies.

(A) The main cluster with 531 sequences is depicted in black and the subcluster of 44 sequences in blue. (B) Enlarged view of the tree structure and the subcluster.
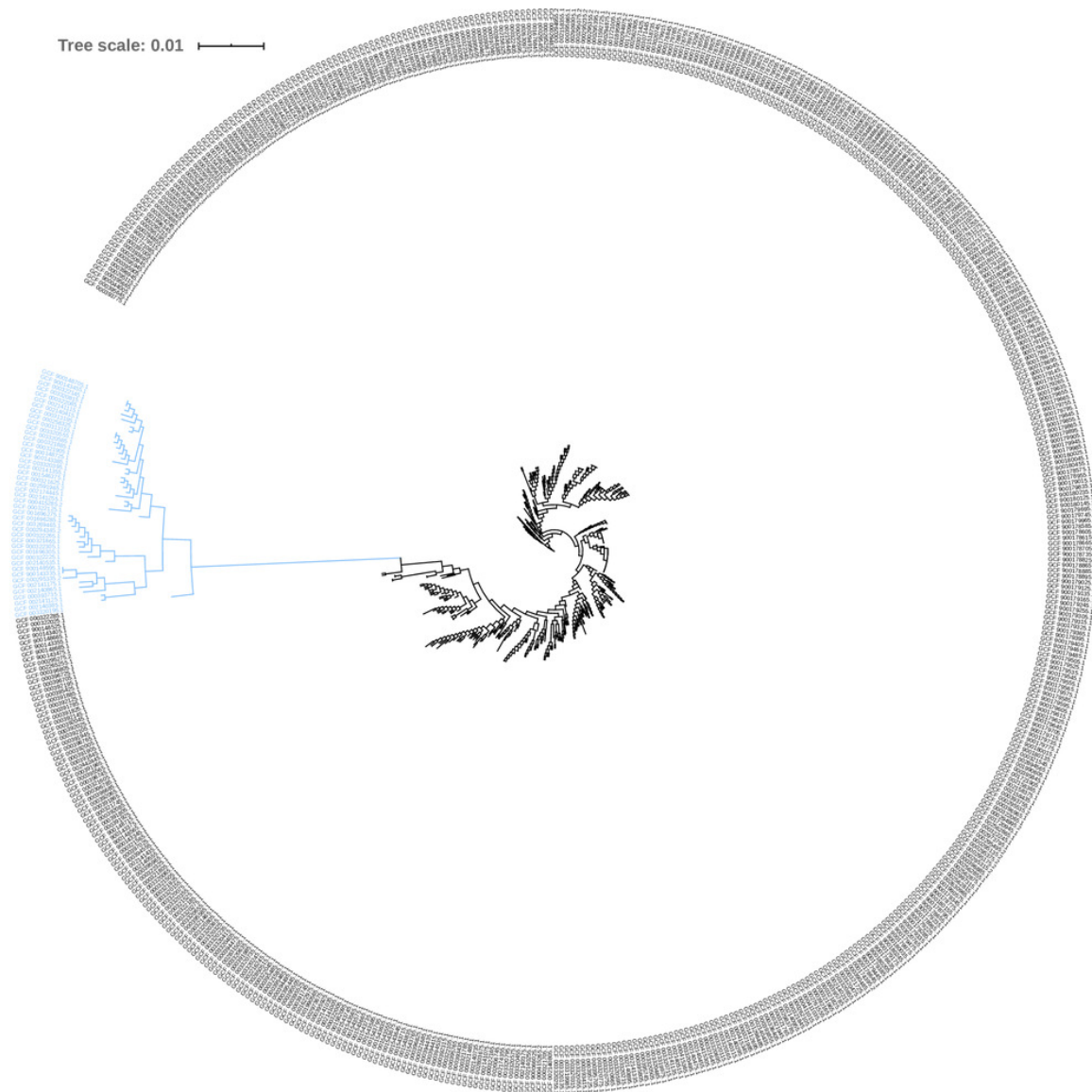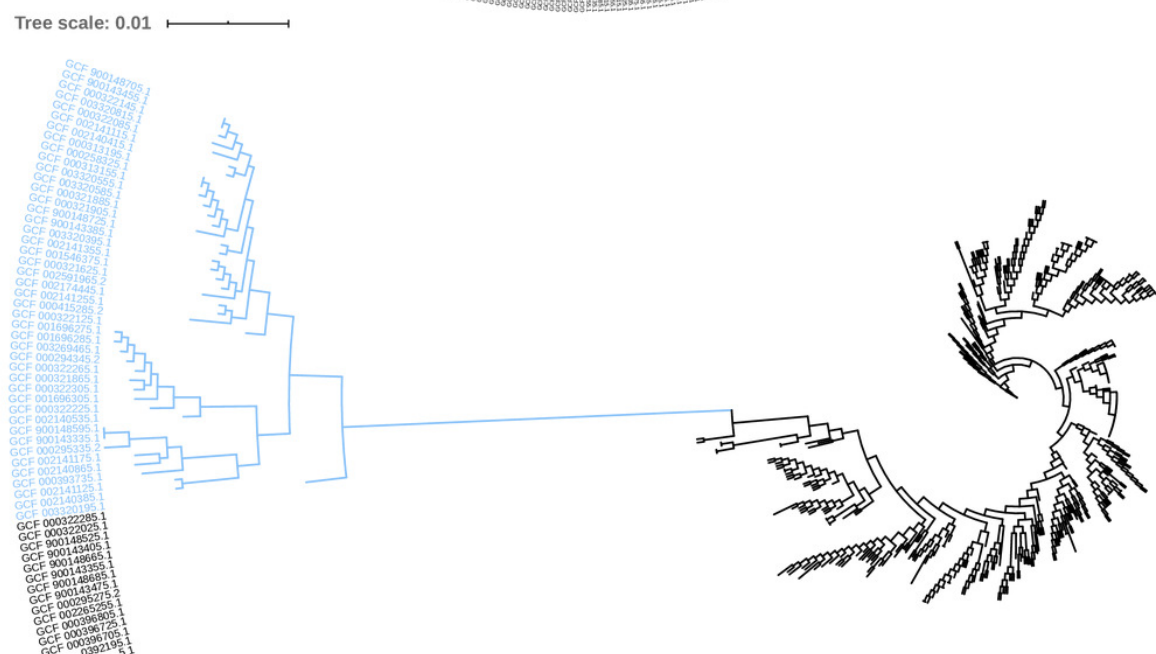
Figure 3

