



Chhandak Basu *Editor*

PCR Primer Design

Second Edition

METHODS IN MOLECULAR BIOLOGY

Series Editor

John M. Walker

School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
<http://www.springer.com/series/7651>

PCR Primer Design

Second Edition

Edited by

Chhandak Basu

Department of Biology, California State University, Northridge, Los Angeles, CA, USA



Editor

Chhandak Basu

Department of Biology

California State University

Northridge

Los Angeles, CA, USA

ISSN 1064-3745

Methods in Molecular Biology

ISBN 978-1-4939-2364-9

DOI 10.1007/978-1-4939-2365-6

ISSN 1940-6029 (electronic)

ISBN 978-1-4939-2365-6 (eBook)

Library of Congress Control Number: 2015931213

Springer New York Heidelberg Dordrecht London
© Springer Science+Business Media New York 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Humana Press is a brand of Springer

Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

In memory of my brother Professor Sangata Basu

Preface

The field of molecular biology and biotechnology has revolutionized since the Nobel Prize winning (1993) work of Dr. Kary Banks Mullis. Dr. Mullis first discovered how to synthesize large amount of DNA from infinitesimal small amount of DNA, and the process is known as Polymerase Chain Reaction (PCR). It is now possible for us to make millions of copies of DNA from minuscule amount of starting DNA within a period of few hours. PCR has been widely used in agriculture, medicine, forensics, paternity testing, molecular ecology, biotechnology etc. and the list will continue to grow. The use of PCR in modern day science is impressive, and it is evident with more than half a million hits in the PubMed Central database with a simple keyword search: "PCR." The tiny PCR tube contains a variety of ingredients including magnesium chloride, dNTPs, the interesting thermostable enzyme *Taq* polymerase (originally isolated from thermophilic bacterium *Thermus aquaticus*), and last but not least the oligonucleotides. The oligonucleotides, also known as primers, play a very important role in successful amplification of a segment of DNA. In other words, a poor primer design may result in less than desirable PCR product. This book specifically focuses on how to design PCR primers for successful DNA amplification. There are 15 chapters in this book, and the chapters cover wide ranges of topics in PCR primer design including primer design strategies for quantitative PCR, for use in forensic biology, for genotyping, for degenerate PCR, for multiplex PCR etc. Besides these, there are also chapters that focus on in silico PCR primer design and primer design using software available over the Internet. This book was a true international effort and scientists from USA, India, Estonia, Spain, Japan, China, Czech Republic, and Brazil contributed to this book. We hope this book will be useful to molecular biology students, researchers, professors, and PCR enthusiasts.

Los Angeles, CA, USA

Chhandak Basu

Contents

<i>Preface</i>	<i>vii</i>
<i>Contributors</i>	<i>xi</i>
1 Fast Masking of Repeated Primer Binding Sites in Eukaryotic Genomes	1
<i>Reidar Andresson, Lauris Kaplinski, and Maito Remm</i>	
2 Primer Design for PCR Reactions in Forensic Biology	17
<i>Kelly M. Elkins</i>	
3 Design of Primers and Probes for Quantitative Real-Time PCR Methods	31
<i>Alicia Rodríguez, Mar Rodríguez, Juan J. Córdoba, and María J. Andrade</i>	
4 Large-Scale Nucleotide Sequence Alignment and Sequence Variability Assessment to Identify the Evolutionarily Highly Conserved Regions for Universal Screening PCR Assay Design: An Example of Influenza A Virus	57
<i>Alexander Nagy, Tomáš Jiřínek, Lenka Černíková, Helena Jiřincová, and Martina Havlíčková</i>	
5 Low-Concentration Initiator Primers Improve the Amplification of Gene Targets with High Sequence Variability	73
<i>Kenneth E. Pierce and Lawrence J. Wangh</i>	
6 Multiplex PCR Primer Design for Simultaneous Detection of Multiple Pathogens	91
<i>Wenchao Yan</i>	
7 Degenerate Primer Design for Highly Variable Genomes	103
<i>Kelvin Li, Susmita Srivastava, and Timothy B. Stockwell</i>	
8 Allele-Specific Real-Time Polymerase Chain Reaction as a Tool for Urate Transporter 1 Mutation Detection	117
<i>Juliet O. Makanga, Antonius Christianto, and Tetsuya Inazu</i>	
9 MultiPLX: Automatic Grouping and Evaluation of PCR Primers	127
<i>Lauris Kaplinski and Maito Remm</i>	
10 In Silico PCR Primer Designing and Validation	143
<i>Anil Kumar and Nikita Chordia</i>	
11 Primer Design Using Primer Express® for SYBR Green-Based Quantitative PCR	153
<i>Amarjeet Singh and Girdhar K. Pandey</i>	
12 Designing Primers for SNaPshot Technique	165
<i>Greiciane Gaburro Paneto and Francisco de Paula Careta</i>	

13	Rapid and Simple Method of qPCR Primer Design	173
	<i>Brenda Thornton and Chhandak Basu</i>	
14	PRIMEGENSw3: A Web-Based Tool for High-Throughput Primer and Probe Design	181
	<i>Garima Kushwaha, Gyan Prakash Srivastava, and Dong Xu</i>	
15	Selecting Specific PCR Primers with MFEprimer	201
	<i>Wubin Qu and Chenggang Zhang</i>	
	<i>Index</i>	215

Contributors

- MARÍA J. ANDRADE • *Food Hygiene and Safety, Meat and Meat Products Research Institute, Faculty of Veterinary Science, University of Extremadura, Cáceres, Spain*
- REIDAR ANDRESON • *Department of Bioinformatics, University of Tartu, Tartu, Estonia; Estonian Biocentre, Tartu, Estonia*
- CHHANDAK BASU • *Department of Biology, California State University, Northridge, Los Angeles, CA, USA*
- FRANCISCO DE PAULA CARETA • *Department of Pharmacy and Nutrition, CCA, Federal University of Espírito Santo, Alegre, ES, Brazil*
- LENKA ČERNÍKOVÁ • *Laboratory of Molecular Methods, State Veterinary Institute Prague, Prague, Czech Republic*
- NIKITA CHORDIA • *School of Biotechnology, Devi Ahilya University, Indore, India*
- ANTONIUS CHRISTIANTO • *Laboratory of Functional Genomics, Department of Pharmacy, College of Pharmaceutical Sciences, Ritsumeikan University, Shiga, Japan*
- JUAN J. CÓRDOBA • *Food Hygiene and Safety, Meat and Meat Products Research Institute, Faculty of Veterinary Science, University of Extremadura, Cáceres, Spain*
- KELLY M. ELKINS • *Chemistry Department, Towson University, Towson, MD, USA*
- MARTINA HAVLÍČKOVÁ • *National Institute of Public Health, Centre for Epidemiology and Microbiology, National Reference Laboratory for Influenza, Prague, Czech Republic*
- TETSUYA INAZU • *Laboratory of Functional Genomics, Department of Pharmacy, College of Pharmaceutical Sciences, Ritsumeikan University, Shiga, Japan*
- HELENA JÍŘINCOVÁ • *National Institute of Public Health, Centre for Epidemiology and Microbiology, National Reference Laboratory for Influenza, Prague, Czech Republic*
- TOMÁŠ JÍŘINEC • *National Institute of Public Health, Centre for Epidemiology and Microbiology, National Reference Laboratory for Influenza, Prague, Czech Republic*
- LAURIS KAPLINSKI • *Department of Bioinformatics, University of Tartu, Tartu, Estonia; Estonian Biocentre, Tartu, Estonia*
- ANIL KUMAR • *School of Biotechnology, Devi Ahilya University, Indore, India*
- GARIMA KUSHWAHA • *Informatics Institute and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA*
- KELVIN LI • *The J. Craig Venter Institute, Rockville, MD, USA*
- JULIET O. MAKANGA • *Laboratory of Functional Genomics, Department of Pharmacy, College of Pharmaceutical Sciences, Ritsumeikan University, Shiga, Japan*
- ALEXANDER NAGY • *Laboratory of Molecular Methods, State Veterinary Institute Prague, Prague, Czech Republic; National Institute of Public Health, Centre for Epidemiology and Microbiology, National Reference Laboratory for Influenza, Prague, Czech Republic*
- GIRDHAR K. PANDEY • *Department of Plant Molecular Biology, University of Delhi South Campus, New Delhi, India*

GREICIANE GABURRO PANETO • *Department of Pharmacy, Federal University of Espírito Santo, Vitória, ES, Brazil; Department of Pharmacy, Nutrition Alto Universitario s/n CCA, Alegre, ES, Brazil*

KENNETH E. PIERCE • *Department of Biology, Brandeis University, Waltham, MA, USA*

WUBIN QU • *Beijing Institute of Radiation Medicine, State Key Laboratory of Proteomics,*

Cognitive and Mental Health Research Center of PLA, Beijing, China

MAIDO REMM • *Department of Bioinformatics, University of Tartu, Tartu, Estonia; Estonian Biocentre, Tartu, Estonia*

ALICIA RODRÍGUEZ • *Food Hygiene and Safety, Meat and Meat Products Research Institute, Faculty of Veterinary Science, University of Extremadura, Cáceres, Spain*

MAR RODRÍGUEZ • *Food Hygiene and Safety, Meat and Meat Products Research Institute, Faculty of Veterinary Science, University of Extremadura, Cáceres, Spain*

SUSMITA SHRIVASTAVA • *The J. Craig Venter Institute, Rockville, MD, USA*

AMARJEET SINGH • *Department of Plant Molecular Biology, University of Delhi South Campus, New Delhi, India*

GYAN PRAKASH SRIVASTAVA • *Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*

TIMOTHY B. STOCKWELL • *The J. Craig Venter Institute, Rockville, MD, USA*

BRENDA THORNTON • *Treasure Coast High School, Delray Beach, FL, USA*

LAWRENCE J. WANGH • *Department of Biology, Brandeis University, Waltham, MA, USA*

DONG XU • *Computer Science Department, Informatics Institute, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA*

WENCHAO YAN • *Animal Quarantine Laboratory, College of Animal Science and Technology, Henan University of Science and Technology, Luoyang, Henan Province, China*

CHENGGANG ZHANG • *Beijing Institute of Radiation Medicine, State Key Laboratory of Proteomics, Cognitive and Mental Health Research Center of PLA, Beijing, China*

Chapter 1

Fast Masking of Repeated Primer Binding Sites in Eukaryotic Genomes

Reidar Andreson, Lauris Kaplinski, and Maito Remm

Abstract

In this article we describe the working principle and a list of practical applications for GenomeMasker—a program that finds and masks all repeated DNA motifs in fully sequenced genomes. The GenomeMasker exhaustively finds and masks all repeated DNA motifs in studied genomes. The software is optimized for PCR primer design. The algorithm is designed for high-throughput work, allowing masking of large DNA regions, even entire eukaryotic genomes. Additionally, the software is able to predict all alternative PCR products from studied genomes for thousands of candidate PCR primer pairs.

Practical applications of the GenomeMasker are shown for command-line version of the GenomeMasker, which can be downloaded from <http://bioinfo.ut.ee/download/>. Graphical Web interfaces with limited options are available at <http://bioinfo.ut.ee/genometester/> and <http://bioinfo.ut.ee/snpmasker/>.

Key words PCR, DNA repeats, Primer design, Microarrays, DNA masking

1 Introduction

Modern genomic technologies allow studying thousands of genomic regions from each DNA sample. Most of these technologies require PCR amplification to achieve sufficiently strong signals. Although many currently available high-throughput technologies use one single pair of universal PCR primers, most applications still need a large number of custom-designed PCR primers. Therefore, there is an urgent need for automatic PCR primer design methods. The automatic PCR primer design is trusted only if it generates primers with a success rate equal to or higher than the manual primer design. Although not all aspects of the PCR process are thoroughly described and modeled, there are certain factors that are known to affect the PCR success rate. One of such factors is primer overlap with repeated regions on the template DNA. The primers that bind to the repeats are more likely to fail or generate false products.

We started studying that problem in early 2001 when we got involved in a large-scale genotyping experiment covering the whole human chromosome 22 [1]. The study analyzed 1,278 single nucleotide polymorphisms (SNPs). For each SNP a separate PCR primer pair was designed. We planned to find the correlations between the PCR primer pair success rate and the sequence properties of the primers. One of the factors studied was the number of predicted primer binding sites. However, it quickly turned out that finding the number of binding sites for thousands of primer pairs from the entire human genome is not an easy task. One of the difficulties was that we did not know exactly what to search for. The binding site of the primer can be defined in various different ways. For example, the binding site can be defined as 8 nucleotides from the primer 3'-end, or 12 nucleotides from the primer 3'-end, or 16 nucleotides from the primer 3'-end or any other number of nucleotides from the primer 3'-end. Also, it defined as the exact match, 100 % identical to the primer or similarity with one or two mismatches. Furthermore, the binding site can be defined as variable length string from the 3'-end of the primer. In this case the length of the primer varies so that the binding energy exceeds certain ΔG level for many different ΔG values. If we want to study all these potential models of binding sites for their effect on prediction of PCR success rate, we first have to calculate the number of binding sites for all the primer pairs with each binding site model. In our case, this meant finding the binding sites for 1,278 primer pairs in hundreds of different ways from the human genome.

We tried to use existing programs for finding and counting primer binding sites from the human genome. The BLAST program is most frequently used for this purpose in multiple applications [2–4]. Unfortunately, the speed of BLAST is not sufficient for counting primer binding sites in large eukaryotic genomes with large number of primers. The speed can be increased by using MEGABLAST [5], SSAHA [6], or BLAT [7], which are specifically designed for homology search from large genomes. Unfortunately, these faster programs are not optimized for primer design tasks and thus recording the number and the location of the predicted primer binding sites requires additional efforts.

Thus, we decided to create software that counts all primer binding sites in the human genome within seconds and reports potential PCR products for thousands of primer pairs. The program, named GenomeTester, is based on exhaustive counting and recording of the locations of all potential binding sites from the human genome. The locations are stored in a binary hash data structure, which allows extremely fast retrieval of the number and the location of all binding sites for any given primer.

To simplify the design of PCR primers in future, we have also added a possibility to mask repeated regions on a template DNA. This helps to avoid the design of PCR primers with extensive

number of binding sites and thus increase the success rate of the designed primers. Masking of repeats on the template DNA is a common approach that is used to mark specific regions in DNA. DUST (<ftp://ftp.ncbi.nlm.nih.gov/pub/tatusov/dust/>) and TandemRepeatsFinder [8] are commonly used for masking simple (short) repeat motifs. RepeatMasker is a universal program that is used for masking out several kinds of repeats and is therefore mostly used for this kind of sequence analysis (Smit AFA, Hubley R, and Green P; <http://www.repeatmasker.org/>). Similarly, BLAST [9] can be used to mask the nonunique regions of the genome [10, 11]. Our masking software, named GenomeMasker, is dedicated to masking of repeated primer binding sites in large genomes. The details of the algorithm are briefly described in Chapter 2. Examples of practical use of our programs are shown in Chapter 3.

2 Working Principle and Data Structures

The efficiency of both GenomeTester and GenomeMasker software is based on preprocessing of genomic sequence into specific data structure—the hash structure.

2.1 GenomeMasker

In GenomeMasker, the hash structure contains list of all repeated sequence motifs from a given genome, encoded into 32 bit integers. The encoding is done by allocating two consecutive bits for each nucleotide in a word. Thus the length of repeated sequence motif X in current implementation is in the range between 8 and 16 nucleotides. Our group is using 16 nt. as default value of X , because it seems to give the best separation between high and low success-rate PCR primers. Most of the following examples in this article are also based on repeat length $X=16$. The sequence motif is defined as “repeated” if given nucleotide sequence occurs in the given genome more than Y times, where Y is an integer chosen by user (e.g., 1, 2, 3, etc.). For example, if motif length $X=16$ nucleotides and tolerated repeat number $Y=1$, then all 16 nt. long sequence motifs that occur more than once in the given genome are put into list of repeats and stored in the hash structure. The entire hash structure of repeated motifs is sorted for faster access and written into blacklist file.

The program GenomeMasker uses this blacklist file as a reference to quickly mask the template sequence for PCR primer design. The GenomeMasker iterates over the whole template sequence (which can be the entire genome) with window length X nucleotides and with step 1 nucleotide. For each window, it checks whether the sequence motif within the window or its reverse complement is recorded in the blacklist file. If the given sequence motif is in the blacklist, the corresponding window in the template sequence is masked (Fig. 1a). Theoretically, primers with partial overlap with

a 16-nucleotide masking (allowed 3'-end overlap with repeat = 0 nt)



b 10-nucleotide masking (allowed 3'-end overlap with repeat = 6 nt)



c 1-nucleotide masking (allowed 3'-end overlap with repeat = 15 nt)



Fig. 1 Principles of masking and subsequent primer design with GenomeMasker software package. The repeated 16-nucleotide motif is *striped*; nonrepeated motifs are shown in *black*. The rejected primers are *striped*, accepted primers in *black*. The number of masked nucleotides can be selected by the user. Full-length masking of repeated motifs is most conservative, allowing design of only two out of six primers on this example (**a**), whereas single-nucleotide masking allows design of four out of six primers (**c**). Please note that the method generates asymmetric masking only if less than 16 nucleotides of repeated 16-nucleotide motif are masked

repeats can work in PCR. Therefore, we can design PCR primers, which partly overlap with repeated motif as long as the last 16 nucleotides from the 3'-end of the primer are not repeated (Fig. 1a). This will lead to asymmetric masking because primers are single-stranded and the 3' end of repeat is different on upper and lower strand. For example, the primer A from upper strand can work fine whereas its reverse complement from lower strand (primer B) may fail because its 3'-end is overlapping with repeat (Fig. 1a). Furthermore, small overlap with the repeated region can be tolerated even if it happens in primers 3'-end. Therefore, we do not have to mask the entire repeat, but just a couple of nucleotides from the repeats 3'-end (Fig. 1b). We have initiated a comprehensive experimental analysis of factors influencing PCR primer success. Our preliminary results indicate that overlaps with repeated regions can be tolerated as long as the last 16 nucleotides of a primer are not repeated. Thus, masking of single nucleotide at the 3'-end of repeat may be sufficient to avoid PCR primers with low success rate (Fig. 1c). Nevertheless, the length of masked region Z can be changed with a special option in GenomeMasker software, allowing each user to select his/her own settings for Z.

2.2 GM_PRIMER3

With GenomeMasker, the sequence can be masked by any user-defined character; however the most useful masking style is with lowercase letters. The lowercase masking maintains the sequence information in masked regions and allows subsequent primer design from the masked sequence even if some primers overlap with masked nucleotides. To take advantage of lowercase-masked sequence, we have modified the well-known program PRIMER3 [12]. The overall functionality and algorithm of the program is the same as in the original PRIMER3, but we have added a new filtering feature that rejects the primer candidates with lowercase letters at their 3' end and new parameters for calculation of melting temperature of primers. Although standard version of the PRIMER3 uses nearest neighbor model to calculate melting temperature, the parameters used for that are rather old [13]. We replaced the nearest neighbor parameters with a newer set [14], added sequence-dependent salt correction formula [15] and correction for concentration of divalent cations [16].

