A peer-reviewed version of this preprint was published in PeerJ on 18 February 2020.

View the peer-reviewed version (peerj.com/articles/8544), which is the preferred citable publication unless you specifically need to cite this preprint.

Dreier M, Berthoud H, Shani N, Wechsler D, Junier P. 2020. SpeciesPrimer: a bioinformatics pipeline dedicated to the design of qPCR primers for the quantification of bacterial species. PeerJ 8:e8544 https://doi.org/10.7717/peerj.8544

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

Matthias Dreier [Corresp., 1, 2], **Hélène Berthoud** [1], **Noam Shani** [1], **Daniel Wechsler** [1], **Pilar Junier** [2]

[1] Agroscope, Bern, Switzerland

[2] Laboratory of Microbiology, University of Neuchâtel, Neuchâtel, Switzerland

Corresponding Author: Matthias Dreier
Email address: matthias.dreier@agroscope.admin.ch

Background. Quantitative real-time PCR (qPCR) is a well-established method for detecting and quantifying bacteria, and it is progressively replacing culture-based diagnostic methods in food microbiology. High-throughput qPCR using microfluidics brings further advantages by providing faster results, decreasing the costs per sample and reducing errors due to automatic distribution of samples and reactants. In order to develop a high-throughput qPCR approach for the rapid and cost-efficient quantification of microbial species in a given system (for instance, cheese), the preliminary setup of qPCR assays working efficiently under identical PCR conditions is required. Identification of target-specific nucleotide sequences and design of specific primers are the most challenging steps in this process. To date, most available tools for primer design require either laborious manual manipulation or high-performance computing systems.

Results. We developed the SpeciesPrimer pipeline for automated high-throughput screening of species-specific target regions and the design of dedicated primers. Using SpeciesPrimer specific primers were designed for four bacterial species of importance in cheese quality control, namely *Enterococcus faecium*, *Enterococcus faecalis*, *Pediococcus acidilactici* and *Pediococcus pentosaceus*. Selected primers were first evaluated *in silico* and subsequently *in vitro* using DNA from pure cultures of a variety of strains found in dairy products. Specific qPCR assays were developed and validated, satisfying the criteria of inclusivity, exclusivity and amplification efficiencies.

Conclusion. In this work, we present the SpeciesPrimer pipeline, a tool to design species-specific primers for the detection and quantification of bacterial species. We use SpeciesPrimer to design qPCR assays for four bacterial species and describe a workflow to evaluate the designed primers. SpeciesPrimer facilitates efficient primer design for species-specific quantification, paving the way for a fast and accurate quantitative investigation of microbial communities.

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

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

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

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

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

Abstract

Background. Quantitative real-time PCR (qPCR) is a well-established method for detecting and quantifying bacteria, and it is progressively replacing culture-based diagnostic methods in food microbiology. High-throughput qPCR using microfluidics brings further advantages by providing faster results, decreasing the costs per sample and reducing errors due to automatic distribution of samples and reactants. In order to develop a high-throughput qPCR approach for the rapid and cost-efficient quantification of microbial species in a given system (for instance, cheese), the preliminary setup of qPCR assays working efficiently under identical PCR conditions is required. Identification of target-specific nucleotide sequences and design of specific primers are the most challenging steps in this process. To date, most available tools for primer design require either laborious manual manipulation or high-performance computing systems.

Results. We developed the SpeciesPrimer pipeline for automated high-throughput screening of species-specific target regions and the design of dedicated primers. Using SpeciesPrimer specific

primers were designed for four bacterial species of importance in cheese quality control, namely *Enterococcus faecium*, *Enterococcus faecalis*, *Pediococcus acidilactici* and *Pediococcus pentosaceus*. Selected primers were first evaluated *in silico* and subsequently *in vitro* using DNA from pure cultures of a variety of strains found in dairy products. Specific qPCR assays were developed and validated, satisfying the criteria of inclusivity, exclusivity and amplification efficiencies.

Conclusion. In this work, we present the SpeciesPrimer pipeline, a tool to design species-specific primers for the detection and quantification of bacterial species. We use SpeciesPrimer to design qPCR assays for four bacterial species and describe a workflow to evaluate the designed primers. SpeciesPrimer facilitates efficient primer design for species-specific quantification, paving the way for a fast and accurate quantitative investigation of microbial communities.

Introduction

Quantitative real-time PCR (qPCR) is a well-established method for the detection and quantification of bacteria in microbiology, for instance in the context of pathogen detection in clinical and veterinary diagnostics and food safety (Cremonesi et al. 2014; Curran et al. 2007; Garrido-Maestu et al. 2018; Ramirez et al. 2009). Culture-based diagnostic methods are progressively being replaced by qPCR due to advantages such as faster results, more specific detection, and the ability to detect sub-dominant populations (Postollec et al. 2011). High-throughput microfluidic qPCR brings further advantages including the fast generation of results, a lower cost per sample and fewer errors due to automatic distribution of samples and reactants. However, in order to work efficiently high-throughput qPCR systems use identical PCR chemistry and PCR conditions for all reactions taking place on a single chip. Therefore, existing qPCR assays are often not suitable and new primers have to be designed (Hermann-Bank et al. 2013; Ishii et al. 2013; Kleyer et al. 2017).