Phylogenetic tree based on the alignment of concatenated core genes of 127 *Pediococcus acidilactici* genome assemblies.

The main cluster with 100 sequences is depicted in black, a subcluster of 26 sequences in blue and the sequence with the largest phylogenetic distance in orange.

Tree scale: 0.01
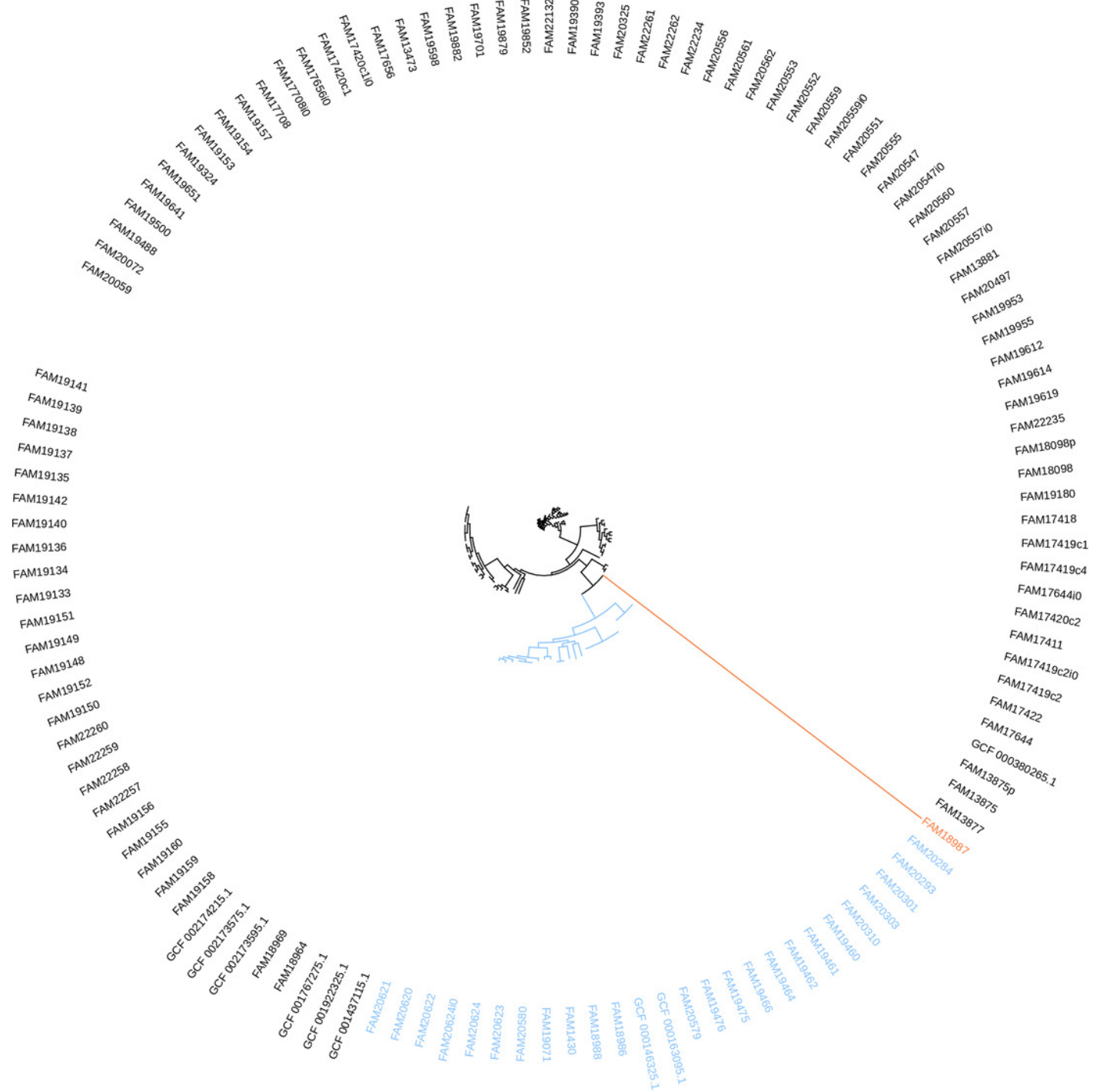
Figure 4

Clustered heatmap of ANIm percentage identity for (A) 575 *Enterococcus faecium* and (B) 129 *Pediococcus acidilactici* genome assemblies.

Purple colored cells in the heatmap correspond to ANIm percentage identity above 95 %, color intensity fades towards the proposed species threshold cutoff. Blue colored cells are below this threshold indicating that the corresponding genome assembly does not belong to the same species. Color bars on top and on the left of