All the changes are now incorporated into PRIMER3 main source code. The modified version of PRIMER3 is available at <http://sourceforge.net/projects/primer3/>, an online version of the program can be used at <http://primer3.ut.ee/>.

2.3 GenomeTester

GenomeTester is a program for counting potential binding sites and potential products for each tested PCR primer. During the preprocessing of the genome sequence, the GenomeTester exhaustively counts all locations of all possible PCR primers. The data structure for preprocessed genomic data is hash structure,

similar to GenomeMasker. The main difference from abovementioned GenomeMasker program is that the GenomeMasker stores only the number of binding sites for each potential primer, whereas GenomeTester also stores all locations (genome coordinates) for each potential primer. The hash structure created during preprocessing is sorted and saved into file(s). The hash structure allows fast identification of PCR primer binding sites and PCR products for large number of primers.

Similar hash structure is used by the program SSAHA, which is designed to run fast sequence searches from genomic sequences. However, the GenomeTester uses slightly different data structure, which allocates equal amount of memory for each location and thus makes it faster in subsequent searches.

3 Practical Applications

3.1 GenomeMasker Application

GenomeMasker is suitable for users who need to design PCR primers that are unique in given genomic DNA. It can be described as an additional filtering stage to avoid primer candidates with low success rate. All this can be done before the actual primer design starts and the output of the GenomeMasker can be used for selecting the successful primer candidates.

GenomeMasker application contains the following executables:

- *glistmaker*—creates a list of repeated (occurring more frequently than user-defined threshold) words in a given genome. This step has to be performed only once for a given set of genomic data and chosen word length.
- *gmasker*—performs a masking procedure for each studied FASTA sequence.

3.1.1 The Usage of *glistmaker*

The GenomeMasker requires preprocessed data structure (blacklist file) as a reference to mask DNA sequences. The executable *glistmaker* creates binary blacklist file for GenomeMasker application from the genomic data in the FASTA format. Here are the options for *glistmaker*:

```
prompt> ./glistmaker -h
Usage: glistmaker OPTIONS inputfilelist outputfile wordsize
-v                                     - Print version and exit
-h                                     - Print this usage screen and exit
-d                                     - Turn on debugging output
-overreplimit NUMBER - Specify overrepresentation cutoff
(default 10)
```

Please see **Note 1** for detailed explanation of input parameters and files. The following is a command-line example of running *glistmaker*:

```
prompt> ./glistmaker -overreplimit 100 human_chr_
list.txt human_repeats.list 12
```

Command shown above will create a blacklist file (*human_repeats.list*) that contains all 12-mer words having more than 100 different locations in all chromosomes included to “*human_chr_list.txt*” file.

3.1.2 The Usage of *gmasker*

After creating the blacklist file, the user can start masking sequence files containing template DNA regions (in FASTA format) with the second executable in GenomeMasker application called *gmasker*. Options for the *gmasker* are shown below:

```
prompt> ./gmasker -h
Usage: gmasker OPTIONS blacklistfile maskingletter maskingtype
[ start end]
-v           - Print version and exit
-h           - Print this usage screen and exit
-d           - Turn on debugging output
-u           - Convert sequence to uppercase before
processing
-nbases NUMBER - How many bases from 3' end to mask
(default 1)
```

Please see **Note 2** for detailed explanation of input parameters and files.

User can define the sequence file that should be masked by *gmasker* and manage output with different ways using Unix pipes (“|”, “<” and “>”). Here are some examples:

```
prompt> cat sequences.fas | ./gmasker human_repeats.list N both >
masked_sequences.fas
prompt> ./gmasker human_repeats.list N both < sequences.fas >
masked_sequences.fas
```

In both cases shown above, “sequences.fas” will be used as an input file (see example sequence in Fig. 2) and the results are written into “masked_sequences.fas”.

```
>user sequence
TTAACGTTTCAAAAAGTTAACAGTACCTATGTCTTCATAGTATTATTACATTCAAGAAATTAAATTTCAGCTACCTTCTTATCCATAGTTCTAA
CTTATGAAATAAGCATGCATTGTTCTGTAAGAAAAATAATTTTATTCTGTTATTAAATAAAATGATATATACATCAGTTGAGTAATAAA
CTTTTCATCTAGATAAGTTTTAAAMTGTGTAATACATTAATAATTATTGTGAGTAGGAATTACCTACATTATATTTAAATAATAGGT
TTTAATAAAATAACATTACAGAGTGCAGTCATTGTTGATTAAGAGTTCTAAACTTTATGTATGTTAGGCTGACCTGAAAGATCCAGAACATCG
ATCTGATTTCCCTTGATCTGTC
```

Fig. 2 Example input file for *gmasker*

a

```
>user sequence (masked_sequences.fas)
TTAACGTTTCAAAAAGTTacagtACCTATGTCTTCATAGTAtttAttaCattcAGaaaTTttaatttacCtaCCCTCTTtATCCATAGTTCTAATCttATGAAat
AagCATGCattGTGtTGAGTAaaaaaatattttattctgttatttaaaaaatgtatataactcataGTTGAGTAaaaaACTtTtcacatcAGATAAGTT
TTTAAAMGTGTAAtacattaaTTAAATTATTGAGTAGGAAttaCctacatttatattttaataaaataggtttttaataaaAtAACATTACAGAGTTGCA
GTCATTGTTGATTAAGatCTcaAataCTTTATGTAGTGACCTGAAAGATCCAGAACATCGATCTGATTTCCCTGATCTGTC
```

b

```
>user sequence
TTAACNNNNCAAAAAGTTANNNNNNACNNNGNNNTNNNGTNNNNNANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNCATGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
TTTANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
GTCATTGTTGATTAAGNNNTCENNANCCTTNNGNTNNNNNGCTGACCTGAAAGATCCAGAACATCGATCTGATTTCCCTGATCTGTC
```

c

```
>user sequence
TTAAC#####AACAGTAC ##### ##### ##### ##### ##### ##### ####TTCTTAT##### ##### ##### ##### ##### ##### #####
##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### TCATC#####
##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### GAGTTGCA
##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### ##### #####
```

Fig. 3 Different masking possibilities with GenomeMasker application. The first section (**a**) illustrates a result of the *gmasker* using a “target” masking type. For the PCR primer design, only the upper strand should be masked on the *left side* of the target region and only the lower strand should be masked on the *right side* of the target region. The *middle part* (bases from 220 to 230) is the target region, which is chosen to be amplified. The character “I” defines that 3’ ends of the repeated words are masked with *lower case letters*. Next two sections demonstrate alternative masking possibilities: (**b**) upper and lower strand (“both”) masking and (**c**) lower strand (“backward”) masking. With special option “-nbases” the user can define how many nucleotides will be masked by *gmasker*. The masking was done by using NCBI build 35.1 human genome assembly. GenomeMasker blacklist was created by using wordsize 16 and the maximum number of allowed locations for each word was set to 10

The final output of GenomeMasker application is a masked FASTA file, where 3’ end of repeated words is replaced with a user-defined character. Figure 3 shows additional masking possibilities and command-line examples to run *gmasker*.

3.2 GM_PRIMER3 Application

To design primers with *gm_primer3*, user needs several supporting programs. Current application includes two Perl scripts, which help to automate the primer design with the *gm_primer3*:

- *fasta_to_p3.pl*—converts FASTA format sequences to *gm_primer3* input format.
- *p3_to_table.pl*—converts *gm_primer3* output to tab-delimited table format with the following columns: name, sense_primer, antisense_primer, product sequence. This table format can be used as an input for the *gtester* executable (see Section 3.3.2) and for MultiPLX program [17].

3.2.1 The Usage of *gm_primer3*

The following parameters need to be changed in *fasta_to_p3.pl* script before primer design can begin:

```
PRIMER_PRODUCT_SIZE_RANGE=100-600
PRIMER_PRODUCT_OPT_SIZE=200
PRIMER_OPT_SIZE=21
PRIMER_MIN_SIZE=18
PRIMER_MAX_SIZE=26
PRIMER_OPT_TM=62
PRIMER_MIN_TM=59
PRIMER_MAX_TM=65
PRIMER_MAX_DIFF_TM=4
PRIMER_OPT_GC_PERCENT=35
PRIMER_MIN_GC=20
PRIMER_MAX_GC=70
PRIMER_SALT_CONC=20
PRIMER_FILE_FLAG=0
PRIMER_EXPLAIN_FLAG=1
PRIMER_MAX_POLY_X=4
PRIMER_NUM_RETURN=1

TARGET=475,51
```

Of course, the user can add additional parameters here; the complete list of possible options and their explanation is available at GM_PRIMER3 Web site (<http://primer3.ut.ee/>). Otherwise the default values will be used for other parameters.

Here is a command-line example to run *gm_primer3* for primer design:

```
prompt> cat masked_sequence.fas | ./fasta_to_p3.pl | ./gm_primer3
| ./p3_to_table.pl > primers.txt
```

The result of successful primer design is a tab-delimited “primers.txt” file that contains primer id, primer, and product sequences. If *gm_primer3* is unable to find good primer candidates, these columns are filled with dashes. What to do with the regions, where *gm_primer3* could not design primers? The first thing is to try relaxing various *gm_primer3* parameters, because too strict parameters may force the *gm_primer3* to reject many good primer candidates (see **Note 3** for more detailed parameter explanations). Secondly, make sure that there are adequate stretches of non-Ns in the regions in which you wish to pick primers. Finally, in some cases the region of interest is full of repeats and therefore *gm_primer3* cannot design primers for a given sequence.

3.3 GenomeTester Application

Having too many binding sites will typically result in failed PCR, but that can be eliminated with GenomeMasker application. A different problem emerges if two primers give several alternative products in PCR reaction. Amplifying more than one product is undesirable because alternative PCR products could cause false positive signals in genotyping. The GenomeTester programs can be used to make sure if designed primers produce single PCR product or not.

GenomeTester application contains the following executables:

- *gindexer*—creates index files containing locations of all the predicted binding sites in a given genome.
- *gtester*—counts all locations of all possible PCR primers and predicts PCR products.

3.3.1 The Usage of *gindexer*

The program *gindexer* is needed to create index files for *gtester* to work. The executable *gindexer* creates binary index files for GenomeTester application from the genomic data (in FASTA format). Here are the options for *gindexer*.

```
prompt> ./gindexer -h

Usage: gindexer OPTIONS inputfile outputfile
-v           - Print version and exit
-h           - Print this usage screen and exit
-d           - Turn on debugging output
-wordsize LENGTH - Specify word size for index (default 16)
```

Please see **Note 4** for detailed explanation of input parameters and files.

Different possibilities to execute *gindexer* are shown below:

```
prompt> ./gindexer ecoli_genome.fas ecoli
prompt> ./gindexer -wordsize 12 chr22.fas chr22
prompt> ./gindexer /home/db/human_DNA/chr/1.fa indexes/1
prompt> ./gindexer /home/db/human_DNA/chr/2.fa indexes/2
...
prompt> ./gindexer /home/db/human_DNA/chr/Y.fa indexes/Y
```

The first example creates index files (*ecolia.location*, *ecolit.location*, *ecolic.location*, and *ecolig.location*) for *Escherichia coli* complete genome. The second example shows the possibility to use smaller word length (-wordsize 12) to create indexes for given chromosome sequence. The last section illustrates the series of executions to create indexes for all human chromosomes.

3.3.2 The Usage of *gtester*

After creating index files for chromosomes or genomic sequences, the user can start using *gtester* program. Here are the command-line options for *gtester*:

```
prompt> ./gtester -h

Usage: gtester OPTIONS primerfile locationsfile
-v           - Print version and exit
-h           - Print this usage screen and exit
-d           - Turn on debugging output
-maxprodlen LENGTH - Specify maximum product length (default
1000 bp)
-limit NUMBER      - Maximum number of binding sites to track
(default 1000)
```

Please see **Note 5** for detailed explanation of input parameters and files.

The basic execution of *gtester* is simple:

```
prompt> ./gtester primers.txt human_indexes_list.txt
```

The file “primers.txt” is the output of GM_PRIMER3 application and “human_indexes_list.txt” is the list of file names for all indexes of human chromosomes. The command above will generate three result files ending with specific suffixes (.gt1, .gt2, .gt3): primers.txt.gt1, primers.txt.gt2, and primers.txt.gt3. Figure 4 illustrates the columns of these three files.

There are also alternative possibilities to use *gtester*:

```
prompt> ./gtester -maxprodlen 10000 -limit 5000 primers.txt
human_indexes_list.txt

prompt> ./gtester primers.txt human_indexes_list.txt
human_repeats.list
```

The first example illustrates the possibility to define the maximum length of products and the number of binding sites. The second example shows how to use a preprocessed blacklist file created by *glistmaker* to reduce the calculation time *gtester*.

```

a
primers.txt.gt1

NAME
NUMBER OF BINDING SITES FOR PRIMER A (left primer)
NUMBER OF BINDING SITES FOR PRIMER B (right primer)
NUMBER OF PRODUCTS

b
primers.txt.gt2

NAME
PRODUCT NUMBER
CHR
LOCATION (start nucleotide)
LENGTH (bp)
TYPE OF PRODUCT
  1: PrimerA-PrimerB (sense strand product)
  -1: PrimerB-PrimerA (antisense strand product)
  2: PrimerA-PrimerA
  -2: PrimerB-PrimerB

c
primers.txt.gt3

NAME OF THE PRIMER PAIR
PRIMER (A or B)
STRAND (1=sense, -1=antisense)
CHR
LOCATION OF THE 5' END OF THE PRIMER

```

Fig. 4 Description of three output files of *gtester*. The *gtester* executable produces three result files: **(a)** a file with the number of primer binding sites and with the number of products, **(b)** a file with the description of all PCR products, and **(c)** a file with the description of all primer binding sites

4 Notes

The first required input file—“inputfilelist”—is a text file containing locations of the chromosome or contig files (in FASTA format) in the user system, one file name per line:

```

/home/db/human_DNA/chr/1.fa
/home/db/human_DNA/chr/2.fa
/home/db/human_DNA/chr/3.fa
/home/db/human_DNA/chr/4.fa
...
/home/db/human_DNA/chr/Y.fa

```

The second file—“outputfile”—is the name of the blacklist file (in binary format) that the *glistmaker* creates.

The third parameter is called “-wordsize” and must be defined by the user. It represents the length of the word (sequence window) that program will use to find and store repeats to the blacklist file. The maximum length of the word can be 16 and minimum 8 nucleotides.

Optional parameter—“-overreplimit”—should be changed if the user wishes to use smaller word size than 16 nucleotides (by default it is 10). Otherwise sequences might be masked with 100 %, because short words are more common in genomes. That cut-off defines the maximum number of different locations any word can have in a given DNA sequence before it is stored to the blacklist file.

4.1 Input Parameters and Files for *gmasker*

The first required input file for *gmasker* is “human_repeats.list”—a preprocessed blacklist file created by *glistmaker*.

The next parameter is called a “maskingletter”. This can be almost any letter, typical examples are “N” or “X”. The only exception is “l” (or “L”), which triggers the lower case masking (3' ends of overrepresented words are in lower case, 3' ends of words with acceptable frequency are in upper case).

The third parameter, which the user must define, is “masking-type”. There are four possible options for defining the type of masking repeats in a sequence: forward, backward, both, and target. As the PCR primer is single-stranded and thus can bind to only one strand, we can mask nucleotides strand-specifically, depending whether we need primers for upper or lower (forward or backward) strand. It can be thought as masking repeat's 3'-end only and leaving some nucleotides from its 5'-end unmasked. Which end of repeat is 3'-end depends on whether we design primers for upper or lower strand (Fig. 1). The third masking type “both” mask repeats using both strands. The most useful option for primer design is “target”. This type masks upper strand in front of a target region and lower strand behind the target region. The target in this case is the region, which should be amplified with PRIMER3. START and END define start and end positions of the target region in a given input sequence.

Parameter “-nbases” defines the number of bases from 3' end of the primer candidate that *gmasker* should mask (default 1). The value of this parameter can vary from 1 to 16.

4.2 Changing Parameters in *gm_primer3*

Most of the *gm_primer3* default options are tuned to suit the situation where the user wants only good primers for very similar PCR conditions and would rather discard regions with no good primers than work with suspicious primer candidates. However, when the user is in a situation, where primers must be designed for very narrow region, parameter relaxation is inevitable. The user should relax the constraints that are least important in particular case and that

are most likely preventing primers from being acceptable. Here is the list of options we have commonly used to relax in our practice:

PRIMER_PRODUCT_SIZE_RANGE	to 1000
PRIMER_MAX_SIZE	to 30
PRIMER_MIN_TM	to 54
PRIMER_MAX_GC	to 80
PRIMER_MAX_POLY_X	to 5
PRIMER_MAX_DIFF_TM	to 5

4.3 Input Parameters and Files for *gindexer*

The first required input file—“inputfile”—is a FASTA file with one single sequence (e.g., the assembled chromosome sequence). Multiple FASTA files (e.g., multiple contig sequences) are not supported at the moment. One possible workaround is that you save each sequence to a separate file and create indexes for all these files. In any case, if you have to index more than one file, it would be practical to write a shell script that does it for each file.

The second parameter “outputfile” is the prefix for the output file. You can add directory names in front of the output file name. For each input file four different binary index files will be created—all words starting with A, C, T, and G nucleotide. For example, for 24 human chromosomes *gindexer* creates a total number of 96 files.

The parameter called “-wordsize” defines the length of the words that *gindexer* is using for creating indexes. Default word length for the indexes is 16, but it can be changed within the range between 8 and 16 nucleotides.

4.4 Input Parameters and Files for *gtester*

The first input file—“primerfile”—is a tab-delimited text file containing the following data columns: primer id, left PCR primer sequence, and right PCR primer sequence. Additional columns will be ignored.

The second input file is “locationsfile”. This is a text file containing file names that were indexed with *gindexer*. Each file name should be on separate line. An example of “locationsfile”:

```
/home/db/human_DNA/genometester/1
/home/db/human_DNA/genometester/2
/home/db/human_DNA/genometester/3
/home/db/human_DNA/genometester/4
...
/home/db/human_DNA/genometester/Y
```

Note that file names do not include extensions; *gtester* requires only chromosomal or genomic file names and adds extensions (a.location, t.location, c.location, and g.location) for each file by itself.

The third input file—“blacklistfile”—is a preprocessed index file created by *glistmaker*. This is an optional parameter that might speed up *gtester*. If this file name is defined, the *gtester* will not test and record locations of the words that are already listed in the blacklist. Please note that this index can only be used if the word length in *gindexer* location indexes is the same as in *glistmaker* blacklist.

The user can define the maximum product length of two primers that *gtester* is searching with a parameter called “-maxprodlen” (default 1,000 bp).

Additionally, the user can define the maximum number of primer binding sites that *gtester* tracks (default 1,000)—“-limit”. For shorter word lengths (less than 16 bp), the user should use larger cutoff value for the maximum number of binding sites.