The main challenges for the successful development of any qPCR assay are the identification of a specific target nucleotide sequence and the design of primers that bind exclusively to that target sequence. Before microbial draft genomes became widely available, the 16S rRNA gene sequence was frequently used as a target sequence. However, the regions that are targeted in the 16S rRNA gene do not provide sufficient resolution to differentiate between closely related bacterial species (Moyaert et al. 2008; Torriani et al. 2001; Wang et al. 2007). Further,

housekeeping genes such as, for instance, *tuf, recA* and *pheS*, were successfully used as target sequences for a variety of bacterial species in fermented foods (Falentin et al. 2010; Masco et al. 2007; Scheirlinck et al. 2009). Today, the steadily increasing number of prokaryotic draft genomes facilitates the identification of new and unique target regions.

Various commercial and open source programs facilitate the design of specific primers for a target sequence, such as the standard tools Primer3 and Primer-BLAST (Untergasser et al. 2012; Ye et al. 2012). Primer3 predicts suitable PCR primers for an input target sequence, while Primer-BLAST combines Primer3 with a BLAST search in a selected nucleotide sequence database to assess the specificity of the primers for the target sequence. Additional tools and pipelines that encompass both the identification of target sequences from bacterial draft genomes and the design of primer candidates include, for instance, RUCS and TOPSI (Thomsen et al. 2017; Vijaya Satya et al. 2010). RUCS is able to identify unique core sequences in a positive set of genomes (target) compared to a negative set of genomes (non-target). It designs primers for the core sequences and validates them with an *in silico* PCR validation method against the positive and negative reference set. TOPSI is an automated high-throughput pipeline for the design of primers, primarily developed for pathogen-diagnostic assays. It identifies sequences present in all input genomes and designs specific primers accordingly.

We aimed to design a series of primers that function with the same qPCR cycling conditions and primer concentrations for later usage in a high-throughput microfluidic qPCR platform. Although TOPSI and RUCS were initially considered for the automated design of primers, TOPSI could not be used because no Linux-based cluster was available. RUCS was easily installed, but we were not able to create primer pairs in initial tests with a small set of positive (target) and negative (non-target) genomes. The example in the original publication of RUCS (using *Escherichia coli* genomes as positive and negative sets) indicates that RUCS works best for very similar genome assemblies in the positive and the negative sets. From this example and the initial test, we inferred that RUCS requires a carefully selected training set of positive and negative genomes to identify target sequences, which is a demanding task in the case of complex microbial systems such as those involved in the production of fermented foods and was therefore not suitable for our high-throughput approach.

This study presents a pipeline for automated high-throughput screening for species-specific target regions combined with the design of primer candidates for these sequences. The process of

primer design is fully automated from the download of bacterial genomes to the quality control of primer candidates. The pipeline runs on a standard computer with a multi-core processor and a minimum of 16 GB RAM. We have applied the SpeciesPrimer pipeline to a set of four bacterial species occurring in cheese and other dairy products and validated the primers *in silico* and *in vitro* by performing qPCR experiments with a variety of target and non-target strains.

Description

Overview

The SpeciesPrimer pipeline consists of three main parts (Table 1). First, genome assemblies are downloaded, annotated and then subjected to quality control. Second, a pan-genome analysis is performed to identify single copy core genes. Conserved sequences of these core genes are then extracted and the specificity for the target species is assessed. Finally, primers are designed for these species-specific conserved core gene sequences and subsequently evaluated in a primer quality control step.

Part 1: Input genome assemblies

The minimal command line input for the pipeline is the species name. Further, a list of non-target species names can be specified (e.g., species found in the investigated ecosystem but that should not be detected in the specific qPCR assay). For downloading genome assemblies from the National Center for Biotechnology Information (NCBI) automatically, a valid e-mail address is required for accessing the NCBI E-utilities services (Sayers 2009). The pipeline works with a pre-formatted NCBI BLAST database (nt), containing partially non-redundant nucleotide sequences. A local copy of the nt database is required. It can be downloaded from NCBI using the update_blastdb.pl script from the BLAST+ package (Altschul et al. 1990), via FTP from the NCBI FTP server or with the pipeline script (getblastdb.py).

The user-provided species name is used to search for genome assemblies in the NCBI database. The Biopython Entrez module (Cock et al. 2009) searches the NCBI taxonomic identity (taxid) for the target species in the taxonomy database and downloads the genome assembly summary report. Afterwards, SpeciesPrimer downloads the genome assemblies in FASTA format from the NCBI RefSeq FTP server using the links specified in the summary report. Finally, the downloaded genome assemblies are annotated with Prokka (Seemann 2014).

The quality of the genome assemblies is a crucial factor for the pan-genome analysis. Genome assemblies deposited with the wrong taxonomic label or low-quality assemblies drastically reduce the number of identified core genes and of conserved sequences for primer design. The initial quality control step is intended to remove such assemblies from the subsequent analysis. For the verification of the taxonomic classification, the user can choose one or several genes from five conserved housekeeping genes (16S rRNA, *tuf, recA, dnaK* and *pheS*). Genome assemblies without an annotation for the specified conserved housekeeping genes or genome assemblies consisting of more than 500 contigs are removed from the downstream pan-genome analysis. The sequences of the specified conserved housekeeping genes are blasted against the local nt database. Genome assemblies pass the quality control if the best BLAST hit for all sequences is a sequence arising from the target species.