the heatmap correspond to the clusters and the colors indicated in the phylogenetic trees (Figure 2 and 3). Dendograms are based on single-linkage hierarchical clustering of the ANIm percentage identities. Row and column names can be found in Dataset S3.
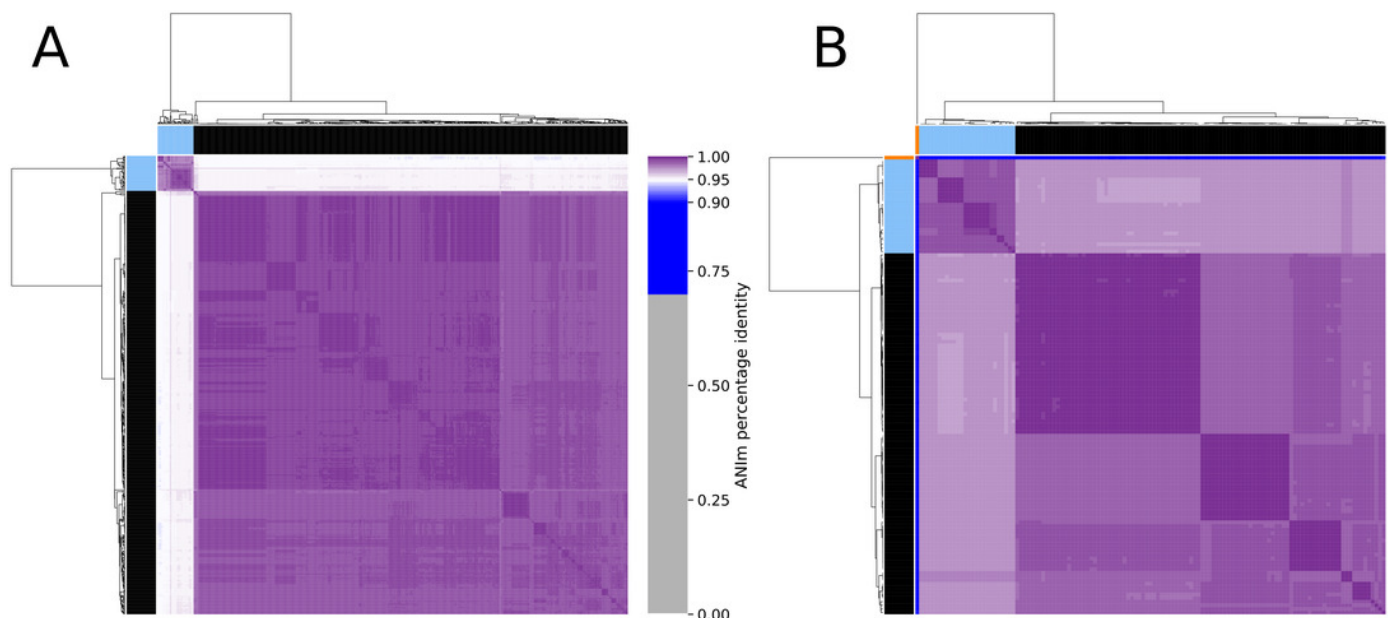
Figure 5

qPCR assay quantification cycle heatmap.

Depicted are all tested non-target strains and their average quantification cycle (technical duplicates) if the melt curve peak was above the threshold. The gray shades represent the Cq values from 10 to 35 (if no fluorescent signal was measured the value was set to Cq 35). Abbreviations: A.: *Acidipropionici*; Cl.: *Clostridium*; Lb.: *Lactobacillus*; Ln.: *Leuconostoc*; Pb: *Propionibacterium*; Pd.: *Pediococcus*; Sc.: *Streptococcus*; NTC: no template control.