Acknowledgements

The development of GenomeMasker package was supported by the Estonian Ministry of Education and Research grant 0182649s04 and grant EU19730 from Enterprise Estonia. Development of primer design software in our group has been funded by Centre of Excellence in Genomics at Estonian Biocentre (EU European Regional Development Fund). The authors thank Katre Palm for a valuable help with English grammar.

References

1. Dawson E, Abecasis GR, Bumpstead S et al (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 418:544–548
2. Rouchka EC, Khalyfa A, Cooper NG (2005) MPrime: efficient large scale multiple primer and oligonucleotide design for customized gene microarrays. *BMC Bioinformatics* 6:175
3. Xu D, Li G, Wu L et al (2002) PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics* 18:1432–1437
4. Weckx S, De Rijk P, Van Broeckhoven C et al (2005) SNPbox: a modular software package for large-scale primer design. *Bioinformatics* 21:385–387
5. Zhang Z, Schwartz S, Wagner L et al (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7:203–214
6. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. *Genome Res* 11:1725–1729
7. Kent WJ (2002) BLAT—the BLAST-like alignment tool. *Genome Res* 12:656–664
8. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27:573–580
9. Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
10. Rouillard JM, Zuker M, Gulari E (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res* 31:3057–3062
11. Van Hujum SA, De Jong A, Buist G et al (2003) UniFrag and GenomePrimer: selection of primers for genome-wide production of unique amplicons. *Bioinformatics* 19:1580–1582
12. Untergrasser A, Cutcutache I, Koressaar T et al (2012) Primer3 – new capabilities and interfaces. *Nucleic Acids Res* 40:e115
13. Breslauer KJ, Frank R, Blocker H et al (1986) Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A* 83:3746–3750
14. SantaLucia J Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA

- nearest-neighbor thermodynamics. Proc Natl Acad Sci U S A 95:1460–1465
15. Owczarzy R, You Y, Moreira BG et al (2004) Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. Biochemistry 43:3537–3554
16. Von Ahsen N, Wittwer CT, Schutz E (2001) Oligonucleotide melting temperatures under PCR conditions: nearest-neighbor corrections for Mg(2+), deoxynucleotide triphosphate, and dimethyl sulfoxide concentrations with comparison to alternative empirical formulas. Clin Chem 47:1956–1961
17. Kaplinski L, Andreson R, Puurand T et al (2005) MultiPLX: automatic grouping and evaluation of PCR primers. Bioinformatics 21:1701–1702

Chapter 2

Primer Design for PCR Reactions in Forensic Biology

Kelly M. Elkins

Abstract

The polymerase chain reaction (PCR) is a popular method to copy DNA in vitro. Its invention revolutionized fields ranging from clinical medicine to anthropology, molecular biology, and forensic biology. The method employs one of many available heat-stable DNA polymerases in a reaction that is repeated many times in situ. The DNA polymerase reads a template DNA strand and using the components of the reaction mix, catalyzes the addition of free 2'-deoxynucleotide triphosphate nitrogenous bases to short segment of DNA that forms a complement with the template via Watson–Crick base pairing. This short segment of DNA is referred to as a PCR primer and it is essential to the success of the reaction. The most widely used application of PCR in forensic labs is the amplification of short tandem repeat (STR) loci used in DNA typing. The STRs are routinely evaluated in concert with 16 or more reactions, a multiplex, run in one test tube simultaneously. In a multiplex, it is essential that the primers work specifically and accurately on the intended reactions without hindering the other reactions. The primers, which are very specific, also can be used to probe single nucleotide polymorphisms (SNPs) in a DNA sequence of interest by single base extension. Primers are often designed using one of many available automated software packages. Here the process of manually designing PCR primers for forensic biology using no-cost software is described.

Key words Molecular biology, Nucleic acids/DNA/RNA, Hydrogen bonding, Oligonucleotide, DNA polymerases, Polymerase chain reaction, PCR primer, STR, SNP

1 Introduction

The polymerase chain reaction (PCR) is employed in diverse fields including clinical medicine, anthropology, and forensic biology for DNA amplification and genetic analyses.

The method employs one of many available heat-stable DNA polymerases. The polymerase reads a template DNA strand and catalyzes the addition of free 2'-deoxynucleotide triphosphate nitrogenous bases (dNTPs) to a primer. The primer is a short segment of RNA or DNA that forms complementary base pairs with the template via Watson–Crick base pairing. Primers must be highly specific for the target sequence to be copied. Specificity can be increased by increasing the length of the primer and by selecting GC-rich sequences. Relatively short RNA primers of a length of 5–10

oligonucleotides are used *in vivo* while longer 18–35 oligonucleotide DNA primers are used *in vitro*. The longer length is important *in vitro* as the entire genome is denatured indiscriminately whereas *in vivo* the DNA template unwound by helicase at the replication fork is highly controlled and the entire genome is not available for amplification. The DNA primers are much more stable and less susceptible to chemical degradation than RNA primers. PCR can be used to amplify any locus in any genome, including those composed of DNA or RNA. A process called reverse transcriptase-PCR (RT-PCR) which substitutes the reverse transcriptase enzyme for DNA polymerase reverse-transcribes the RNA into cDNA.

PCR primers are required to copy DNA and bracket the DNA region of interest. In designing PCR primers, it is important to define the desired region and design primers upstream and downstream of the locus of interest. For the shortest amplicon, primers are designed from the 5' DNA region directly upstream from the locus (e.g., a STR, SNP, or other locus) and the 3' region directly downstream for the 5' and 3' primers, respectively. DNA polymerase grows both primers in the 5'-3' direction. The 5' primer is complementary to the bottom or minus strand of the DNA double helix and will start the new top or plus strand using the bottom strand as a template. The 3' primer is complementary to the top or plus strand of the DNA double helix and will start the new bottom or minus strand using the top strand as a template. The length of the amplicon is computed by adding the number of nucleotides in the region, including the primers, from end to end [1, 2].

In a multiplex, additional loci can be simultaneously copied by adding sets of primers to the reaction mixture provided the primers are structurally compatible in a common buffer. Multiplex PCR was first described in 1988 and over 50 loci have been demonstrated to amplify in a multiplex. Extension times are increased so the polymerase can fully copy all targets. Different dye labels can be added to the 5' end of the primer to differentiate amplicons, even those of the same length, by the unique fluorescence [3].

Of primary interest in forensic science is copying short tandem repeat (STR) microsatellite loci as they are highly variable and suitable for human DNA typing applications. Several commercial kits have been developed for the purpose identifying STR alleles at loci on the autosomes and sex chromosomes for use in human identification. Primers may also be designed to evaluate the base identity in single nucleotide polymorphisms (SNPs) in with much smaller PCR amplicons than STRs. As SNPs occur approximately every 1,000 bases in the human genome (ten times more frequent than STRs) with mutation rates a hundred 1,000 times lower than STRs, this category of genetic marker can be used for forensic DNA analysis in mass disaster and paternity cases. In forensic science, these have been used in mitochondrial DNA profiling and phenotype applications but Y chromosome SNPs (Y-SNPs) and X

chromosome SNPs (X-SNPs) have also been identified. PCR is also used in whole genome sequencing, whole genome amplification, and differentiation of body fluids and normal and cancer tissue applications. In other fields, PCR is also used to introduce site-directed mutants and foreign genetic material to the chromosome [1, 2].

Commercial DNA quantitation (e.g., Quantifiler® and Plexor® HY for human identification) and multiplex kits for STR DNA typing (e.g., AmpF/STR® Identifiler® and PowerPlex® 16) contain reaction pre-made mixtures and primers for multiple loci to be used simultaneously and require only the addition of template DNA [1]. They are attractive for their ease of use, certified quality, and capability for inter-laboratory data sharing and comparison. Routine screening and analysis is greatly assisted by the use of multiplex PCR. Researchers may also use PCR primers previously designed and employed by other groups and published in the literature. However, both of these approaches are expensive and confine analyses to only the loci and species covered by the kit or previous research. By designing their own primers, researchers have flexibility in their studies are not confined to using only primers used by other groups and published in the literature or to commercially available primers. New loci may be amplified and evaluated.

Primers are often designed using one of many available automated software packages including Primer 3 (Whitehead Institute), Primer3Plus, PrimerQuest (IDT), QuikChange Primer Design Program (Agilent), Primer Express (Applied Biosystems) Oligo (Molecular Biology Insights), Primer 1.2 (Indiana), Primer-BLAST (NCBI), and Primer Premier (PREMIER Biosoft), to name a few [4–9]. Others including OligoAnalyzer (IDT) and PerlPrimer calculate primer parameter values but will not automatically suggest primers for the locus of interest [10]. While most are no-cost, some do have a cost associated with their use. Each target region requires a unique set of two PCR primers. The primers are selected for a variety of characteristics including having similar melting/annealing temperatures, disinterest in forming primer dimers with themselves or each other, and specificity for the target. Although the software saves time and requires little user experience or expertise, they can be restrictive if an investigator is looking to produce an amplicon of a desired length or prepare a multiplex with primers for other loci, which requires an override of some of the parameters. In these cases, the researcher gains flexibility by designing the primers and only using the software to evaluate characteristics of the primers including melting temperature, GC content, primer dimer and heterodimer formation, and hairpin formation. It is important for researchers and forensic professionals to be able not only to use primers supplied in ready-made kits but also to be able to design a single set of primers as needed and create multiplexes of primers as necessary.

Table 1 shows the parameters evaluated in designing PCR primers and the optimal result or range. Optimal PCR primers will have a length of 18–35 nucleotide bases, annealing temperatures in the 55–72 °C range, melting temperatures (T_m) within 5 °C of each other, 40–60 % overall guanine–cytosine (GC) content, less than four identical base repeats within the primer, minimal primer

Table 1
Sample PCR primer results for D18S51

Parameter	Desired result	OligoAnalyzer result	Requirement met?
<i>D18S51 5' primer: 5'-CTC TGA GTG ACA AAT TGA GAC CTT G-3'</i>			
Length of primer (nucleotide bases)	18–35	25	Yes
Melting temperature (°C)	55–72	55.6	Yes
Percent GC content (%)	40–60	44.0	Yes
Primer dimer (contiguous bases)	<4	4	No
No long runs with the same base (contiguous nucleotide bases)	<4	3	Yes
Hairpin (°C)	Melting temperature ≪ annealing temperature	23.6	Yes
Unique oligo sequence search	Best match in Blast		Yes
<i>D18S51 3' primer: 5'-CTG GTG TGT GGA GAT GTC TTA CAA T-3'</i>			
Length of primer (nucleotide bases)	18–35	25	Yes
Melting temperature (°C)	55–72	56.7	Yes
Percent GC content (%)	40–60	44.0	Yes
Primer dimer (contiguous bases)	<4	3	Yes
No long runs with the same base (contiguous nucleotide bases)	<4	2	Yes
Hairpin (°C)	Melting temperature ≪ annealing temperature	19.4	Yes
Unique oligo sequence search	Best match in Blast		Yes
<i>Between the 5'/3' primers</i>			
Distance between primers on target DNA (nucleotide bases)	<450 (<2,000)	162	Yes
ΔT_m 5'/3' primer pairs (°C)	≤5	1.1	Yes
5'/3' hetero-dimer	<4	4	No

dimer and hairpin formation, and produce an amplicon that does not exceed 2,000-base pairs [1]. They will also demonstrate specificity for the desired locus when probed with NCBI-Primer BLAST.

There are many relevant loci in forensic biology. They include loci for which primers are available and others that have not been probed for forensic applications. All are accessible for study using PCR. Here the process of manually designing PCR primers to amplify a locus of interest in forensic biology in silico using no-cost software is described.

2 Materials

2.1 Websites

1. The procedure requires a computer or Web-device with an Internet connection. The websites used in this procedure include NCBI (<http://www.ncbi.nlm.nih.gov/>), OligoAnalyzer (<http://www.idtdna.com/Home/Home.aspx>), and NCBI Primer-Blast [7] (<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>).

2.2 Obtaining a DNA Sequence to Amplify from the Desired Locus

These items are necessary to obtain the double-stranded DNA sequence to be amplified by PCR, flanking regions and NCBI accession number. If you have a locus of interest, skip **item 1**. If you have the NCBI accession number for the locus of interest, skip **item 2**. If you have the double-stranded sequence of the locus of interest and the flanking DNA region, proceed to Subheading 3.

1. Select a STR, SNP, or other desired locus to amplify. STRBase is a good resource for forensic loci of interest [11] (see Note 1).
2. Obtain the NCBI accession number for the locus from the data in STRBase or a research paper describing the gene or locus of interest. The locus used in the example is NCBI Accession number L18333.
3. Obtain the selected SNP or STR DNA sequence and flanking region from the NCBI databank (<http://www.ncbi.nlm.nih.gov/>). The NCBI accession number, if known, or the gene name, locus or chromosome location is inputted in the search box. As an example using the NCBI accession number, in the search box at the top of the page, type the NCBI accession number “L18333.” This corresponds to the Combined DNA Index System (CODIS) STR locus D18S51 located on chromosome 18 used in human DNA typing [1]. Select “nucleotide” from the drop-down menu and then select the search box to perform the search. The top of the resulting page should read “Human chromosome 18 STS UT574, sequence tagged site”.
4. Scroll down to the word “ORIGIN,” located on the left side of the page. Figure 1 contains the sequence of the segment of the genome on chromosome 18 that contains the STR locus D18S51.

D18S51 NCBI Nucleotide Sequence*(Accession L18333)*

1 aattgagcnc aggagttaa gaccagcctg ggtaacacag tgagacccct gtctctacaa
 TTAACTCGNG TCCTCAAATT CTGGTCGGAC CCATTGTGTC ACTCTGGGA CAGAGATGTT
 61 aaaaatacaa aaatnagttg gcatggtgg cacgtgcctg tagtctcagc tacttgagg
 TTTTTATGTT TTTANTCAAC CCGTACCACC GTGCACGGAC ATCAGAGTCG ATGAACGTCC
 121 gctgaggcag gaggagttct tgagcccaga aggttaaggc tgcagtgagc catgttcatg
 CGACTCCGTC CTCCTCAAGA ACTCGGGTCT TCCAATTCCG ACGTCACTCG GTACAAGTAC
 181 ccactgcact tcactctgag tgacaaattt agaccttgc tcagaaagaa agaaagaaag
 GGTGACGTGA AGTGAGACTC ACTGTTAAC TCTGGAACAG AGTCTTCTT TCTTCTTTC
 241 aaagaaagaa agaaagaaag aangaaagaa agaaagtaag aaaaagagag ggaaagaaag
 TTTCTTCTT TCTTCTTTC TTNCTTCTT TCTTTCATTC TTTTCTCTC CCTTCTTTC
 301 agaaanagna aanaaatagt agcaactgtt attgtaagac atctccacac accagagaag
 TCTTNTCNT TTNTTATCA TCGTTGACAA TAACATTCTG TAGAGGTGTG TGGTCTCTTC
 361 ttaattttaa tttaaacatg ttaagaacag agagaagcca acatgtccac cttaggctga
 AATTAAAATT AAAATTGTAC AATTCTGTC TCTCTCGGT TGTACAGGTG GAATCCGACT
 421 cggttgttt atttgtgtt tgctggtag tcgggtttgt tattttaaa gtagcttatac
 GCCAAACAAA TAAACACAAC AACGACCATC AGCCCAAACA ATAAAAAATT CATCGAATAG
 481 caataacttca ttaacaattt cagtaagtta tttcatctt caacataaat acgnacaagg
 GTTATGAAGT AATTGTTAAA GTCATTCAAT AAAGTAGAAA GTTGTATTAA TGCNTGTTCC
 541 atttcttctg gtcaagacca aactaatatt agtccatagt aggagctaatt actatcacat
 TAAAGAAGAC CAGTTCTGGT TTGATTATAA TCAGGTATCA TCCTCGATTA TGATAGTGTA
 601 ttactaagta ttctatttgc aatttgactg tagcccatag cttttgtcg gctaaagtga
 AATGATTCA AAGATAAACG TTAAACTGAC ATCGGGTATC GGAAAACAGC CGATTCACT
 661 gcttaatgct gatcgactct agag
 CGAATTACGA CTAGCTGAGA TCTC

Fig. 1 Genome sequence for NCBI Accession Number L18333 for CODIS STR D18S51 locus region on human chromosome 18 with AGAA repeat (allele 13) shown in alternating grayscale with example underlined primers producing a 162 base pair product. The arrows show the primer direction

The DNA nucleotides are spaced in groups of ten and numbered on the left side in order of occurrence in the sequence. The sequence represents only the top strand or the 5'-3' sequence. Rather than work directly from the Web pages, it is helpful to highlight the entire one-letter nucleotide DNA sequence and copy the sequence to a text file to a word processing program of your choice [10]. Give the file a name and save your work (*see Note 2*).

5. Search for the selected locus. In the D18S51 STR example, the tetranucleotide repeat consists of “AGAA” repeated 13 times (allele 13) [1]. The repeats have been located and highlighted in gray in Fig. 1 (*see Note 3*). Note that at position 263, the base is “n” (*see Note 4*).
6. This complementary sequence can be quickly determined using the free software at <http://arep.med.harvard.edu/labgc/adnan/projects/Utilities/revcomp.html> (*see Note 5*). Copy the full DNA sequence and paste it in the box. Carefully remove all of the tracking numbers. Select complement. The complementary sequence will be returned (*see Note 6*). Figure 1 shows the complement (capital letters) below the retrieved NCBI L18333 sequence.

3 Methods

3.1 Design and Evaluation of PCR Primers

1. Copy an 18–35 nucleotide segment of the 5' or top strand directly before (upstream) of the STR repeat and a 18–35 nucleotide segment of the 3' or bottom strand directly after (downstream) of the STR repeat into the text file. These sequences will be evaluated for potential use as the 5' and 3' PCR primers, respectively (*see Note 7*), using OligoAnalyzer [10]. The primers evaluated in this example are underlined in Fig. 1.
2. Rewrite the bottom strand sequence to read from 5' to 3' (reverse of current direction) (*see Note 8*).
3. Evaluate the potential primers using free Web-based analysis tool OligoAnalyzer 3.1 (<http://www.idtdna.com/analyizer/applications/oligoanalyzer/>) (*see Note 9*). First, paste the 5' primer into the box. In the example, this is the underlined top strand sequence in Fig. 1. The default settings are used for all calculations. Select “Analyze”. Record the results for the length of the 5' primer, melting temperature, and GC content in the text file (*see Note 10*).
4. Perform a homodimer or “self-dimer” analysis and a “hairpin” analysis on the 5' primer. Record the results as shown in Table 1 (*see Note 11*).