Part 2: Identification of target sequences for primer design

A pan-genome analysis is performed using Roary (Page et al. 2015) to identify the core genes of the target species. Based on the results of the pan-genome analysis, single copy core genes are identified. The gene_presence_absence.csv produced by Roary reports the presence (or absence) of every annotated gene for every input genome assembly. Single copy core genes are the genes for which the number of assemblies harboring the sequence and the number of total identified sequences equals the number of total input assemblies. An sqlite3 database containing all annotated sequences of all assemblies is compiled (https://github.com/EnzoAndree/tutorials/blob/patch-1/DBGenerator.py). This database is queried for single copy core genes and the nucleotide sequences are saved in multi-FASTA format. Each multi-FASTA file contains the sequences of one single copy core gene from each input genome assembly. These sequences are aligned using the probabilistic multiple alignment program Prank (Löytynoja 2014). A consensus sequence with ambiguous bases is then created using consambig from the EMBOSS package (Rice et al. 2000). The alignments and extraction of the consensus sequence are performed in parallel for several core genes using GNU parallel (Tange 2011). Continuous consensus sequences longer than the minimal PCR product length, harboring less than two ambiguous bases in the range of 20 bases are used for the subsequent steps of the pipeline.

These conserved consenus sequences are used for a BLAST search against the local nt database using the discontiguous BLAST algorithm and an e-value cutoff of 500. For all hits in the

BLAST results, the species name is extracted from the sequence description and compared with the names in the species list (non-target species). If any species name in the species list matches a hit in the BLAST results the corresponding query sequence is discarded, otherwise the sequence is classified as specific for the target and considered for primer design.

Part 3: Primer design

Primer3 is used to design primers for the unique single copy core gene sequences. As pipeline default the optimum primer melting temperature is set to 60 °C and the maximal primer length is set to 26 bases, all other settings are the default settings of the primer3web version (http://primer3.ut.ee, accessed November 29, 2018). The minimal and maximal amplicon size of the PCR product can be specified individually for every target species through the command line options. The other parameters for primer3 cannot be changed individually, but the general primer3 settings can be changed by modifying the primer3 settings file.

The primer quality control consists of three parts, an in silico PCR to evaluate the specificity of the primer for the template, an estimation of secondary structures of the amplicon sequence and the estimation of the potential to form primer dimers. The specificity check for each primer pair is performed with MFEprimer 2.0 (Qu et al. 2012). For the evaluation of the specificity, three indexed databases are generated: the target template database, the non-target sequence database and the target genome database. The target template database consists of the unique conserved core gene sequences used as template for primer design. The non-target sequence database is compiled from sequences of non-target species, which show similarities to the primer sequences. To identify these sequences, a BLAST search with all primers against the local nt database is performed. BLAST hits with a species name in the description matching a name in the user-specified non-target species list are selected. These selected sequences and 4000 base pairs up- and downstream are extracted from the nt database using the blastdbcmd tool. The target genome database is composed of maximal 10 of the input genome assemblies. If the assembly summary report from the automatic download of genome assemblies from NCBI is available the genome assemblies as complete as possible are preferred (assembly status: complete > chromosome > scaffold > contig). The target sequence database is used to evaluate the maximum primer pair coverage (PPC), a value used by MFEprimer 2.0 to score the ability of the primer pair to bind to a DNA template. The maximum value of the PPC is 100, all primer pairs with a PPC value lower than the specified threshold (mfethreshold, default = 90) for their template are excluded. Next,

MFEprimer 2.0 is used to score the binding of the primer pairs to the sequences of the non-target sequence and the target genome database. The difference of the PPC for the DNA template and the specified threshold (Δthreshold = PPC – mfethreshold) is used as a threshold for the maximum PPC value a primer pair is allowed to have for a non-target sequence. Strong secondary structures at the 5'- or the 3'- end of the PCR product could impair efficient primer binding. Therefore, the PCR products of the primer pairs are submitted to mfold (Zuker et al. 1999) to exclude PCR products with strong secondary structures at the annealing temperature of 60 °C. Moreover, as primer dimers can yield unspecific signals during the qPCR run, the 3'- ends of the primer pairs are checked for their potential to form homo- or hetero-dimers using a Perl script (MPprimer_dimer_check.pl) from MPprimer (Shen et al. 2010).

The pipeline output is a list containing the primer name, primer pair coverage (MFEprimer) and penalty values, primer and template sequences and melting temperatures (Primer3). Further, a report of the genome assembly quality control, a file containing the pipeline run statistics, the 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 and 32 GB of RAM. The used docker image is available from 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 (Data S1). Genome assemblies from the strain collection of Agroscope 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 (Ye et al. 2012). The primers were blasted against the nucleotide collection BLAST database (nr) limited to sequences from bacteria. 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.

In vitro validation

The inclusivity of the primer pairs was assayed by performing qPCR with 2 ng DNA of 20 to 25 strains of the target species in technical duplicates. The linear amplification of genomic DNA and PCR efficiency was examined by ten-fold dilution series of the type strain DNA in a range from 10^6 to 10^1 genome copies per reaction. DNA concentration for the corresponding number of genome copies was estimated by taking the genome size of the type strain (https://www.ncbi.nlm.nih.gov/genome) and an average weight of $1.096 \cdot 10^{-21}$ g per base pair. The exclusivity of the primer pairs was assayed by performing qPCR with 2 ng DNA from various bacterial species in technical duplicates found in dairy products in four qPCR runs including three strains of the target species as positive control.

Bacterial strains

Strains stored within the Agroscope Culture Collection at -80 °C in sterile reconstituted skim milk powder (10 %, w/v), were reactivated and cultivated according to the conditions specified in Data S2.

DNA extraction

Unless otherwise noted, all reagents were purchased from Merck, Darmstadt, Germany. Bacterial pellets harvested from 1 ml culture by centrifugation (10000 x g, 5 min, room temperature) were used for DNA extraction. For a pre-lysis treatment, the bacterial cells were incubated in 1 ml of 50 mM sodium hydroxide for 15 min at room temperature. Afterwards cells were collected by centrifugation (10000 x g, 5 min, room temperature) and then treated with

lysozyme (2.5 mg/ml dissolved in 100 mM Tris(hydroxymethyl)aminomethane, 10 mM ethylendiaminetetraacetic acid (Calbiochem, San Diego, USA), 25 % (w/v) sucrose, pH 8.0) for 1 hour at 37 °C. After the pre-lysis treatment, the bacterial cells were collected by centrifugation (10000 x g, 5 min, room temperature). Cell lysis and genomic DNA extraction was performed using the EZ1 DNA Tissue kit and a BioRobot® EZ1 workstation (Qiagen, Hilden, Germany) according to the manufacturer's instructions and eluted in a volume of 100 µl. The DNA concentration was measured using a NanoDrop® ND-1000 Spectrophotometer (NanoDrop Technologies, Thermo Fisher Scientific, Waltham, MA, USA).

Quantitative real-time PCR

The qPCR assays were performed in a total reaction mix volume of 12 µl containing 6 µl 2x SsoFast™ EvaGreen® Supermix with low ROX (Biorad, Cressier, Switzerland), 500 nM of forward and reverse primers, respectively, and 2 µl of DNA. Each sample was measured in technical duplicates. The qPCR cycling conditions were an initial denaturation at 95 °C for 1 minute followed by 35 cycles of 95 °C for 5 seconds and 60 °C for 1 minute. For the melting curve analysis, a gradient from 60 – 95 °C with 1 °C steps per 3 seconds was performed. All qPCR assays were run on a Corbett Rotor-Gene 3000 (Qiagen). The analysis was performed using Rotor-Gene 6000 Software 1.7 with dynamic tube normalization and a threshold of 0.05 for quantification cycle (Cq) value calculation, the five first cycles were ignored for the determination of the Cq values. The peak calling threshold for the melt curve analysis was set to -2 dF/dT and a temperature threshold was set 2 °C lower than the positive control peak.

Average nucleotide identity calculations

Average nucleotide identity (ANI) calculations were performed with OrthoANIu (Yoon et al. 2017)). All *Enterococcus faecium* genome assemblies were compared to the *E. faecium* reference sequence (NC_017960.1), while all *Pediococcus acidilactici* genome assemblies were compared to the *P. acidilactici* reference sequence (NZ_CP015206.1). The genome assemblies were grouped based on the phylogeny tree of the core gene sequences built with FastTree (Price et al. 2010), the ANI values were collected and the average, minimum and maximum was calculated for each group.

Results

Primer design

Primer design for four bacterial species commonly found in cheese was performed with the SpeciesPrimer pipeline. The pipeline runs were completed in two to eight hours, excluding the time required for downloading and annotation of the genome assemblies. Depending on the number of genome assemblies, downloading and annotation of the genome assemblies took from 24 minutes (27) to 12 hours 27 minutes (575). The average time for downloading and annotation was two seconds and one minute six seconds, respectively. The analysis of the *Enterococcus faecalis*, *Enterococcus faecium*, *Pediococcus acidilactici* and *Pediococcus pentosaceus* assemblies resulted in 15, 2, 2 and 160 identified primer pair candidates, respectively (Table 2). The primer pair candidates for *E. faecalis* and *P. pentosaceus* were filtered for the highest primer pair coverage score (*E. faecalis*: 2*; P. pentosaceus*: 29); for *P. pentosaceus* only the two primer pairs with the lowest primer pair penalty values were selected.