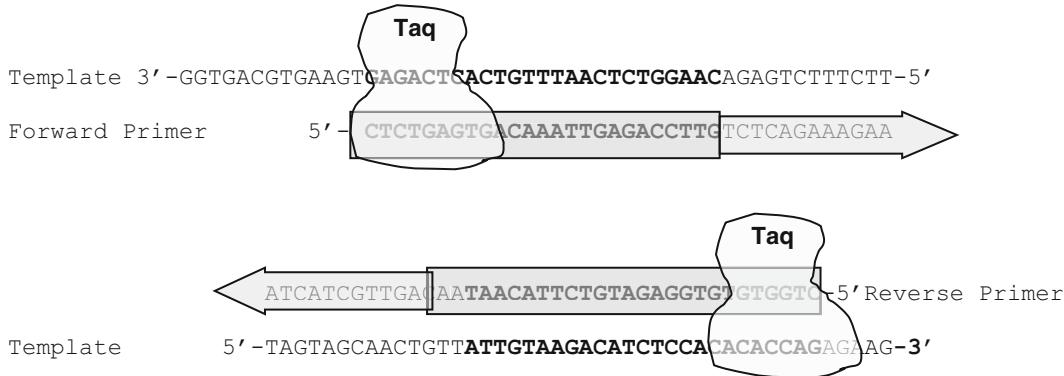


Fig. 2 Template with PCR primers (**bold**) annealing and direction of extension by *Taq*. The amplicon length is 162 base pairs including the primers

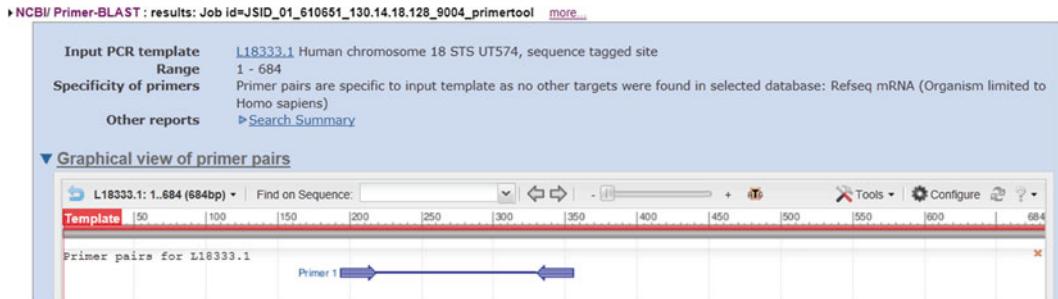
5. Repeat **steps 3** and **4** for the 3' primer. Record the results as shown in Table 1 (*see Note 12*).
6. Perform a “hetero-dimer” analysis on the 5' and 3' primers to evaluate their complementarity that may inhibit their abilities to act independently in chemical reactions (*see Note 13*). The results for the sample D18S51 primers are shown in Table 1 (*see Note 14*).
7. The amplification of the D18S51 primers by *Taq* DNA polymerase is shown in Fig. 2.

3.2 Evaluating Primers Compatibility for Multiplex Reactions

1. Create a set of primers for a second locus of interest as described in Subheading 3.1.
2. Using OligoAnalyzer 3.1, perform a heterodimer analysis for the four primers to be multiplexed (one for the first locus and two for the second locus) (e.g., 5'-1 with 5'-2, 3'-1 with 3'-2, 5'-1 with 3'-2 and 5'-2 with 3'-1) [10] (*see Note 15*).

3.3 Evaluating Nonspecific Priming by NCBI Primer-BLAST

1. Primer specificity is important and can be checked by comparing the primer sequences to the human (*Homo sapiens*) genome (or other genome of interest) for site complementarity using NCBI Primer-BLAST (<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>) [7]. Enter the NCBI accession number for the locus in the box at the top left of the screen. Then copy the primers to be checked under “Primer Parameters”. Using all of the other default parameters, scroll to the bottom of the screen and select “Get Primers”. This tool employs the software program Primer 3 to design primers computationally. Primer-BLAST will also design primers for a locus of interest automatically using user-defined or default parameters. Primers are specific if the program reports, “Primer pairs are specific to input template as no other targets were found in selected database.” Results for the primers used in the D18S51 example are shown in Fig. 3.



Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CTCTGAGTGACAAATTGAGACCTTG	Plus	25	194	218	59.82	44.00	5.00	1.00
Reverse primer	CTGGTGTGGAGATGTCTTACAAT	Minus	25	355	331	60.80	44.00	4.00	2.00
Product length	162								

Fig. 3 Results from NCBI Blast using the human genome to verify specificity of the primers for the desired locus in humans

3.4 Designing PCR Primers for Differentiation of SNPs

- SNPs can be probed by real-time PCR using primers with 3' ends that encompass the SNP using three primers [12, 13]. First, a SNP locus needs to be selected. For example, the following primers were designed to probe the most highly variable human mtDNA SNP at position 16519 (T/C) located between highly variable region (HV) HVI and HVII on the mitochondrial chromosome used in forensic science [2, 14].
- Identify forward and reverse primers for the locus as previously described in Subheading 3.1. The forward or 5' primer design will remain unchanged from what was previously described. For mtDNA SNP 16519 the forward primer is 5'-ACCACCATCCTCCGTGAAT-3' and begins at position 16399 on the chromosome [2].
- Next, prepare two reverse primers. The first primer is a direct complement to the template at the final base position of the SNP at the 3' end. Incorporate a base mismatch at the third base from the 3' end of both of the reverse primers to enhance specificity of the primers in PCR. The base incorporated in this example is a "C" and is shown in lower case. This works by destabilizing the extension of the doubly mismatched primer [2, 12, 13].
Reverse primer 1: 5'-CGTGTGGCTATTAGGCTTT AcGA-3'.
- Adjust one of the reverse primers to have a second base mismatch to the template at the final base position at the end of the reverse primer. This primer will complement the alternate SNP (C in this example) [2, 12, 13].

mtDNA SNP 16519 T/C NCBI Nucleotide Sequence*(Accession NC_012920)*

```

16381 TCAGATAGGG GTCCCTTGAC CACCATCCTC CGTGAATCA ATATCCGCA CAAGAGTGCT
AGTCTATCCC CAGGAACTG GTGGTAGGAG GCACTTTAGT TATAGGGCGT GTTCTACGA

16441 ACTCTCCTCG CTCCGGGCC ATAACACTT GGGTAGCTA AAGTGAACG TATCCGACAT
TGAGAGGAGC GAGGCCGGG TATTGTGAAC CCCCATCGAT TTCACATTGAC ATAGGCTGTA

16501 CTGGTTCCCTA CTTCAGGGT ATAAAGCCTA AATAGCCC ACGTTCCCT TAAATAAGAC
GACCAAGGAT GAAGTCCAG TATTCGGAT TTATCGGGT TGCAAGGGGA ATTTATTCTG

```

Fig. 4 The mtDNA SNP 16519T/C primers are diagrammed on the mitochondrial chromosome in Fig. 4. The SNP is indicated by the *bold letters* enclosed in the *box* and the primer annealing locations are *underlined*

Reverse primer 2: 5'-CGTGTGGCTATTAGGCTTT
AcGG-3'.

5. Add an 11-base pair GC-rich sequence (clamp) to the 5' end of one of the reverse primers. Here it was added to the doubly mismatched primer. The GC-rich clamp on the 5' end is underlined [2, 12].

Reverse primer 2: 5'-CGCGGCCGGCC-CGTGTGGCT
ATTAGGCTTAcGG-3'.

Alternatively, long and short GC-clamps can be added to both reverse primers, respectively [13].

6. The base identity in the GC-clamp SNP is detected by the differing lengths of the amplicons, which in turn causes a difference in the melting temperatures (approximately 4 °C) that can be probed using real-time PCR. The primers in the example produce a 145-base pair amplicon [2]. The primers are diagrammed on the mitochondrial chromosome in Fig. 4.
7. Check for primer specificity to the locus using NCBI Primer-BLAST as previously described in Subheading 3.3. Omit the GC-clamp sequence when pasting in the primer sequences to the boxes.

3.5 Obtaining PCR Primer Reagents

1. Primers may be purchased commercially from a variety of manufacturers. The author has used Integrated DNA Technologies (IDT) extensively as a source for PCR primers for the past 10 years. Custom primers purchased using the 25-nmole quantity and standard desalting methods have worked well for the lab's routine experiments. The final primer concentration in single template reaction should be in the range of 0.05–1 μM. The primers are quantified using UV-Vis spectroscopy and diluted to 5 μM stocks prior to use. 1 μL of primer stock is used for a 25 μL reaction for a final concentration of 0.2 μM. Higher concentrations may cause spurious amplification products and may cause secondary priming [3, 15] (see Note 16).

4 Notes

1. NIST's STRBase (<http://www.cstl.nist.gov/strbase/>) [11] is an excellent source of data for STR (http://www.cstl.nist.gov/strbase/promega_primers.htm) [16] and (<http://www.cstl.nist.gov/strbase/primer2.htm>) and SNP loci (<http://www.cstl.nist.gov/strbase/SNP.htm>) and their NCBI/GenBank accession numbers. The Web addresses may change. These worked at the time of publication.
2. Courier font of 10 point size is the easiest to read. Saving the sequence to a file simplifies designing the primers and introducing and tracking changes later.
3. It is helpful to use the “Find” function in the word processor to find the STR repeat or a segment of the gene or region of interest. Searching for one repeat allows you to quickly scan for the sequence of interest. When one repeat is located, look to the left and the right of the sequence for additional repeats [2, 10]. In this example, D18S51 should have 13 repeats of “AGAA” although one base could not be determined and is denoted “n”.
4. This means that the base could not be determined from the data from which this sequence was determined. Do not use a base sequence with a “n” in designing primers. The software will report an error and the primer may lose specificity depending upon the base chosen and the true base identity.
5. As DNA consists of a double helix formed by Watson–Crick base pairing, the complementary sequence can be quickly computed. Every guanine will hydrogen bond to cytosine and every adenine will hydrogen bond to thymine. Using the computer program speeds up the task. OligoAnalyzer will also write the complement for DNA sequences.
6. Copy the sequence into the text file so that the complement is directly beneath the top strand. The complementary strand is also referred to as the bottom strand or the 3'–5' strand.
7. Common mistakes include ordering the 3' (bottom strand) in the 3'–5' direction instead of the 5'–3' direction and designing the 3' primer from the top strand instead of the bottom strand. These mistakes will cause primers to fail.
8. Short amplicons are best copied with PCR. Primers closest to the repeat or locus that meets all of the criteria are best (Table 1) [1]. However, the primers may need to be repositioned to meet the melting temperature, homo-dimer, hetero-dimer, and hairpin requirements. If desired, PCR primers can be prepared from sequences further upstream or downstream

from the locus to increase the length of the amplicon. In the example, the 3' primer is positioned further downstream due to the long contiguous A and AG stretches.

9. This procedure uses OligoAnalyzer to calculate attributes of primers (e.g., melting temperature, primer dimer formation, GC-content, etc.) for primer design [10]. Other programs such as PerlPrimer also perform these calculations. As described in the introduction, there are numerous automated programs available for PCR primer design that require only the locus of interest. However, to understand the primer design process and effects of adjustments, it is helpful to design a few manually. A reliable internet connection is required for all steps unless the software is available for local installation.
10. Higher melting temperatures for the primers will promote more specific binding to the template and reduce potential side reactions [3]. The melting temperature was low initially (52.7°C) for the sequence selected so more bases were added (with a preference for more Gs and Cs) to increase the temperature to 55.6°C .
11. The melting temperature of the hairpin should be significantly lower than the annealing temperature for the primer in PCR that the hairpin is denatured in the reaction.
12. Notice that the selected 3' primer is further downstream because of the AT-rich sequence directly downstream of the STR repeat.
13. It is helpful to keep an electronic notebook for this procedure and especially for designing and redesigning PCR primers using Notepad or Microsoft Word or a comparable word-processing program to record the melting temperature, primer length, and number of continuous base pairs in homo- and heterodimers for each iteration of each primer. It is also helpful to record a few notes of what was changed (e.g., length, frameshift, lack of specificity) and why to track what was previously tested and the outcomes.
14. The size of the overall PCR product (primers and full sequence between two primers) should be less than 2,000 base pairs for end-point PCR. Short amplicons of less than 150 base pairs perform best in real-time PCR. Longer sequences will use up dNTPs more quickly which may cause incomplete strands. Increasing the quantity of dNTPs may increase the length of the amplicons produced but may decrease fidelity. Longer sequences also require combined longer annealing/extension times due to the increased sequence length [1–3, 15].

15. Primer design may require only minutes or extend to hours. There are no chemical hazards associated with PCR primer design as the work is performed in silico. Iteratively, shift the potential primer region and reevaluate the primers with the Web tools if the primers do not meet the optimal criteria [10].
16. The PCR primers must be verified experimentally. Primers are priced by the length (approximately USD \$0.35 per base). Shorter primers cost less. By using a shorter primer (e.g., 18–22 base pairs), the cost is minimized provided specificity is not compromised. For optimal results, the concentration of magnesium, dNTPs, primers, DNA polymerase must all be optimized although premade PCR master mixes often yield sufficient product to evaluate amplification of the correct product. Low concentrations of DNA template may cause stochastic results in PCR reactions so it is recommended that 1 ng of DNA be used in the PCR reaction. Amplification of the correct product (sequence and length) can be verified using DNA sequencing although the melting temperature in real-time PCR and using post-PCR capillary or slab gel electrophoresis can be used to verify amplicon length. 3 % NuSieve agarose gels and POP4 capillary electrophoresis work well for discriminating amplicon lengths for STRs while single base differences are better differentiated by POP6 or 6 % polyacrylamide gels. Multiplexing requires optimization of PCR reagents as well as primers. It is extremely time-consuming to design PCR primers that will work in tandem by hand; the time can be reduced by using primer design software. Even if the primers appear that they will work, experimental testing must be performed. Multiplex reactions require increased 2'-dideoxynucleotide triphosphates (dNTPs) concentrations and increased extension times for optimal results. The actual primer concentrations will need to be optimized for optimal results as some primers are preferentially extended (e.g., weak vs. strong loci) by the DNA polymerase. Decreasing the extension temperature has been shown to increase the yield for some loci [3, 15].

Acknowledgment

Thank you to the editors for inviting me to submit this chapter. This work was supported by start-up funds provided by Towson University (K.M.E.). The author is grateful to Zoë Krohn, Suzanne Gray, and Alison Eychner for testing the protocol as written.

References

1. Butler JM (2005) Forensic DNA typing, second edition: biology, technology, and genetics of STR markers. Elsevier Academic Press, Burlington, MA
2. Elkins KM (2012) Forensic DNA biology: a laboratory manual. Elsevier Academic Press, Waltham, MA
3. Henegariu O, Heerema NA, Dlouhy SR, Vance GH, Vogt PH (1997) Multiplex PCR: critical parameters and step-by-step protocol. *Biotechniques* 23:504–511
4. Kim TD (2000) PCR primer design: an inquiry-based introduction to bioinformatics on the World Wide Web. *Biochem Mol Biol Edu* 28:274–276
5. Lima AOS, Garcés SPS (2006) Intrageneric primer design: Bringing bioinformatics tools to the class. *Biochem Mol Biol Edu* 34: 332–337
6. Thornton B, Basu C (2011) Real-time PCR (qPCR) primer design using free online software. *Biochem Mol Biol Educ* 39:145–154
7. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden T (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13:1–11
8. Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols: methods in molecular biology. Humana, Totowa, NJ, pp 365–386
9. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM (2007) Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res* 35:W71–W74
10. Elkins KM (2011) Designing PCR primer multiplexes in the forensic laboratory. *J Chem Educ* 88:1422–1427
11. Ruitberg CM, Reeder DJ, Butler JM (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Res* 29:320–322
12. Papp AC, Pinsonneault JK, Cooke G, Sadée W (2003) Single nucleotide polymorphism genotyping using allele-specific PCR and fluorescence melting curves. *Biotechniques* 34: 1068–1072
13. Wang J, Chuang K, Ahluwalia M, Patel S, Umblas N, Mirel D, Higuchi R, Germer S (2005) High-throughput SNP genotyping by single-tube PCR with Tm-shift primers. *Biotechniques* 39:885–893
14. Coble MD, Just RS, O'Callaghan JE, Letmanyi IH, Peterson CT, Parsons TJ (2004) Single nucleotide polymorphisms over the entire mtDNA genome that increase the forensic power of testing in Caucasians. *Int J Legal Med* 118:137–146
15. Elkins KM, Kadunc RE (2012) An undergraduate laboratory experiment for upper-level forensic science courses: the use of TPOX single locus primers to amplify human DNA by real-time PCR with SYBR green detection. *J Chem Educ* 89:784–790
16. Masibay A, Mozer TJ, Sprecher C (2000) Promega Corporation reveals primer sequences in its testing kits [letter]. *J Forensic Sci* 45: 1360–1362

Chapter 3

Design of Primers and Probes for Quantitative Real-Time PCR Methods

Alicia Rodríguez, Mar Rodríguez, Juan J. Córdoba, and María J. Andrade

Abstract

Design of primers and probes is one of the most crucial factors affecting the success and quality of quantitative real-time PCR (qPCR) analyses, since an accurate and reliable quantification depends on using efficient primers and probes. Design of primers and probes should meet several criteria to find potential primers and probes for specific qPCR assays. The formation of primer-dimers and other non-specific products should be avoided or reduced. This factor is especially important when designing primers for SYBR® Green protocols but also in designing probes to ensure specificity of the developed qPCR protocol. To design primers and probes for qPCR, multiple software programs and websites are available being numerous of them free. These tools often consider the default requirements for primers and probes, although new research advances in primer and probe design should be progressively added to different algorithm programs. After a proper design, a precise validation of the primers and probes is necessary. Specific consideration should be taken into account when designing primers and probes for multiplex qPCR and reverse transcription qPCR (RT-qPCR).

This chapter provides guidelines for the design of suitable primers and probes and their subsequent validation through the development of simplex qPCR, multiplex qPCR, and RT-qPCR protocols.

Key words Quantitative real-time PCR, Primers, Probes, Software and databases, Validation, Reverse transcription real-time PCR

1 Introduction

Quantitative real-time PCR (qPCR) is widely and successfully used in clinical and biological fields for quantification of nucleic acid sequences (DNA or RNA). This is a sensitive and specific technique in which the DNA amount is monitored during the reaction by using fluorescent dyes that are incorporated into the PCR product. The increase in the fluorescent signal is directly proportional to the number of PCR product molecules generated [1]. The fluorescence monitoring through a qPCR reaction can be detected by nonspecific dyes, such as SYBR® Green, or by sequence-specific primers or probes coupled to fluorescent dyes, including hydrolysis probes, molecular beacons, fluorescence resonance energy transfer (FRET)

probes, and Scorpions primers [2]. SYBR® Green chemistry and TaqMan® hydrolysis probes are the most frequently used methodologies for developing qPCR protocols.

The success in any of the developed qPCR protocol depends on the suitability of the designed primers and probes, since the specificity of the technique is closely related to the annealing of primers to their complementary targets [3] and, afterwards, to the probe hybridization into newly synthesized DNA. In SYBR® Green assays, the proper design of primers is especially critical because the dye intercalates into double-stranded DNA without distinguishing between specific and nonspecific qPCR products [4–6]. In TaqMan® qPCR, the optimal design of probes is essential for their hybridization to the amplified target sequence to increases the specificity of the assay [4, 5].