The phylogeny trees of the core gene alignments from *E. faecium* and *P. acidilactici* were created using Roary and FastTree (Figure 1). The unrooted tree from the concatenated core genes of *E. faecium* shows the phylogenetic distance of two distinct groups of sequences, a main cluster with 531 sequences and a subcluster with 44 sequences. The tree made with the concatenated core gene sequences of *P. acidilactici* shows the phylogenetic distance of one sequence from all other sequences. From this observation, the existence of different species or subspecies was suspected. Calculation of the average nucleotide identity (ANI) has been proposed as a valuable tool to determine species boundaries (Richter & Rossello-Mora 2009). Therefore, we performed ANI calculations for the genome assemblies and the reference sequence *for E. faecium* (NC_017960.1) using the tool OrthoANIu. The genome assemblies of the *E. faecium* subcluster have an average ANI of 94.67 %. The ANI between the genome assembly of the *P. acidilactici* strain FAM 18987 and the NCBI reference sequence for *P. acidilactici* (NZ_CP015206.1) was only 89.21 %. The OrthoANI values (Table 3) of the assemblies in the subclusters of *E. faecium* (94.15 - 95.60 %) are at the border and the value for the *P. acidilactici* strain FAM 18987 (89.21%) is below the proposed species threshold cutoffs (95 - 96 %) (Kim et al. 2014; Richter & Rossello-Mora 2009). *P. acidilactici* strain FAM 18987 should therefore probably be assigned to a new species or subspecies. However, for certain species also lower boundary cutoffs might be reasonable (Ciufo et al. 2018). According to the

current taxonomic classification, we proceeded with the assumption that these genome assemblies reflected the actual diversity of strains and thus included the assemblies for the primer design.

Two test cases were generated to exemplify the influence of the input genome assemblies on the pipeline results. Firstly, a single genome assembly with a wrong taxonomic label was used as input in addition to the correctly labelled genome assemblies. Introducing a genome assembly with a wrong taxonomic label (GCF_000415325.2, *E. faecalis*) into the pool of *E. faecium* genome assemblies resulted in a decrease of identified core genes (from 1131 to 43) and provided no species-specific sequence. Secondly, the genome assembly of the *P. acidilactici* strain (FAM 18987) that was distinct from the other assemblies in the phylogenetic tree with an ANI to the reference sequence below 90 % was excluded from the pipeline run. This resulted in an increased number of identified core genes (from 921 to 1238), of species-specific sequences (from 54 to 516) and of reported primer pairs (from 2 to 53). The results of the two test cases illustrate that the SpeciesPrimer pipeline performs best on closely related genome assemblies with a good overall quality.

In silico **validation**

Two parameters were selected as criteria for the primer validation using web-based BLAST. First, the BLAST hits for the predicted PCR product sequence should only match the target species. If sequences of other bacterial species matched to parts of the sequence, the corresponding primer pairs were discarded, unless more than three mismatches were found in each primer-binding region for the forward and reverse primers. Second, the primer binding sites in the target sequences were not allowed to have mismatches in the 3'-end region. The criterion for the primer validation by Primer-BLAST was that no predicted PCR products for other bacterial species had been reported by Primer-BLAST. Primer pairs exclusively binding to the target sequence of the target species were classified as specific. The results of the in silico validation are summarized in Table 4. With the exception of Ec_faeca_g3060_1_P0 and Ec_faeci_cysS_3_P1, all primer pairs showed a perfect match to their target sequences. For primer pair Ec_faeca_g3060_1_P0, the first three nucleotides of one sequence out of 690, are missing in the forward primer-binding region. For Ec_faeci_cysS_3_P1, only one sequence out of 1058 aligned sequences showed a single nucleotide transition in the reverse primer-binding region (Data S3).

In vitro validation

The specificity of the qPCR assays was assessed with 21 to 25 strains of the target species (inclusivity) and 121 non-target bacterial strains found in dairy products (exclusivity). The qPCR assay performance was assessed by 10-fold dilution series of type strain DNA from 10^6 to 10^1 copies per reaction. The results of the qPCR runs were interpreted as positive if both qPCR reactions (duplicates) reached the fluorescent threshold before quantification cycle 35 and the peak of the melting curve analysis was above the peak calling threshold (-2 dF/dT). A summary of the results is shown in Table 5. The primer sequences can be found in Table S1. The inclusivity of the qPCR assays was 100 % for the assays Ec_faeca_acuI, Ec_faeca_g3060, Ec_faeci_cysS, Pd_acidi_asnS, Pd_acidi_g1164, Pd_pento_nagK and Pd_pento_g4364. Only one qPCR assay, Ec_faeca_purD was negative for one of the tested target strains.

Out of the 121 non-target strains analyzed to determine the exclusivity of the qPCR assays (Figure 2), all strains were negative for Ec_faeca_acuI and Pd_acidi_asnS. Both assays targeting *E. faecium* were positive solely for one *L. fermentum* strain (FAM 20347). Later it was found that the stock culture of this strain was contaminated with an *E. faecium* strain (data not shown). The assay Pd_pento_nagK targeting *P. pentosaceus* was positive for two out of three tested *Leuconostoc lactis* strains, the fluorescence signal reached the threshold after Cq 26, and the melting curve analysis showed a peak at 85 °C, while the positive control samples for this assay displayed a peak at 83.5 °C. Nine out of the 121 non-target strains were positive for the Ec_faeca_g3060 qPCR assay, for these samples the fluorescence signals reached the threshold after Cq 26 and had a melting curve peak at a higher temperature than the target PCR product. The assays Pd_acidi_g1164 and Pd_pento_g4364 were positive for five and eight non-target strains, respectively. Notably, all three tested *Lactobacillus paracasei* strains were positive for the Pd_acidi_g1164 assay, the fluorescence signal reached the threshold around Cq 21 and 22 and they showed a distinct melting curve peak at 86 °C.

The qPCR assays displayed linear results between 10^1 and 10^6 genome copies per reaction. The calculated efficiency of the qPCR assays was between 92 and 100 %. The linear regression equations ($Cq = slope * \log(copies) + intercept$) had slopes between -3.329 and -3.523 and correlation coefficients of 0.990 or above.

Discussion

After setup of the SpeciesPrimer docker container, the download of the local BLAST database

and the selection of the SpeciesPrimer run settings, no further manual handling was required to get primer pair candidates for all four bacterial species after a total time of 44 hours and 30 minutes. The number of input genomes and subsequently the number of retrieved primer pairs for the specificity check have the highest impact on speed. During the specificity check, blasting the primer sequences optimized for short sequences (blastn-short) and the subsequent compilation and indexing of the non-target sequence database are the most time consuming steps.

The results of the SpeciesPrimer pipeline for the four target species ranged from two to 160 identified primer pair candidates. Several factors can influence the number of identified primer pairs, such as the quality of the input genome assemblies, assemblies with wrong taxonomic labels and the genetic diversity within the species. A low-quality assembly with missing sequences or contaminations can decrease the number of identified core genes. The initial quality control helps to minimize the risk that such assemblies are included in the pipeline runs. However, also an increased sequence diversity, either due to sequencing errors, assembly errors or real diversity, limits the number and the length of identified conserved sequences. Subsequently this affects the yield of reported primer pairs, since the pipeline selects highly conserved sequences for primer design.

The specificity of the designed primers was evaluated in silico by BLAST with a more extensive database (RefSeq Genome) than the one used for the specificity check during primer design. The validation showed that the specificity of the tested amplicons was high and no other species than the target species had an identical sequence. Most target sequences in the database showed a perfect match for the primers in the primer-binding region. For all tested primer pairs, solely the expected PCR products for the target species and no amplicons for other sequences were predicted by Primer-BLAST. The results of Primer-BLAST indicate that the reported primer pairs were very specific, even though the species list used for the specificity evaluation during primer design covered only 259 non-target species.

In this work, 21 to 25 target strains for each target species and 121 non-target strains have been tested to assess inclusivity and exclusivity of the qPCR assays, respectively. The *in vitro* validation of primer pairs has shown that the *in silico* validation is not always able to predict non-target PCR products. The fluorescence signals occurring at late quantification cycles (Cq > 30) are probably due to PCR products with suboptimal primer binding. Testing the qPCR assays

in mixtures and communities could be interesting to assess if these PCR products also accumulate in presence of target DNA. The specificity could be sufficient in mixtures due to competition for the primers and the difference in primer binding and amplification efficiency. For many research applications, qPCR assays with a low signal in negative samples are acceptable, assuming that low-level signals can be distinguished from low concentrations of target species DNA by the melting curve analysis. Further, for many applications, the annealing temperature can be optimized by empirical determination of a suitable annealing temperature and the primer concentration can be adjusted to improve the specificity of the assay (www.bio-rad.com/en-ch/applications-technologies/qpcr-assay-design-optimization). We did not try to optimize our assays with these measures, because the aim was to design primers for high-throughput qPCR, requiring the exact same PCR conditions. For the tested qPCR conditions the most specific qPCR assays were Ec_faeca_acuI (*E. faecalis*), Ec_faeci_cysS (*E. faecium*), Pd_acidi_asnS (*P. acidilactici*) and Pd_pento_nagK (*P. pentosaceus*). Further work will be necessary in order to make these qPCR assays fully operational for the quantification of bacteria in a complex system such as food. For instance, suitable qPCR standards should be designed and validated, so that the limit of detection of each assay can be determined.

Primer-BLAST and RUCS allow designing primers for different applications, but demand experience and manual manipulations. Primer-BLAST designs primers and performs specificity checks, but requires a user provided target sequence. In the case of RUCS manual manipulation and experience is needed to prepare the positive and negative reference sets. Compared to primer-BLAST and RUCS, the task SpeciesPrimer performs is really specialized, namely to design primers for species-specific sequences. In contrast, SpeciesPrimer requires no previous knowledge about the input genome assemblies and no manual manipulation of sequences has to be performed. The ability of SpeciesPrimer to run on standard computers with good performance instead of specialized high-performance computers, will hopefully allow primer design for a wide range of scientists. Docker containers simplify the installation procedure and should allow non-bioinformaticians to setup and use the SpeciesPrimer pipeline.

Conclusions

In this work, we presented the SpeciesPrimer pipeline, which is a fully automated pipeline from the download of bacterial genomes, the identification of conserved species-specific core genes to primer design and subsequent quality control of primer candidates. Primers for four bacterial

species were designed and validated and have shown to perform adequately under the same qPCR conditions.