Although the design of primers for qPCR is not substantially different from those for standard PCR, they need to meet special criteria for the reaction success [7]. Thus, they should allow strictly the synthesis of a single amplicon with good efficiency (ideally two copies of template after every PCR cycle) and without formation of primer-dimers. This is necessary for an accurate and reliable quantification of the target sequence.

To design primers and probes for qPCR, multiple software programs and websites are available being numerous of them free. These tools can be used to design primers and probes, test for non-specific priming, and assess the formation of secondary structures which might form between primers, probes, templates, or the amplification product [6]. Some of them are described in Subheading 3 of this chapter.

A proper design of primers and probes for qPCR requires sequential steps involved in this process including the selection of target sequences and primer and probe candidates followed by a validation process. The overall procedure of the qPCR primer and probe design is shown in Fig. 1 and is described in this chapter.

The design of primers and probes is especially critical for multiplex qPCR assays since more than one primer pair and probe set is included in the same reaction for amplifying two or more target sequences and consequently the probability of mispriming is higher [1].

On the other hand a special attention to primer and probe design is required for reverse transcription qPCR (RT-qPCR) which is used for gene expression analysis. In this method, RNA is transcribed into cDNA prior to its quantification using qPCR [8]. RT-qPCR can take place in one or two steps. The data can be quantified by absolute or relative methods which determine the selection of the proper primers as described in Subheading 5 of this chapter.

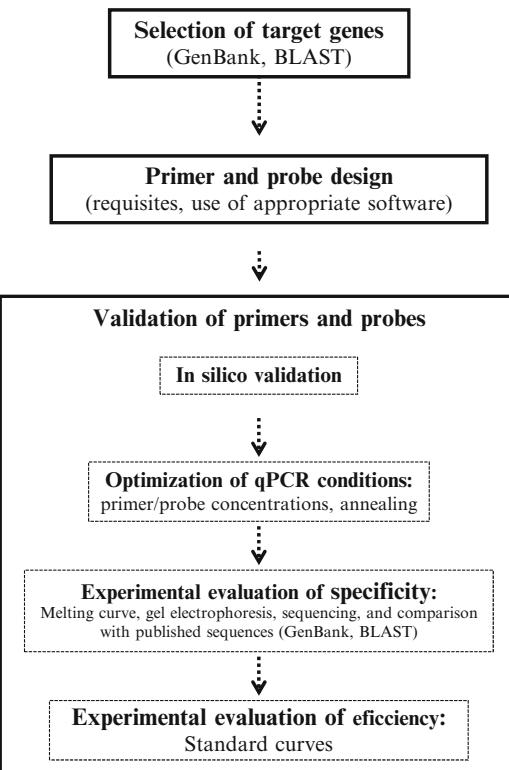


Fig. 1 Flowchart schematizing the procedure used for designing and validating primers and probes for quantitative real-time PCR protocols

2 Parameters to Be Considered When Designing Primers and Probes for Quantitative Real-Time PCR

The use of optimal primer and probe sequences is one of the critical steps for a successful qPCR assay. The criteria for primer and probe design are extensively described in the literature [2, 9–12]. A preliminary and key step when designing primers and probes for qPCR is the selection of target sequences of nucleic acid where they can hybridize. This is particularly important in the area of microbiology where targeting a gene highly conserved among different species can be used in broad-based detection strategies, while targeting a DNA sequence unique for a particular species or even strain can provide a highly specific test [13]. In this way, the target could be a sequence unique for a monophyletic group of microorganisms or could be based on a group of functional genes encoding a specific enzyme.

Once the selection of the target sequences has been done, the next step is to find potential primers or probes targeting regions on the corresponding gene sequences. This can be done manually

Table 1

Default requisites for designing primers and probes for quantitative real-time PCR assays (adapted from refs. 2, 10, 86)

Requisites	Primers	Probes
GC content	30–80 %	
Calculated primer/probe T_m	50–60 °C (always >55 °C) T_m of primers should not differ >2 °C	68–70 °C (8–10 °C above T_m primer)
Runs of identical nucleotides	Maximum 3 (No G bases)	
Primer/probe length	15–30 bp	
PCR product length	50–150 bp (optimum <80 bp)	
Distance forward primer to probe	50 bp	
Primer-dimers, hairpins	Avoid	Avoid
3' end rule (3' instability)	Maximum two G or C in the last 5 bp	–
Autoquenching	–	No G on the 5' end
GC ratio	–	C>G
Degree of degeneracy of bases	Avoid	Avoid

using sequence alignment program or automatically using primer design software. In both techniques it should be checked that the suggested primer and probe set achieves the following criteria: amplicon length, melting temperature (T_m), primer and probe length, GC content, self-complementary, primer-dimer and hairpin formation, degree of degeneracy, 5' end stability, and 3' end specificity. An overview of these criteria is given in Table 1 and they are described below.

Amplification products smaller than 150 bp are highly recommended for high efficiency of qPCR [11], although lengths below 80 bp are advisable. However, amplicons up to 400 bp may amplify efficiently. Shorter qPCR products amplify more efficiently than longer ones and are less susceptible to potential secondary structure within them [14]. This is because they are more likely to be denature during the extension step (at 92–95 °C), allowing the primers and probe to complete more effectively the binding to their complementary target sequences. Concretely, the length of the amplification products should be similar in multiplex qPCR protocols [6]. In addition, GC rich regions in the target sequence should be avoided since they are more difficult to amplify [15].

The T_m of the primers and probes consisting of the temperature at which 50 % of them and its complement are hybridized [6] is also an important parameter in the development of qPCR assays. A T_m value of the primers ranging from 55 to 60 °C is recommended for qPCR [12]. In addition, the T_m of sense and antisense

primers should be similar to avoid hairpins [16]. In particular, the T_m of the primer pair should not be more than 1 °C different from each other [6]. Since amplification primers are extended as soon as they bind to their complementary sequences, probe needs a T_m greater than primers to ensure strong binding of itself during the annealing phase [4, 17]. Concretely, the T_m of the probe should be 8–10 °C higher than those of the primers. The T_m of the primers and probe directly depends upon their length and their percentage of GC content [18].

The optimal size of qPCR primers and probes usually ranges from 15 to 30 nucleotides [19]. Shorter primers may decrease the specificity [16]. Despite longer primers could be match better to the target sequence the PCR amplification efficiency could be lower [11]. This reduction in the efficiency of the reaction, especially when using environmental DNA samples, may lead to a significant reduction in the yield and quality of the qPCR product [20]. In addition, the use of longer probes allows more mismatches, does not improve the sensitivity, can exhibit less efficient quenching, and produces lower fluorescence yields whilst use of shorter probes decrease the specificity [10, 11].

Regarding the GC content in the primers and probe sequences, a GC percentage between 30 and 80 % is recommended. In spite of the fact that primers with higher GC content should stabilize probe hybridization [10], they may not denature easily during PCR provoking a decrease of the amplification efficiency. Furthermore, poly-C and poly-G regions in the primers should be avoided since they can make up a tetraplex structure, which is very stable and cannot be transcribed by the polymerase [21]. Non-specific priming can be minimized by selecting primers that have only one or two G/C within 3' end last five nucleotides [2], since a higher GC content at this end of the primer may prevent the complete annealing of the remainder of the primer sequence and reduce the specificity of the reaction [16]. However, it is recommended that primers have a G or C as nucleotide on the 3' end to ensure their correct and strong binding to the template [11]. The presence of a G as nucleotide at the 5' end of the probe should be excluded to avoid a continuous quenching even after probe cleavage, which resulting in reduced normalized fluorescence values [22]. Furthermore, the probe should contain more C than G because of high change of normalization fluorescence (ΔR_n). It allows low positive signals which can be more easily differentiated from the background signal [23].

Primers and probes with a high possibility of self-complementarity, particularly close to the 3' end, should be avoided because secondary structures, such as hairpins, can be formed and interfere the extension step. Moreover, intramolecular and intermolecular interactions between the primers can generate primer-dimers which should be considered in the design process [24, 25]. This is a common artifact in qPCR reactions which occurs when two primers

bind to each other instead of to the template. In addition, the probe should never overlap with or be complementary to either of the primers. On the other hand, it is advisable to avoid more than five interactions between the primers, especially at the 3' end position.

The presence of degenerate nucleotides in primers and probes should be excluded in the design process. Differences in the GC content at degenerate positions in the primer target regions of the template DNA could affect the amplification [26, 27]. However, sometimes a certain degree of degeneracy is necessary in order to prevent some under-estimation of target sequences when that has non-conserved regions.

Concerning multiplex qPCR assays, the above mentioned criteria must be considered. However it should be taken into account that in this type of reactions multiple templates and several primer and probe sets are in the same reaction. The presence of multiple primers and probes may lead to interactions with each other and the possibility of mispriming with other templates. For this reason, it is important to ensure that the different primer and probe sets do not exhibit complementarity to one another [28]. Thus a special care must be taken to design proper primers and probes and to select appropriate reporter dyes and quenchers for the probes. Regarding the last concern, three criteria should be considered: (a) the probes should be labelled with reporter dyes whose fluorescence spectra are well separated or show only minimal overlap, (b) selected combinations of reporter dyes and quenchers should be compatible with the detection abilities of the real-time cycler, and (c) non-fluorescent quenchers should be used [8, 28].

3 Software and Other Bioinformatics Tools to Design Primers and Probes for qPCR

In the design of primer and probe from a common gene-specific region, all known sequences in the public databases should be first selected and then aligned to find conserved regions. This may suppose a high time-consuming activity if it is not automated. In addition, although primers and probes seem to generate acceptable results at first, many home-made or “do-it-yourself” primers often come up short in their specificity, qPCR amplification efficiency, reproducibility, and sensitivity. Therefore, there are currently many online and commercial bioinformatics tools for routine use. Software automatically checks for the best primers and probes for considered parameters and provides a list of them. They often control most of the default requirements for primers and probes previously described in Subheading 2 (CG content, primer and probe length, primer and probe T_m , CG 3' end terminal enforcement, etc.). Additionally several programs take into account other main parameters for a more accurate and comprehensive selection of primers and probes, such as the general nucleotide structure of primers such as linguistic complexity (nucleotide arrangement

and composition), specificity, the melting temperature of the whole primers and the melting temperature at the 3' and 5' termini, self-complementarity, and secondary binding [29].

There are multiple primer and probe design tools available on the net that allow producing high quality primers (Table 2). Though most of them are freely available, they have variable quality and some of them are not well-maintained. This often results in missing links and sites that may have been useful previously but they may not be functional at a later date [30]. These programs can be used to generate potential primers and probes, check for non-specific hybridization, and evaluate the formation of secondary structures. On the other hand, it has to be taken into account that the use of these online programs requires practice since online guides may not be available to support novice users in designing primers and probes [6]. In general, these tools are very effective, yielding success rates well in excess of 95 % in the hands of experienced users [31].

Several companies supplying primers and probes offer Web-based tools for their design and free applications are also available on the net [30]. Despite the fact that most of the algorithms considered by them have been conceived for standard PCR, they are also helpful for qPCR primer and probe design [32]. According to Gubelman et al. [33] an ideal qPCR primer and probe design program should at least include the following features: (a) all annotated splice variants of each gene to enable either gene or transcript specific expression profiling should be considered, (b) for RT-qPCR assays at least one primer needs to span exons to avoid amplification of contaminating genomic DNA, (c) the specificity of primers and probes needs to be automatically assessed by similarity search, (d) no cumbersome post-processing should be required to retrieve the best primer combination, and (e) the location of primer pairs within their genomic context should be visualized for easy and final evaluation by the end user.

3.1 Software and Programs for Designing of Primers and Probes for qPCR

Next some of the most suitable software for supporting in the design of primers and probes for qPCR are going to be briefly described.

OLIGO software is the first computer application that performed on the market for designing primers and probes. However *OLIGO* went through many transformations to latest software in 2010, *OLIGO* 7. Based on nearest-neighbor thermodynamics, *OLIGO*'s search algorithms find optimal primers for PCR, qPCR (TaqMan® probes), and sequencing. *OLIGO* 7 searches also for hybridization, ligase chain reaction probes and molecular beacons and even siRNAs [34].

Primer3 is one of the most commonly used primer design software [35]. It is a frequently updated, open-source project and used by many Web-based applications to develop useful functions for primer and probe design [12]. Its popularity is likely due to several factors that include the availability of a relatively easy-to-use Web

Table 2
Some of the available commercial and online software for quantitative real-time PCR primer and probe design and their websites

Name	URL	References
ABI PRISM Primer Express	http://www.lifetechnologies.com/order/catalog/ product/4363993?ICID=search-product	[48]
AlignMiner	http://www.scbi.uma.es/alignminer/	[59]
ConservedPrimers 2.0	http://probes.pw.usda.gov/ConservedPrimers/	[50]
EasyExonPrimer	http://129.43.22.27/~primer/EasyExonPrimer.html	[53]
EcoPrimer	http://www.grenoble.prabi.fr/trac/ecoPrimers	[52]
DATFAP	http://cgi-www.daimi.au.dk/cgi-chili/datfap/frontdoor.py	[56]
DFold	http://dfold.cgb.ki.se/	[31]
Gemi	http://sourceforge.net/projects/gemi/	[46]
GETPrime	http://updepla1srv1.epfl.ch/getprime/	[33]
Java Web Tools	http://primerdigital.com/tools/	[58]
MultiPriDe (Multiple Primer Design)	Available upon request to aziesel@emory.edu	[43]
OLIGO 7	http://www.oligo.net/	[34]
PerlPrimer	http://perlprimer.sourceforge.net/	[38]
PRaTo	http://prato.daapv.unipd.it/	[32]
Primer3	http://biotools.umassmed.edu/bioapps/primer3_www.cgi	[35]
Primer3Plus	http://primer3plus.com	[37]
Primer3web	http://primer3.wi.mit.edu http://bioinfo.ut.ee/primer3/	[36]
PrimerBank	http://pga.mgh.harvard.edu/primerbank/	[57]
Primer-Blast	http://www.ncbi.nlm.nih.gov/tools/primerblast/index.cgi?LINK_LOC=BlastHomeAd	[87]
PrimerCE	http://tch.hebau.edu.cn/shengm/download/down.html	[54]
PrimerDesign	http://www.hiv.lanl.gov/tools/primer/main	[47]
PUNS (Primer-UniGene Selectivity)	http://okeylabimac.med.utoronto.ca/PUNS	[40]
RTPrimerDB	http://medgen.ugent.be/rtpimerdb/	[55]
QPrimerDepot	http://primerdepot.nci.nih.gov/ http://mouseprimerdepot.nci.nih.gov/	[88]
QuantPrime	http://www.quantprime.de/	[42]
RASE	http://designs.lgfus.ca/cgi-bin/bsp_designs/index.pl	[45]
TOPSI	http://www.bhsai.org/downloads/topsi.tar.gz	[44]

service, robust engineering, open access to the program source code, suitability for use in high-throughput pipelines for genome-scale research, and the simplicity for incorporating into or interoperating with other software [36]. This software has been recently enhanced with new Web interfaces, *Primer3Plus* and *Primer3Web* [36, 37]. The most notable enhancements incorporate accurate thermodynamic models in the primer design process, both to improve melting temperature prediction and reduce the likelihood that primers will form hairpins or dimers [36].

PerlPrimer is a cross-platform graphical user interface application for the design of primers for qPCR as well as standard PCR, bisulfite PCR, and sequencing. This program combines accurate primer-dimer prediction algorithms with powerful tools such as sequence retrieval from Ensembl genome database and the ability to BLAST search primer pairs [38, 39]. Using the default settings, PerlPrimer searches for small amplicons (100–300 bp) which span an intron–exon boundary and possess at least one primer hybridizing across an intron–exon boundary.

PUNS (Primer-UniGene Selectivity) software is a CGI/Perl-based Web server to perform in silico PCR on PCR primer sequences. PUNS server simulates PCR reactions by running BLAST analysis on user-entered primer pairs against both the transcriptome and the genome to assess primer specificity. PUNS is particularly suited for the identification of highly selective primers for microarray experiments which are usually carried out either by semiquantitative or by RT-qPCR [40]. The use of PUNS in primer design follows a three-step process. Firstly users enter primers into the database. Users then submit their primer sequences for a BLAST analysis [41]. Finally, the information in a primer pair is combined by an in silico PCR which identifies potential amplicons by both identity and size. The in silico PCR report allows deciding if potential primer pairs are accepted or rejected for experimental use [40].

QuantPrime is an intuitive and user-friendly, fully automated for using in primer pair and probe design for qPCR analyses. QuantPrime can be used online or on a local computer after downloading. QuantPrime specifically tests primer pairs for qPCR, developed to satisfy needs of advanced users in low to high-throughput transcript profiling experiments, while keeping the user interface very simple providing important features missing in other available software and Web services. The public QuantPrime server is currently set up with publicly available transcriptome and genome annotations from 295 different eukaryotic species [42]. The parameter flexibility for designing and specificity testing offered in QuantPrime makes it straightforward to be used in the design of oligonucleotides for additional quantification applications, such as qPCR with hydrolysis probes (e.g., TaqMan® probes, Scorpion primers) or quantitative in situ hybridization of mRNA. Such protocols are added to QuantPrime as program gather experimental data and feedback from users.

MultiPriDe (Multiple Primer Design), a Perl tool that accepts batch lists of Gene database identifiers, collects available intron and exon position data critical to RT-qPCR primer development and supplies these sites as identified targets for the Primer3 utility to maximize successful primer design [43].

TOPSI (Tool for PCR Signature Identification) is a computationally efficient, fully integrated tool for the design of qPCR-based pathogen diagnostic assays. The TOPSI pipeline efficiently designs qPCR primers and probe sets common to multiple bacterial genomes by obtaining the shared regions through pairwise alignments between the input genomes [44]. TOPSI uses pairwise alignments to identify sequences that are common to multiple genomes and compares these sequences with non-target genomes to identify unique segments suitable for designing signatures.

RASE (Real-Time PCR Annotation of Splicing Events) is a pipeline that allows accurate identification of a large number of splicing isoforms in human cell lines and tissues [45]. The RASE automatically designs specific primer pairs for 81 % of all alternative splicing events in the NCBI build 36 database. With this program a quick identification of splicing isoform signatures can be obtained in different types of human tissues. However, this program does not enable the design of gene-specific primers. In addition, its associated Web interface only supports low-throughput experiments [33].

GETPrime is a primer database supported by a novel platform that uniquely combines and automates several features critical for optimal qPCR primer and probe design. These include the consideration of all gene splice variants to enable either gene-specific or transcript-specific expression profiling, primer specificity validation, automated best primer pair selection according to strict criteria, and graphical visualization of the latter primer pairs within their genomic context [33]. This program is very useful due to the fact that it combines and automates all of the important features required to address the increasing demands in qPCR primer design for high-throughput qPCR experiments, especially those requirement to target genes in gene- or transcript-specific fashion without post-processing [33].