A standard computer with good performance, good quality genome assemblies, a local copy of the nt BLAST database and a list of non-target bacterial species are the only requirements for primer design with SpeciesPrimer. A complete image, of a Linux OS with all dependencies and the pipeline scripts, is available from Dockerhub. To simplify primer design for users not familiar with command line tools, a graphic user interface is provided in the latest version of SpeciesPrimer. SpeciesPrimer facilitates efficient primer design for species-specific quantification, paving the way for a fast and accurate quantitative investigation of microbial communities.

Acknowledgements

We would like to thank Marco Meola and Remo Schmidt for critically reading the manuscript and many helpful comments and Daniel Marzohl, Nadine Sidler, Elvira Wagner and Kotchanoot Srikham for their valuable technical help.

References

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

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

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

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 Molbak 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

Kleyer H, Tecon R, and Or D. 2017. Resolving Species Level Changes in a Representative Soil Bacterial Community Using Microfluidic Quantitative PCR. *Front Microbiol* 8:2017. 10.3389/fmicb.2017.02017

Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. In: Russell DJ, ed. *Multiple Sequence Alignment Methods*. Totowa, NJ: Humana Press, 155-170.

Masco L, Vanhoutte T, Temmerman R, Swings J, and Huys G. 2007. Evaluation of real-time PCR targeting the 16S rRNA and recA genes for the enumeration of bifidobacteria in probiotic products. *Int J Food Microbiol* 113:351-357. http://dx.doi.org/10.1016/j.ijfoodmicro.2006.07.021

Moyaert H, Pasmans F, Ducatelle R, Haesebrouck F, and Baele M. 2008. Evaluation of 16S rRNA Gene-Based PCR Assays for Genus-Level Identification of Helicobacter Species. *J Clin Microbiol* 46:1867-1869. 10.1128/jcm.00139-08

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, and Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31:3691-3693. 10.1093/bioinformatics/btv421

Postollec F, Falentin H, Pavan S, Combrisson J, and Sohier D. 2011. Recent advances in quantitative PCR (qPCR) applications in food microbiology. *Food Microbiol* 28:848-861. http://dx.doi.org/10.1016/j.fm.2011.02.008

Price MN, Dehal PS, and Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. 10.1371/journal.pone.0009490

Qu W, Zhou Y, Zhang Y, Lu Y, Wang X, Zhao D, Yang Y, and Zhang C. 2012. MFEprimer-2.0: a fast thermodynamics-based program for checking PCR primer specificity. *Nucleic Acids Res* 40:W205-208. 10.1093/nar/gks552

Ramirez M, Castro C, Palomares JC, Torres MJ, Aller AI, Ruiz M, Aznar J, and Martin-Mazuelos E. 2009. Molecular detection and identification of Aspergillus spp. from clinical samples using real-time PCR. *Mycoses* 52:129-134. 10.1111/j.1439-0507.2008.01548.x

Rice P, Longden I, and Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16:276-277. https://doi.org/10.1016/S0168-9525(00)02024-2

Sayers E. 2009. The E-utilities In-Depth: Parameters, Syntax and More. *Available at https://www.ncbi.nlm.nih.gov/books/NBK25499/*.

Scheirlinck I, Van der Meulen R, De Vuyst L, Vandamme P, and Huys G. 2009. Molecular source tracking of predominant lactic acid bacteria in traditional Belgian sourdoughs and their production environments. *J Appl Microbiol* 106:1081-1092. 10.1111/j.1365-2672.2008.04094.x

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068-2069. 10.1093/bioinformatics/btu153

Shen Z, Qu W, Wang W, Lu Y, Wu Y, Li Z, Hang X, Wang X, Zhao D, and Zhang C. 2010. MPprimer: a program for reliable multiplex PCR primer design. *BMC Bioinformatics* 11:143. 10.1186/1471-2105-11-143

Tange O. 2011. GNU Parallel - The Command-Line Power Tool. *login: The USENIX Magazine* 36:42-47.

Thomsen MCF, Hasman H, Westh H, Kaya H, and Lund O. 2017. RUCS: rapid identification of PCR primers for unique core sequences. *Bioinformatics* 33:3917-3921. 10.1093/bioinformatics/btx526

Torriani S, Felis GE, and Dellaglio F. 2001. Differentiation of Lactobacillus plantarum, L. pentosus, and L. paraplantarum by recA gene sequence analysis and multiplex PCR assay with recA gene-derived primers. *Appl Environ Microbiol* 67:3450-3454. 10.1128/aem.67.8.3450-3454.2001

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, and Rozen SG. 2012. Primer3--new capabilities and interfaces. *Nucleic Acids Res* 40:e115. 10.1093/nar/gks596

Vijaya Satya R, Kumar K, Zavaljevski N, and Reifman J. 2010. A high-throughput pipeline for the design of real-time PCR signatures. *BMC Bioinformatics* 11:340. 10.1186/1471-2105-11-340

Wang LT, Lee FL, Tai CJ, and Kasai H. 2007. Comparison of gyrB gene sequences, 16S rRNA gene sequences and DNA-DNA hybridization in the Bacillus subtilis group. *Int J Syst Evol Microbiol* 57:1846-1850. 10.1099/ijs.0.64685-0

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, and Madden TL. 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13:134. 10.1186/1471-2105-13-134

Yoon SH, Ha SM, Lim J, Kwon S, and Chun J. 2017. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Van Leeuwenhoek* 110:1281-1286. 10.1007/s10482-017-0844-4

Zuker M, Mathews DH, and Turner DH. 1999. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In: Barciszewski J, and Clark BFC, eds. *RNA Biochemistry and Biotechnology*. Dordrecht: Springer Netherlands, 11-43.