Gemi is an automated, fast, and easy-to-use bioinformatics tool with a user-friendly interface to design primers and probes based on multiple aligned sequences. This tool can be used for the purpose of both conventional and qPCR and can deal efficiently with great number of large size sequences [46]. The main criterion used by Gemi to identify primers and probes is the nucleotide sequence belonging to conserved DNA, but it provides the dissociation temperature, length, and GC percentage in the final output file for each to select primers or probes. The application executes directly on a computer and provides a simple and user-friendly interface allowing an easily and quickly primer design. This tool can be particularly useful in the microbiology field [46].

PrimerDesign is a novel computer program for designing primers and probes for highly variable DNA targets. The design takes into account genetic variation and several user-specified as well as automatic design features related to the aim of a particular study and the intended experimental setting. It has been reported as useful tool for designing primers and probes for biological systems with high levels of genetic variation [47]. The overall software procedure proceeds through interconnected steps: (a) the target locations for primers and probes are determined guided by sequence entropy estimates and complexity, (b) primer melting temperatures are optimized, (c) bio-barcodes and adaptors are added, and (d) risks of dimerization are estimated. Each interconnected step informs the subsequent steps. In addition, if previous steps have to be reoptimized, the information to next steps occurs automatically [47].

ABI PRISM Primer Express 3.0 is a primer and probe design tool. It allows designing oligonucleotides for qPCR applications using a customized application specific document for absolute/relative quantification and allelic discrimination assays. Besides the primers, ABI PRISM Primer Express 3.0 helps in designing the labelled probes, selection of the appropriate reagents, use of universal thermal cycling parameters, and use of default primer and probe concentrations (or optimizing if necessary). The Primer Express software includes a Primer Test document that allows evaluating primers for their T_m , secondary structure, and primer-dimer formation [48].

DFold is a software that creates PCR primers without stable secondary structures [31]. DFold combines the use of Primer3 [35] for assessing of PCR primers and the MFold package [49] for predicting secondary structures.

ConservedPrimers 2.0 was developed as application able to design large numbers of PCR primers in exons flanking one or several introns on the basis of orthologous gene sequences in genetically closed species. This program has been developed for designing intron-flanking primers for large-scale single nucleotide polymorphism (SNP) discovery and marker development [50]. This tool uses non-redundant expressed sequence tags (EST) and related genomic sequences as inputs. Intron-flanking primers are then designed based on the intron-exon information using the Primer3 core program [35] or BatchPrimer3 [51].

EcoPrimer is a software which fulfills all the requirements for designing new barcode regions suitable for metabarcoding studies. This software has the ability to scan large training databases, since it is used to design highly conserved primers to amplify variable DNA regions [52].

EasyExonPrimer is a Web-based software that automates the design of PCR primers to amplify exon sequences from genomic DNA. It uses Primer3 [35] to design PCR primers based on the genome builds and annotation databases available at the University of

California, Santa Cruz (UCSC) Genome Browser database (<http://genome.ucsc.edu/>). It masks repeats and known SNP sites in the genome and designs standardized primers using optimized conditions [53].

PrimerCE is a reliable primer design program that specifically fulfills the need for gene cloning aimed at produce proteins. The main applications of PrimerCE include inspection of restriction enzyme recognition sequence, open reading frame verification, stop codon inspection, base adjustment, primer optimization, sequence assembly, and protein analysis [54].

3.2 Other Commercial Bioinformatics Tools for Designing Primers and Probes for qPCR

Besides software for designing qPCR primers and probes, there are additional bioinformatics tools and also databases gathering a huge amount of validated primers and probes which prevent to spend time in their design and experimental optimization and validation. Several of these additional tools are described below.

RTPrimerDB is a public database for primer and probe sequences used in qPCR assays engaging popular chemistries (SYBR® Green, TaqMan®, and Molecular Beacon) to reduce time-consuming primer and probe design and experimental optimization. In addition, this program introduces a certain level of uniformity and standardization among different laboratories [55]. RTPrimerDB includes records with user submitted assays that are linked to genome information from reference databases and quality controlled using an in silico assay evaluation system. The primer evaluation tools intended to assess the specificity and detect features that could negatively affect the amplification efficiency are combined into a pipeline to test custom designed primer and probe sequences. An improved feedback system guides users and submitters to enter practical remarks and details about experimental evaluation analyses [55].

DATFAP (Database of Transcription Factors with Alignments and Primers) is a free, Web-based, and very user-friendly browsing tool based on a new database of more than 55,000 EST (expressed sequence tag) sequences from 13 plant species, classified as transcription factors. Further, the database offers primers and probes designed for qPCR as well as homology alignments and phylogenies for the sequence analysis [56]. DATFAP is equipped with a sophisticated search facility and specific primers for almost all sequences, DATFAP constitutes a valuable tool to researchers in all areas of plant molecular biology working with transcription factors. No other multi-species transcription factor database offers such easy interspecies and intraspecies navigation in the network of related transcription factors [56].

PrimerBank is a robust bioinformatics process for primer design. The algorithm has been used to design many qPCR primers to cover the most known human and mouse genes, all of which

are freely accessible via the PrimerBank website. PrimerBank primers have been designed and validated to perform at an invariant annealing temperature of 60 °C. The expression profiles of thousands of genes can be simultaneously determined, making the primers useful for high-throughput nanoliter-scale qPCR platforms. In addition, PrimerBank contains a high number of experimentally validated primers, comprising the largest collection of its kind in the public domain [57].

qPrimerDepot is a qPCR primer database which provides optimized primers for all genes in the human and mouse Reference Sequence collection (RefSeq). The primers are designed to amplify desired targets under unified annealing temperature in order to facilitate their application in large-scale high-throughput assays. In addition, qPrimerDepot allows designing specific primers to perform gene expression studies using RT-qPCR.

PRaTo is a simple to use and easy interpret Web-based tool that enables checking and ranking of primer pairs because of their attitude for an optimal and reliable performance when used in qPCR experiments. It can be used as a stand-alone tool or in association with software for primer and probe design or for calculating oligonucleotide properties [32].

Java Web tool (jPCR) is based on the FastPCR software for Windows [58] and provides a more flexible approach for designing primers and probes for many applications. It checks if either primers or probes have secondary binding sites in the input sequences that may give rise to an additional PCR product. The jPCR tool eliminates intraoligonucleotide and interoligonucleotide reactions before generating a list with primer pair candidates. This is very important for qPCR efficiency since production of stable and inhibitory primer-dimers is predicted and can be avoided, particularly the complementarity in the 3' end of primers whence the DNA polymerase will extend [29]. Primer-dimer prediction is based on the analysis of non-gap local alignment and the stability of both the 3' end and the central part of the primers.

AlignMiner is a Web-based application to detect matching (convergent) and divergent regions in alignments of conserved sequences focusing particularly on divergence. Virtually without exception, all available tools focus on conserved segments or residues. Though small divergent regions are biologically important for specific qPCR, genotyping, etc., they have received little attention. As a consequence, they must be selected empirically by the researcher [59]. This software tries to cover the gap in bioinformatics function by evaluating divergence, rather than similarity, in alignments of closely related sequences. Hence, it is expected that its usage will ensure an objective selection of the best-possible divergent region when closely related sequences are analyzed, saving researchers' time of analysis [59].

3.3 Advances in the Selection of Software for Designing Primers and Probes for qPCR

The optimal design of primers and probes for qPCR using some of the above programs is essential to ensure specific and efficient amplification of the amplification products. Thus, the advantages and disadvantages of the above described software should be carefully checked before selection, with special emphasis in avoiding primer containing secondary structures. Taylor et al. [60] recommend the use of Primer-Blast, a NCBI's program that uses the algorithm Primer3 [35]. The program MFold has been reported as appropriate to be used to analyze amplicons for potential secondary structures in RT-qPCR primer design [60].

Furthermore, new advances in research about primer and probe design should be added progressively to different algorithm programs. In this sense, these programs should consider annealing failure caused by single nucleotide variant (SNV) situated inside the primer sequences. Novel allele dropout mechanism causing genotyping errors originated by a non-primer-binding-site SNV has been recently reported by Lam and Mak [61]. These authors emphasize the need of the next generation of primer and probe design software to be able to analyze the secondary structure of primers, probes, and template sequence taking SNV in all the sequences to avoid secondary hairpin structure formation of the PCR products and amplification failure.

4 Validation of Primers and Probes Designed for qPCR

After designing qPCR primers and probes by using the available tools, in silico validation (BLAST specificity analysis, <http://blast.ncbi.nlm.nih.gov/Blast.cgi>) has to be conducted to confirm the specificity of targeted gene sequences. BLAST algorithm allows carrying out sequence-similarity searches against several databases, returning a set of gapped alignments with links to full database records [10]. The query coverage and the maximum identity should be 100 %. The BLAST program reports a statistical significance, called the “expectation value” (*E*-value) for each alignment. This is an indicator of the probability for finding the match by chance. *E*-values ≤ 0.01 normally suggest homologous sequences [41, 62]. The *E*-value is a widely accepted measure for assessing potential biological relationship [10]. In addition to BLAST, other in silico tools could be used for validating the designed primers and probes as previously mentioned in Subheadings 3.1 and 3.2.

Furthermore, in silico PCR tools could be used for predicting the potential PCR products and searching of possible mispriming of the designed primers or probes as it has been described in the previous section. Despite the fact that in silico tools provide valuable feedback, the specificity of the qPCR assay using the designed primers and probes has to be validated empirically with direct experimental evidence as described below [63].

The next factor to take into consideration for optimal qPCR results should be reagent optimization including primer and probe concentration. To select the optimal concentrations of them for qPCR it is necessary to check the obtained amplification plots and select the combination showing the lowest value of quantification cycle (C_q), the cycle in which fluorescence reaches a defined threshold [64], and the highest fluorescent signal for a fixed target concentration [65]. Primer concentrations are normally between 50 and 300 nM owing to the fact that higher concentrations could promote mispriming, nonspecific amplification product accumulation, and lower concentrations primer exhaustion [2]. The optimal concentration for both primers could be different in a qPCR protocol [65]. The optimal probe concentration should be estimated after optimizing primer concentration. Probe concentrations normally vary between 50 and 250 nM [66], 250 nM being the optimal one. When the concentration is too low, no fluorescent signal will be observed and if it is too high a high fluorescent background could be detected [65].

After optimizing primer and probe concentration in the qPCR protocol, the optimal cycling conditions must be determined. Although the optimal annealing temperature is determined by the primer design software, it can differ greatly from the experimental annealing temperature [65]. Thus, an optimization could be necessary. It is recommended testing several annealing temperatures, starting around 5 °C below the T_m , to determine the optimal experimental annealing conditions [67] with which the efficiency of the qPCR method meets the criteria listed below.

The specificity of a qPCR protocol can be affected by the presence of nonspecific amplification products produced by primers binding to apparently random sites in the sample DNA other than the intended target or sometimes to themselves forming primer-dimers. Specificity of amplification products can be checked by analyzing the melting curves, also called dissociation curves, generated in those qPCR protocols based on double-stranded DNA-binding dyes including SYBR® Green, since they can bind to primer-dimers and other reaction artifacts producing a fluorescent signal [4, 5]. The melting curves can be carried out in all reported software programs for performing qPCR reactions immediately after amplification [68]. The specific amplicon in absence of primer-dimers appears as one single and narrow peak in the obtained melting curve (Fig. 2a) [10, 68]. If unintended amplification products are present they show a relatively lower T_m value than that expected for the amplicon or broader peaks are visualized in the melting curve (Fig. 2a) [1, 18]. Nonspecific PCR products melt at lower temperature values than the desired products mainly because of GC content and length [1, 18, 69]. The presence of primer-dimers can be experimentally demonstrated by comparing the T_m value of the checked template with that of no template

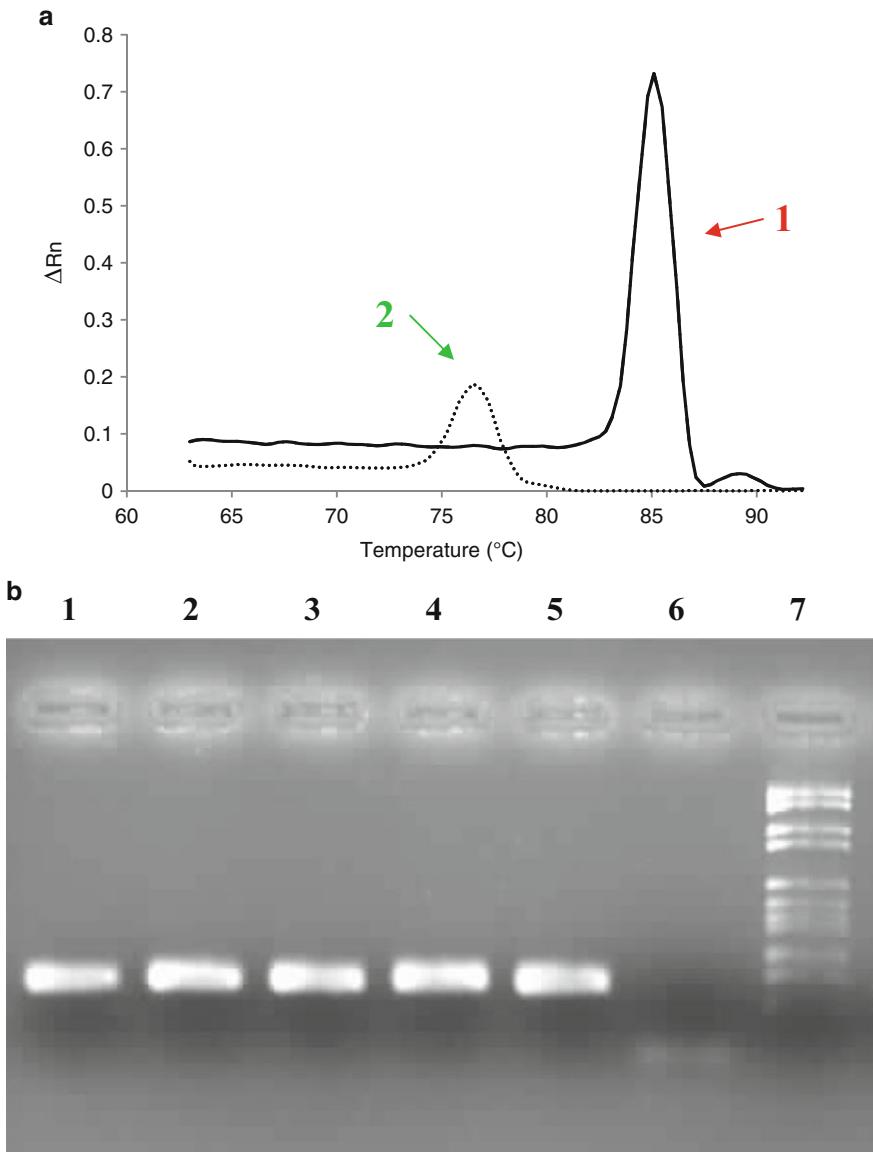


Fig. 2 (a) Melting curves showing specific amplification of the target sequence and primer-dimer formation in the no template control (NTC) sample. 1: Sample containing DNA; 2: NTC sample. (b) Agarose gel analysis to investigate primer-dimer formation. Lines 1–5: qPCR products obtained by using SYBR® Green methodology; Line 6: NTC sample with the presence of primer-dimer observed as diffuse band at the bottom of the gel; Line 7: DNA molecular size marker of 2.1–0.15 kbp (Roche Farma S.A.)

control (NTC) [1], because this artifact is much more common when template is not present. When using SYBR® Green chemistry or other double-stranded DNA binding dyes the absence of reaction artifacts has to be confirmed by using gel electrophoresis analysis since it has a higher resolution than melting curve analysis (Fig. 2b) [69]. Only a PCR product of the expected size must be

visualized in the gel when nonspecific products are detected. In spite of the fact that nonspecific PCR products do not affect to the fluorescent signal in probe-based assays the analysis of PCR product using gels is necessary since the PCR results will still be affected by the presence of nonspecific amplification. Furthermore, for a precise verification of the amplification, the specific PCR product must be sequenced, followed by a comparison of the obtained sequence with published sequences in GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>; [10]) using some of the bioinformatics tools detailed in Subheading 3.

Nonetheless, despite efforts for designing proper primers, primer-dimers or other nonspecific amplification products could be generated [1]. In this case, reaction conditions could be modified for reducing this kind of artifacts. Thus, a PCR protocol incorporating a hot-start, where an inactive DNA polymerase is activated at the start of qPCR by incubation at high temperature, could be used when performing a SYBR® Green protocol [8, 70]. This allows avoiding an early extension of primer complexes by the DNA polymerase. When this kind of enzyme is not used, reactions could be prepared on ice and the thermal cycler preheated to 95 °C before adding the reaction tubes or plates [67]. Another possibility for reducing the primer-dimer presence consisting of performing the fluorescence acquisition at a temperature higher than primer-dimer T_m , but lower than the T_m of the expected amplicon could solve this problem [71].

After that, the efficiency of the qPCR protocol has to be evaluated because an unsuitable measurement of selected target sequence fully invalidates the assay. The efficiency of a qPCR reaction should be 100 %, meaning during the exponential amplification two copies from every available templates are generated with each cycle [64]. Under experimental conditions efficiency may be as close to this value as possible. However several factors including primer characteristics may influence on it. Thus, one of the factors affecting the ability of qPCR for quantification is the efficiency of the designed primers and probes. When they are inefficient they should provoke imprecise qPCR efficiency.

Estimation of the efficiency of a qPCR method is based on constructing standard curves. Most qPCR instruments have software able to elaborate automatically a standard curve and calculate the efficiency of the reaction. If it is not available, the standard curve can be constructed by plotting the C_q values against a series of increasing and known concentrations of the template (tenfold serial dilutions of nucleic acid). For this at least four but preferably six or more points should be included [72]. The amplification efficiency can then be calculated from the formula $E = [10^{-1/S}] - 1$ where S is the slope of the standard curve [63, 73]. Generally slopes between -3.1 and -3.6 with PCR efficiency values in the range of 90–110 % are considered satisfactory (Fig. 3) [10, 72].

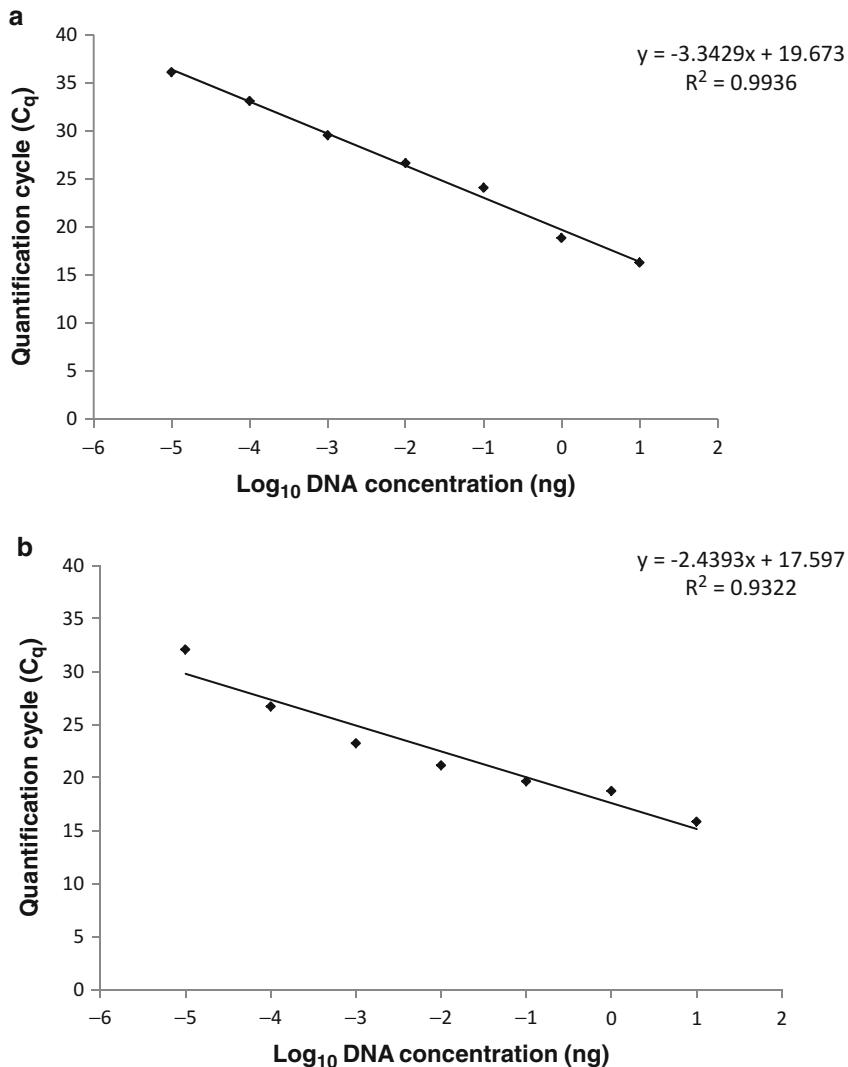


Fig. 3 Examples of an optimized standard curve (a) and an inappropriate standard curve (b) constructed with DNA standards of known concentration. Quantification cycle (C_q) values are plotted against the logarithm of seven 10-fold dilutions of standard DNA. Standards are measured in triplicate for each concentration

Furthermore, the optimal correlation coefficient (R^2) derived from the standard curve has to be between 0.99 and 0.999 [10].

A special consideration should be made for multiplex qPCR, since more problems related to primers and probes could occur such as a higher formation of nonspecific products [2]. Thus, when optimizing the protocol additional steps, such as including a higher amount of magnesium or of the hot-star enzyme if it is used, could be necessary [1], since they become limiting in later cycles and the amplification of the less efficient or less abundant target is compromised. Consequently, it is advisable to perform a primer-limiting assay to find the primer concentration giving the lowest possible C_q .

value for the more abundant target without distorting the C_q value of the less abundant target [66]. Regarding the validation of the designed primers and probe for multiplex qPCR assays, before their combination in a multiplex PCR assay primers and probes for each target should be validated in single runs and their individual efficiencies determined. After that the efficiency of the overall multiplex qPCR assay should be performed [1]. The values obtained for a given target in the individual and multiplex assays should not differ significantly. If the C_q values from the individual and multiplex assays are significantly different, reactions need to be optimized [28].

5 Design of Primers and Probes for Gene Expression Studies Using Real-Time Reverse Transcription PCR (RT-qPCR)

RT-qPCR has proven to be a powerful method to gene expression analysis [25, 74] which is increasingly important in a variety of clinical and biological research fields [75]. To avoid missing any gene expression, primers must detect every alternative transcript and splicing variant of the target genes [76].

RT-qPCR can be performed in one-step or in two-step. In a one-step procedure the transcription and the amplification of the target sequences are carried out in one reaction. However in the two-step protocols, cDNA synthesis is firstly obtained by RT of RNA and a cDNA aliquot is then used for amplifying the specific target [77].

Priming of the cDNA can be performed using oligo-dT, random primers, or target-specific primers depending on whether one-step or two-step is used and the choice of primer can provoke marked variation in calculated mRNA copy number [77, 78]. Therefore, one-step RT-qPCR is always performed with gene-specific primers [8]. In two-step RT-qPCR, the three types of primers or their mixtures could be used in the RT step prior to cDNA amplification using gene specific primers by means of qPCR. Gene-specific primers yield the most specific cDNA and provide the most sensitive quantification method [77, 79].

The method selected for analyzing the data derived from RT-qPCR will also influence upon the design of the primers and probes. The data analyses can be either of absolute levels to determine the absolute transcript copy number or relative levels to measure differences in the expression level of a specific target between different samples [80]. For absolute quantification, only target gene specific primers and/or probes are necessary since a RNA standard curve of the gene of interest is required. However, for relative quantification, target and endogenous specific primers and/or probes must be designed [81].

The design of gene-specific primers and/or probes to be used in the RT-qPCR should fulfil certain requirements apart from those

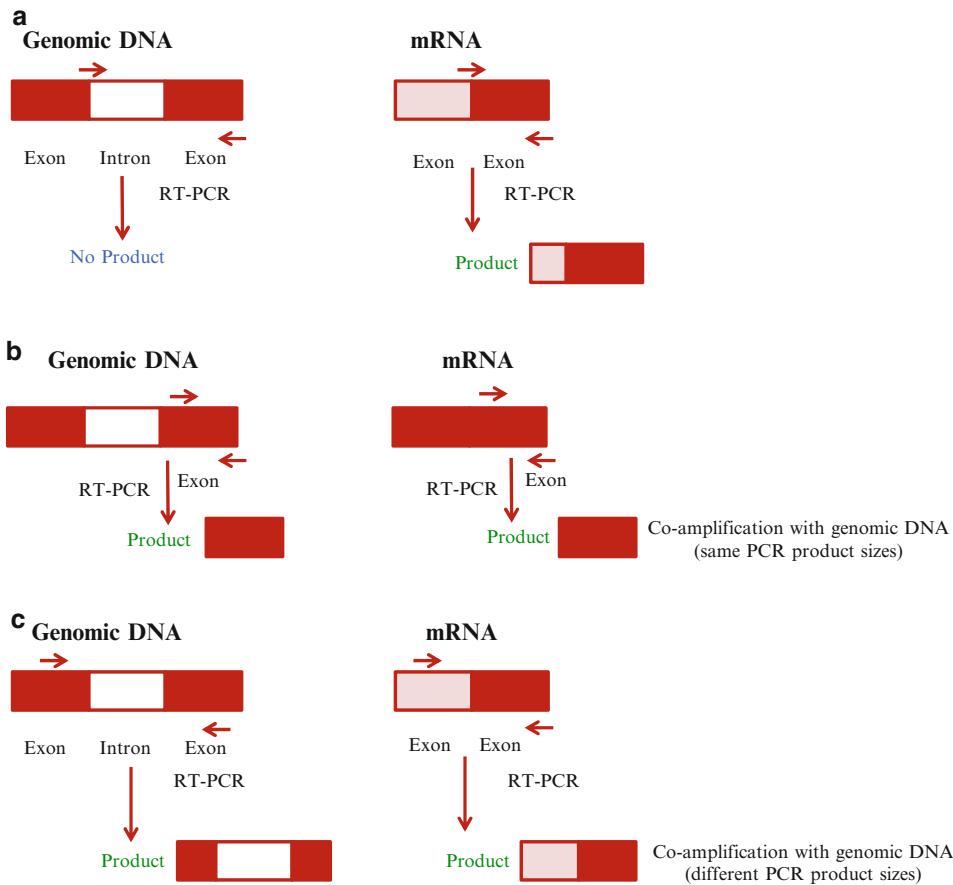


Fig. 4 Primer design to remove amplification of potential contaminating genomic DNA when using reverse transcription real-time PCR (adapted from ref. 8). (a) Forward primer crosses an intron–exon boundary. (b) Forward and reverse primers span within the same exon. (c) Forward and reverse primers span two different exons containing an intron

detailed in the above sections. These primers and/or probes could be designed using the specific software for them detailed in Subheading 3 and then validated as described in Subheading 4. For this kind of qPCR a proper primer and probe design is even more important, since the specific target (and the reference one when it is required) sequences should be unique, the length of the amplification product should be between 75 and 150 bp with a GC content of 50–60 %, and not containing secondary structures [60]. In addition, the primers and probes either should span an exon–exon splice junction enabling amplification and detection of RNA sequences only or they should be designed within the same exon (Fig. 4) [82]. In the first kind of primer and probe design, genomic DNA can be excluded as a template in a RT-qPCR reaction because primers or probe will bind to cDNA synthesized from sliced mRNAs but not to genomic DNA [8]. Nevertheless in the second one

contaminating genomic DNA could serve as a template resulting in a co-amplification with the cDNA (Fig. 4) and it is necessary to decide if genomic DNA is sufficiently negligible being necessary to treat the template RNA with RNase-free DNase [83]. Alternatively, primers and probes for RT-qPCR protocols can be designed to flank a region containing at least one intron (Fig. 4) [8]. Products amplified from cDNA without introns will be smaller than those amplified from genomic DNA which contains introns. If possible, a target with very long introns should be selected. Therefore, the RNA target may then be preferentially amplified because of the higher PCR efficiency of this shorter PCR product without introns. As previously described, if genomic DNA contamination is detected a treatment of the RNA sample with RNase-free DNase should be performed. Otherwise, the primers and probes should be redesigned to avoid amplification of genomic DNA. In addition, the use of Mn²⁺ rather than Mg²⁺ minimizes any problems caused by amplification of reannealed DNA [84]. Thus, the correct design of primers and/or probes for RT-qPCR assays could prevent co-amplification of genomic DNA avoiding a reduction of the assay sensitivity and specificity by competition of the intended PCR product and the product derived from genomic DNA [10].

The optimization of the concentration of primers and/or probes used in RT-qPCR reactions is crucial for performing gene expression analyses. The optimal primer and probe concentrations are ranging between 50–200 nM and 100 nM, respectively [85]. Depending on the selected method to perform the RT-qPCR (one- or two-step) and analyses of data (absolute or relative quantification) derived from it, different optimization steps taking into account the potential intramolecular and intermolecular interactions between primers, probes, and/or templates would be required. Thus, the one-step RT-qPCR requires the same primer concentration for RT and qPCR, reducing flexibility in primer concentrations optimal for multiplexing. In addition, in this kind of RT-qPCR both gene-specific primers have a higher tendency to dimerize at 42–50 °C RT conditions. This can be especially problematic in reactions using DNA-binding dyes for detection [70]. However, in two-step RT-qPCR, the qPCR primer concentration may be optimized for multiplexing, without having any adverse effect on RT [70]. Concerning the method for data analyses from RT-qPCR in order to analyze unique specific target sequence, the absolute one must be performed using an individual assay, and the relative one could be performed using either individual or multiplex assays, designing specific primers and probes taking into consideration all criteria previously described [83].

The validation process of the RT-qPCR assays for gene expression studies using the primers and/or probes previously designed is influenced by the chosen method for analyzing gene expression data. For validating a RT-qPCR when the absolute quantification is

used, a RNA standard curve plotting C_q values against several concentrations of the obtained cDNA is required to calculate the number of copies. This standard curve should be evaluated according to the criteria described for qPCR in Subheading 4. If the qPCR efficiency obtained is not in the optimal range, either new primers and/or probes should be designed or reagent and thermal conditions should be optimized. However, for an appropriate validation of a RT-qPCR which uses the $2^{-\Delta C_T}$ method for relative quantification two assumptions should be met [81]. If these assumptions are not fulfilled, then new primers and/or probes should be selected and redesigned. The correct selection of the reference gene and the design of primers and probes targeting the above gene are essential for carrying out the relative expression analysis. Thus properly selected reference genes will normalize differences in the amount and quality of starting material as well as in the reaction efficiency. Normalization uses reference genes with the assumptions that their expression is: (a) similar between all samples in a given study, (b) is resistant to experimental conditions, and (c) undergoes all steps of the qPCR with the same kinetics as the target gene [85].

Finally, the RT-qPCR assays for relative expression analyses could be performed using individual or multiplex assays. To undertake a multiplex assay several requirements should be met as follows: (a) the expression level of the reference gene must be greater than that of the target gene and (b) the gene that is more highly expressed (reference gene) should be setup with its primers at a limiting level. This ensures an accurate quantification of both genes since the competition between targets is excluded. In order to test that the reference gene is more abundantly expressed than the target gene, it should be tested that samples span the expected range of target gene expression. If the experiment is not successful, new primers and/or probes from the tested reference gene or other new ones consistently expressed in the sample have to be evaluated or the primers and/or probes designed on the basis of the reference and target genes may be run in separate wells (individual assays) [83].

6 Conclusions

Guidelines for designing primers and probes for qPCR are revised in this chapter, since this is the one of the most critical factor affecting the success and ability for quantifying of this PCR technique. The parameters to be considered when designing primers and probes have been profoundly described, highlighting special criteria which should be met if these primers and probes are used for multiplex qPCR. A brief description of some of the numerous available software programs and bioinformatics tools for designing primers and probes has been given. However new advances in research focused on this subject should be progressively added to

different algorithm programs for a higher suitability of the designed primers and probes. Once primers and probes have been designed, as detailed in this chapter a special attention has to be done to their validation process for obtaining successful results of the qPCR. Finally, a special remark has been done in the design of proper primers and probes for RT-qPCR protocols and their validation.

Acknowledgments

We acknowledge financial support of this work by projects “AGL2010-21623” and “Carnisenusa CSD2007-00016—Consolider Ingenio 2010” of the Spanish Government and GR10162 of the Government of Extremadura and FEDER.

References

1. Invitrogen (2008) Real-time PCR: from theory to practice. <http://corelabs.cgrb.oregon-state.edu/sites/default/files/Real%20Time%20PCR.From%20Theory%20to%20Practice.pdf>. Accessed 6 Nov 2013
2. Rodríguez-Lázaro D, Hernández M (2013) Real time PCR in food science: introduction. *Curr Issues Mol Biol* 15:25–38
3. Rosadas C, Cabral-Castro MJ, Vicente AC et al (2013) Validation of a quantitative real-time PCR assay for HTLV-1 proviral load in peripheral blood mononuclear cells. *J Virol Methods* 193:536–541
4. Holland PM, Abramson RD, Watson R et al (1991) Detection of specific polymerase chain reaction product by utilizing the 50–30 exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc Natl Acad Sci U S A* 88: 7276–7280
5. Heid CA, Stevens J, Livak KJ et al (1996) Real time quantitative PCR. *Genome Res* 6: 986–994
6. Thornton B, Basu C (2011) Real-time PCR (qPCR) primer design using free online software. *Biochem Mol Biol Educ* 39:145–154
7. Nolan T, Hands RE, Bustin SA (2006) Quantification of mRNA using real-time RT-PCR. *Nat Protoc* 1:1559–1582
8. Qiagen (2010) Critical factors for successful real-time PCR. <http://www.qiagen.com/es/resources/resourcedetail?id=f7efb4f4-fbcf-4b25-9315-c4702414e8d6&lang=en>. Accessed 9 Nov 2013
9. Yu Y, Lee C, Kim J et al (2005) Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. *Biotechnol Bioeng* 89:670–679
10. Raymaekers M, Smets R, Maes B et al (2009) Checklist for optimization and validation of real-time PCR assays. *J Clin Lab Anal* 23:145–151
11. Lim J, Shin SG, Lee S et al (2011) Design and use of group-specific primers and probes for real-time quantitative PCR. *Front Environ Sci Eng* 5:28–39
12. Chuang LY, Cheng YH, Yang CH (2013) Specific primer design for the polymerase chain reaction. *Biotechnol Lett* 35:1541–1549
13. Hanna SE, Connor CJ, Wang HH (2005) Real-time polymerase chain reaction for the food microbiologist: technologies, applications, and limitations. *J Food Sci* 70:49–53
14. Toouli CD, Turner DR, Grist SA et al (2000) The effect of cycle number and target size on polymerase chain reaction amplification of polymorphic repetitive sequences. *Anal Biochem* 280:324–326
15. McConlogue L, Brow MA, Innis MA (1988) Structure-independent DNA amplification by PCR using 7-deaza-20-deoxyguanosine. *Nucleic Acids Res* 16:9869
16. Mitsuhashi M (1996) Technical report: Part 1. Basic requirements for designing optimal oligonucleotide probe sequences. *J Clin Lab Anal* 10:277–284
17. Wittwer CT, Herrmann MG, Moss AA et al (1997) Continuous fluorescence monitoring of rapid cycle DNA amplification. *Biotechniques* 22:130–131
18. Ririe KM, Rasmussen RP, Wittwer CT (1997) Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. *Anal Biochem* 245:154–160
19. Wu JS, Lee C, Wu CC et al (2004) Primer design using genetic algorithm. *Bioinformatics* 20:1710–1717