517

Table 1(on next page)

Overview of the SpeciesPrimer pipeline workflow and the used software.

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

1

Table 2(on next page)

Pipeline input and run statistics.

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)	**9:04**	**12:27**	**1:55**	**0:24**
Pipeline statistics				
Running time (h:min)	6:11	8:05	1:55	4:25
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

Table 3(on next page)

Summarized results of the average nucleotide identity (ANI) calculations.

	E. faecium main cluster	*E. faecium* subcluster	*P. acidilactici* main cluster	*P. acidilactici* subcluster
Assemblies	530	44	125	1
ANI (%)				89.21
average	99.43	94.67	98.28	
maximum	99.86	95.60	98.83	
minimum	98.19	94.15	96.88	

1

Table 4(on next page)

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27870v1 | CC BY 4.0 Open Access | rec: 23 Jul 2019, publ: 23 Jul 2019

Target species	Primer pair	PPC	BLAST (perfect/total)	primer-BLAST (perfect/total)
E. faecalis	Ec_faeca_acuI_1_P0	100	specific (694/694)	specific (24/24)
	Ec_faeca_g3060_1_P0	100	specific (689/690)	specific (24/24)
E. faecium	Ec_faeci_cysS_3_P1	96.7	specific (1057/1058)	specific (63/63)
	Ec_faeci_purD_2_P0	93.3	specific(1083/1083)	specific (63/63)
P. acidilactici	Pd_acidi_asnS_2_P0	90.1	specific (19/19)	specific (5/5)
	Pd_acidi_g1164_1_P0	93.3	specific (23/23)	specific (5/5)
P. pentosaceus	Pd_pento_nagK_1_P0	100	specific (15/15)	specific (7/7)
	Pd_pento_g4364_1_P0	100	specific (15/15)	specific (7/7)

1

Table 5(on next page)

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	0/121	9/121	0*/121	0*/121	0/121	5/121	2/121	8/121
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
R^2	0.998	0.997	0.990	0.996	0.997	0.995	0.996	0.997

1 * Contamination of stock culture of the strain FAM20347 with *E. faecium*.

Figure 1

Phylogeny of core gene alignments.

(A) *E. faecium.* (B) *P. acidilactici*. The phylogenies are displayed with SeaView (http://doua.prabi.fr/software/seaview) in circular view.

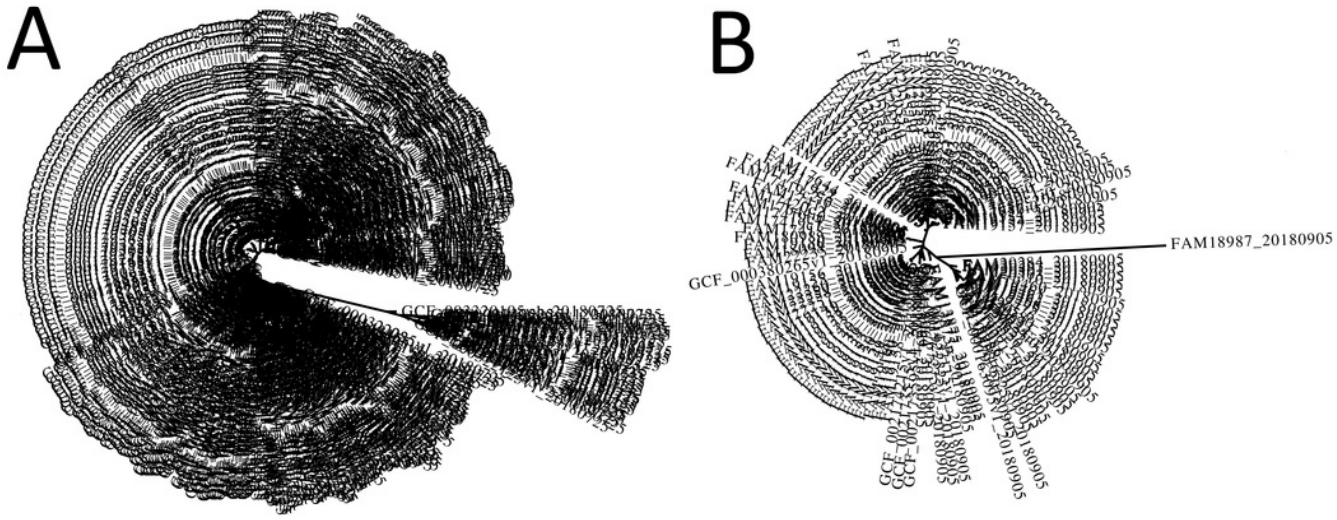

Figure 2

qPCR assay quantification cycle heatmap.

Depicted are all tested non-target strains and their average quantification cycle (technical duplicates). Bars represent results with a melt curve peak above the threshold and a Cq value below Cq 35. 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.