20. Marchesi JR (2001) Primer design for PCR amplification of environmental DNA targets. In: Rochelle PA (ed) Environmental molecular microbiology: protocols and applications. Horizon Scientific Press, Wymondham, pp 43–54
21. Simonsson T, Pecinka P, Kubista M (1998) DNA tetraplex formation in the control region of c-myc. *Nucleic Acids Res* 26:1167–1172
22. Giulietti A, Overbergh L, Valckx D et al (2001) An overview of real-time quantitative PCR: applications to quantify cytokine gene expression. *Methods* 25:386–401
23. Gunson RN, Collins TC, Carman WF (2006) Practical experience of high throughput real time PCR in the routine diagnostic virology setting. *J Clin Virol* 35:355–367
24. Saiki RK (1989) The design and optimization of the PCR. In: Erlich HA (ed) PCR technology: principles and applications for DNA amplification. McMillan Publishers (Stockton Press), New York, NY, pp 7–22
25. Kubista M, Andrade JM, Bengtsson M et al (2006) The real-time polymerase chain reaction. *Mol Asp Med* 27:95–125
26. Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* 64:3724–3730
27. Linhart C, Shamir R (2005) The degenerate primer design problem: theory and applications. *J Comput Biol* 12:431–456
28. Biorad (2013) qPCR assay design and optimization. <http://www.bio-rad.com/en-es/applications-technologies/qpcr-assay-design-optimization>. Accessed 24 Oct 2013
29. Kalendar R, Lee D, Schulman AH (2011) Java web tools for PCR, in silico PCR, and oligonucleotide assembly and analysis. *Genomics* 98:137–144
30. Abd-Elsalam KA (2003) Bioinformatic tools and guideline for PCR primer design. *Afr J Biotechnol* 2:91–95
31. Fredman D, Jobs M, Strömqvist L et al (2004) DFold: PCR design that minimizes secondary structure and optimizes downstream genotyping applications. *Hum Mutat* 24:1–8
32. Nonis A, Scortegagna M, Nonis A et al (2011) PRaTo: a web-tool to select optimal primer pairs for qPCR. *Biochem Biophys Res Commun* 415:707–708
33. Gubelmann C, Gattiker A, Massouras A et al (2011) GETPrime: a gene- or transcript-specific primer database for quantitative real-time PCR. *Database* 2011:bar040. doi:[10.1093/database/bar040](https://doi.org/10.1093/database/bar040)
34. Rychlik W (2007) OLIGO 7 primer analysis software. In: Yuryev A (ed) Methods in molecular biology, vol 402, PCR primer design. Humana, Totowa, NJ, pp 35–59
35. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365–386
36. Untergasser A, Cutcutache I, Koressaar T et al (2012) Primer3: new capabilities and interfaces. *Nucleic Acids Res* 40:e115
37. Untergasser A, Nijveen H, Rao X et al (2007) Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res* 35:W71–W74
38. Marshall OJ (2004) PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics* 20:2471–2472
39. Marshall OJ (2007) Graphical design of primers with PerlPrimer. In: Yuryev A (ed) Methods in molecular biology, vol 402, PCR primer design. Humana, Totowa, NJ, pp 403–414
40. Boutros PC, Okey AB (2004) PUNS: transcriptomic- and genomic-in silico PCR for enhanced primer design. *Bioinformatics* 20: 2399–2400
41. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
42. Arvidsson S, Kwasniewski M, Riaño-Pachón DM et al (2008) QuantPrime: a flexible tool for reliable high-throughput primer design for quantitative PCR. *BMC Bioinformatics* 9:465
43. Ziesel AC, Chrenek MA, Wong PW (2008) MultiPriDe: automated batch development of quantitative real-time PCR primers. *Nucleic Acids Res* 36:3095–3100
44. Vijaya SR, Kumar K, Zavaljevski N et al (2010) A high-throughput pipeline for the design of real-time PCR signatures. *BMC Bioinformatics* 11:340
45. Brosseau JP, Lucier JF, Lapointe E et al (2010) High-throughput quantification of splicing isoforms. *RNA* 16:442–449
46. Sobhy H, Colson P (2012) Gemi: PCR primers prediction from multiple alignments. *Comp Funct Genomics* 2012:783138. doi:[10.1155/2012/783138](https://doi.org/10.1155/2012/783138)
47. Brodin J, Krishnamoorthy M, Athreya G et al (2013) A multiple-alignment based primer design algorithm for genetically highly variable DNA targets. *BMC Bioinformatics* 14:255
48. Applied Biosystems (2004) Primer Express software version 3.0. getting started guide. <http://www.bu.edu/picf/files/2010/11/Primer-express-30.pdf>. Accessed 10 Jan 2005
49. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31:3406–3415
50. You FM, Huo N, Gu YQ et al (2009) ConservedPrimers 2.0: a high-throughput pipeline for comparative genome referenced intron-flanking PCR primer design and its application in wheat SNP discovery. *BMC Bioinformatics* 10:331

51. You FM, Huo N, Gu YQ et al (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* 9:253
52. Riaz T, Shehzad W, Viari A et al (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Res* 39:e145
53. Wu X, Munroe DJ (2006) EasyExonPrimer: automated primer design for exon sequences. *Appl Bioinformatics* 5:119–120
54. Cao Y, Sun J, Zhu J et al (2010) PrimerCE: designing primers for cloning and gene expression. *Mol Biotechnol* 46:113–117
55. Lefever S, Vandesompele J, Speleman F et al (2009) RTPrimerDB: the portal for real-time PCR primers and probes. *Nucleic Acids Res* 37:D942–D945
56. Fredslund J (2008) DATFAP: a database of primers and homology alignments for transcription factors from 13 plant species. *BMC Genomics* 9:140
57. Wang X, Spandidos A, Wang H et al (2012) PrimerBank: a PCR primer database for quantitative gene expression analysis, 2012 update. *Nucleic Acids Res* 40:D1144–D1149
58. Kalendar R, Lee D, Schulman AH (2009) FastPCR software for PCR primer and probe design and repeat search. *Genes Genomes Genomics* 3:1–14
59. Guerrero D, Bautista R, Villalobos DP et al (2010) AlignMiner: a web-based tool for detection of divergent regions in multiple sequence alignments of conserved sequences. *Algorithms Mol Biol* 5:24
60. Taylor S, Wkem M, Dijkman G et al (2010) A practical approach to RT-qPCR: publishing data that conform to the MIQE guidelines. *Methods* 50:S1–S5
61. Lam CW, Mak CM (2013) Allele dropout caused by a non-primer-site SNV affecting PCR amplification: a call for next-generation primer design algorithm. *Clin Chim Acta* 421:208–212
62. Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A* 87:2264–2268
63. Bustin SA, Benes V, Garson JA et al (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* 55:611–622
64. Mallona I, Weiss J, Egea-Cortines M (2011) pcrEfficiency: a web tool for PCR amplification efficiency prediction. *BMC Bioinformatics* 12:404
65. Edwards KJ (2004) Performing real-time PCR. In: Edwards K, Logan J, Saunders N (eds) *Real-time PCR, an essential guide*. Horizon Bioscience, Norfolk, pp 71–83
66. Applied Biosystems (2010) Real-time PCR systems. Reagent guide. https://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_052263.pdf. Accessed 7 Jul 2010
67. Promega Corporation (2009) Protocols & applications guide. http://www.promega.com/~media/files/resources/paguide/letter/paguide_us.pdf?la=en. Accessed 21 Oct 2013
68. Pfaffl MW (2004) Quantification strategies in real-time PCR. In: Bustin SA (ed) *A-Z of Quantitative PCR* (IUL Biotechnology, No. 5). International University Line (IUL), San Diego, CA, pp 87–112
69. Lee MA, Squirell DJ, Leslie DL et al (2004) Homogeneous fluorescent chemistries for real-time PCR. In: Edwards K, Logan J, Saunders N (eds) *Real-time PCR, an essential guide*. Horizon Bioscience, Norfolk, pp 31–70
70. Life Technologies Corporation (2012) Real-time PCR handbook. http://find.lifetechnologies.com/Global/FileLib/qPCR/RealTimePCR_Handbook_Update_FLR.pdf. Accessed 6 Nov 2013
71. Rajeevan MS, Ranamukhaarachchi DG, Vernon SD et al (2001) Use of real-time quantitative PCR to validate the results of cDNA array and differential display PCR technologies. *Methods* 25:443–451
72. Kavanagh I, Jones G, Nayab SN (2011) Significance of controls and standard curves in PCR. In: Kennedy S, Oswald N (eds) *PCR troubleshooting and optimization: the essential guide*. Caister Academic Press, Norfolk, pp 67–78
73. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* 1:29–45
74. Gadkar VY, Filion M (2013) New developments in quantitative real-time polymerase chain reaction technology. *Curr Issues Mol Biol* 8:1–6
75. Ishii T, Sootome H, Shan L et al (2007) Validation of universal conditions for duplex quantitative reverse transcription polymerase chain reaction assays. *Anal Biochem* 362:201–212
76. Quellhorst, G., Rulli, S. (2008) A systematic guideline for developing the best real-time PCR primers. *SABiosci.* <http://www.sabiosciences.com/manuals/RT2performanceWhitePaper.pdf>. Accessed 26 Aug 2013
77. Bustin SA, Nolan T (2004) Analysis of mRNA expression by real-time PCR. In: Edwards K, Logan J, Saunders N (eds) *Real-time PCR, an essential guide*. Horizon Bioscience, Norfolk, pp 125–184
78. Zhang J, Byrne CD (1999) Differential priming of RNA templates during cDNA synthesis markedly affects both accuracy and reproducibility of

- quantitative competitive reverse-transcriptase PCR. *Biochem J* 337:231–241
79. Lekanne Deprez RH, Fijnvandraat AC, Ruijter JM et al (2002) Sensitivity and accuracy of quantitative real-time polymerase chain reaction using SYBR green I depends on cDNA synthesis conditions. *Anal Biochem* 307:63–69
80. VanGuilder HD, Vrana KE, Freeman WM (2008) Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques* 44:619–626
81. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods* 25:402–408
82. Wang X, Seed B (2003) A PCR primer bank for quantitative gene expression analysis. *Nucleic Acids Res* 31:e154
83. Applied Biosystems (2008) Guide to performing relative quantitation of gene expression using real-time quantitative PCR. http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_042380.pdf. Accessed 2 Jun 2008
84. Bauer P, Rolfs A, Regitz-Zagrosek V et al (1997) Use of manganese in RT-PCR eliminates PCR artefacts resulting from DNase I digestion. *Biotechniques* 22:1128–1132
85. Bustin SA (2000) Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol* 25:169–193
86. Rodríguez A (2012) Desarrollo de métodos de PCR en tiempo real para la detección y cuantificación de mohos productores de micotoxinas en alimentos. Doctoral Thesis. University of Extremadura, Spain
87. Sayers EW, Barrett T, Benson DA et al (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 40:D13–D25
88. Cui W, Taub DD, Gardner K (2007) qPrimerDepot: a primer database for quantitative real time PCR. *Nucleic Acids Res* 35:D805–D809

Chapter 4

Large-Scale Nucleotide Sequence Alignment and Sequence Variability Assessment to Identify the Evolutionarily Highly Conserved Regions for Universal Screening PCR Assay Design: An Example of Influenza A Virus

Alexander Nagy, Tomáš Jiřinec, Lenka Černíková,
Helena Jiřincová, and Martina Havlíčková

Abstract

The development of a diagnostic polymerase chain reaction (PCR) or quantitative PCR (qPCR) assay for universal detection of highly variable viral genomes is always a difficult task. The purpose of this chapter is to provide a guideline on how to align, process, and evaluate a huge set of homologous nucleotide sequences in order to reveal the evolutionarily most conserved positions suitable for universal qPCR primer and hybridization probe design. Attention is paid to the quantification and clear graphical visualization of the sequence variability at each position of the alignment. In addition, specific problems related to the processing of the extremely large sequence pool are highlighted. All of these steps are performed using an ordinary desktop computer without the need for extensive mathematical or computational skills.

Key words Alignment, Entropy, Primer, Probe, Inclusivity, Influenza, PCR, Real-time PCR, qPCR

1 Introduction

Currently, conventional or real-time PCR (qPCR) is the leading technique for virus detection in diagnostic microbiology. PCR and qPCR assays were developed to detect a plethora of various viruses and are implemented in many clinical and diagnostic laboratories as the first-line tool for the screening and discrimination between positive and negative specimens. Conventionally, such assays were designed as universal to enable virus detection at defined taxonomic levels, e.g., Influenza A virus genus [1], *Paramyxoviridae* family [2], *Coronavirinae* subfamily [3], and all serotypes of Bluetongue virus [4] or Foot-and-mouth disease virus [5], etc.

An essential feature of such a screening PCR assays is to provide quick, sensitive, and specific virus detection, with the ability to include as many genetic variants of a given virus taxon as possible.

This ability is, however, significantly limited by a continuous accumulation of mutations during the virus evolution. Therefore, careful oligonucleotide selection targeting the most conserved regions of the viral genome is of utmost importance in the development of a diagnostic PCR assay to prevent or minimize false negativity due to primer or probe binding failure. Accordingly, the first step in designing a universal PCR assay is to look for the evolutionarily highly conserved regions in the viral genome. However, this approach is frequently underestimated in literature. In addition, the inclusivity (*see Note 1*) of the screening PCR assays has not been addressed. Therefore, nobody can really tell how many virus strains are detected by a given diagnostic PCR assay and how many remain undetected.

Intensive development and wide accessibility of sequencing technology during the last decade has led to a significant accumulation of virus sequence information which is stored in public databases. Besides general sequence databases collecting sequences from all fields of biology, for instance the GenBank [6], specialized virus databases or sub-databases have been established for the most studied viruses, e.g., VSD, IVSD, GISAID, HCV database, HIV database, HPV database [7–12]. Currently, the databases contain tens or hundreds of thousands of sequences of a given virus taxon, gene, or genome segment. Such extensive and heterogeneous sequence data, representing a broad collection time period, various host species, and geographic areas, provides a suitable basis for the analysis of sequence variability and identification of the highly conserved positions for a screening PCR assay design. On the other hand, it enables to infer the inclusivity of a preexisting diagnostic assay by challenging the primer and probe sequences against large-scale sequence data.

The purpose of this chapter is to provide a guideline on how to align, process, and evaluate a huge set of homologous nucleotide sequences. Attention was paid to the quantification and clear graphical visualization of the sequence variability at each position of the alignment. All of these steps were performed using an ordinary desktop computer without the need for extensive mathematical or computational skills. In addition, specific problems related to the processing of the extremely large nucleic acid sequence pool were highlighted. The methodology is demonstrated on the M segment of influenza A (IA) virus as an example and follows the principle presented in our previous work [13], however, in a more detailed manner and on the basis of a larger dataset. In general, the methodology is applicable to the identification of sequence variability of any kind of homologous nucleic acid sequences of interest.

The procedure contains the following steps:

1. Sequence downloading.
2. Multiple sequence alignment.

3. Alignment validation.
4. Quantification and visualization of sequence variability.

Since the rapid rate of sequence accumulation does not go hand in hand with the development of bioinformatic analysis tools, the presented methodology is far from being comprehensive. Rather, it integrates particular functions of different freely accessible intuitive software tools and Web-based applications.

2 Materials

Conventionally, the IA virus screening PCR methods were targeted on the M segment which is generally considered to be highly conserved across various IA virus subtypes and host species. The M segment is a bicistronic negative sense RNA molecule of 1,027 nucleotides in length and encodes two membrane proteins, M1 and M2. The long-term evolution of the entire segment is characterized by a continuous accumulation of point mutations, without insertion or deletion mutagenesis [14, 15].

2.1 Software Tools and Web-Based Applications

Sequence database. The IA virus M sequences were downloaded from the GISAID's EpiFlu™, a non-annotated database, which contains the world's most complete collection of IA virus sequences (see Note 2). The database is operated by the Max Planck Institute in Germany: <http://platform.gisaid.org/epi3/frontend#23443a> and is accessible through registration.

Multiple sequence alignment. For multiple sequence alignment, a server-based application of the MAFFT program (Multiple Alignment using Fast Fourier Transform; [16]), provided by the Computational Biology and Research Centre, Japan, was implemented. The application is freely accessible through: <http://mafft.cbrc.jp/alignment/server/>.

Alignment validation. Alignment visualization, sequence editing, and alignment trimming were performed by using the BioEdit [17] a biological sequence alignment editor, freely accessible through: <http://www.mbio.ncsu.edu/bioedit/bioedit.html>.

Text Editors. Simple plain text file editing was performed in Notepad developed by the Microsoft Corporation.

Entropy calculation. The information entropy was calculated using the EntropyCalculator designed in our laboratory and available as open source, along with the tutorial, on the following Web sites: <http://www.szu.cz/entropycalculator>.

Graph drawing. The information entropy plot and coverage plot drawing was performed by using the Excel program, a spreadsheet application developed by the Microsoft Corporation.

3 Methods

3.1 Sequence Download

To obtain as much sequence data as possible, the entire M sequence information content of the EpiFlu™ database (*see Note 2*) was downloaded in a FASTA format (*see Note 3*, Fig. 1). The sequences were saved regardless of IA virus subtype, collection date, sampling locality, or M sequence length in four separate files: avian, human, swine, and equine (Table 1).

1. Access the EpiFlu™ database through <http://platform.gisaid.org/epi3/frontend#23443a>.
2. Select all search options in the filter set (influenza type A, HA, and NA subtypes, host, geographic locations, collection date, and segment) (*see Note 4*).
3. Select the download format and customize the FASTA header (*see Note 5*).
4. Save the files separately for avian, human, swine, and equine reservoirs as FASTA to the location of your choice (*see Note 6*).
5. Open the FASTA files in a text editor. Select a full-length sequence in the correct direction from the list and put it as the first sequence for reference (*see Note 7*).
6. Save the data. The data structure of the presented procedure was summarised in Table 1.

3.2 Multiple Sequence Alignment

In bioinformatics, a multiple sequence alignment compares two or more protein or nucleic acid sequences in order to identify regions of sequence similarity. Within the alignment, the sequences are displayed in rows where the homologous positions or characters are aligned in successive columns.

For a group of short or closely related sequences, the alignments are easy to generate, even by eye. However, increase in the number and diversity of the input data makes the alignment extraordinarily difficult to perform. Several algorithms were constructed to generate high-quality sequence alignments.

```
>EPI_ISL_15686 | A/Cygnus olor/Czech Republic/5170/06 | A / H5N1 | 2006-03-20 | MP
gtcgaaacgtacgtttctctatcccgctcaggccccctaaagccgagatcgccgacaaaacttgaggatgtcttcgc
aggaaaagaacaccgatctcgaggctctatggagtggctaaagacaagccaatctgtcacctctgactaaaggatgt
tgggattgttgcattgcaccgtcaccgtggcccaagtggcggaggactgcacgttagcggatgccttgcataatggc
aatggagatccaaataatatatggatggcactgttgcataatggatggatgtttgcataatggatggatgtttgc
ggaggtcgactcactcaacccgtgcactcgccagttgcattgttgcataatggatggatgtttgcataatggatgg
cgaaagtggctttggcctagttgttgcacttgttgcattgttgcataatggatggatgtttgcataatggatgg
actatcaccaacccactaatcaggcatgagaacagaatggtgcggccagcactacagactaaggctatggagcagatggc
gggatcaagtggcaggcaggccatggagggtgcctaattcaggcttaggcaggatgtgcaggcaatgagaacaaattg
ggactcatcctaactcttagtgcgttgcataatcttgcaggcttaccagaaacgaatggagatgtttgcattgttgc
cagatgcaggcattcaagtgtatccctttgttgcggcataatggatcttgcacttgttgcattgttgcaggatgttt
atgtttttttcaatgcattatgcgtcgcttaataatgggtttgaaaagaggccccttacggaaaggagtacctgag
tctatgggaaagagatcggcaggaaacagcagaatgt
```

Fig. 1 FASTA file. A representative sequence in FASTA format opened in a text editor

Table 1
Summary of the downloaded sequences

Host	Number of sequences	Date	FASTA file size (Mb)
Avian	10,378	2nd April 2013	11.1
Human	21,651	8th April 2013	32.3
Swine	3,260	6th May 2013	3.5
Equine	159	6th May 2013	0.17
Sum	35,448		

However, for a data set that consists of thousands of sequences, they are not applicable due to dramatic increase in the CPU (Central Processing Unit) time of desktop computers.

An alternative method for multiple sequence alignment called MAFFT has been developed by Katoh and colleagues [16]. The method implements fast Fourier transform which drastically reduces the CPU time. Currently the MAFFT is provided as a Web-based service or as a download version. For our purpose, the Web-based application is suitable, enabling for example to align a set of 10,000 MP sequences in a matter of 20 min, depending on the Internet connection speed and database occupancy however.

1. Launch the MAFFT server through <http://mafft.cbrc.jp/alignment/server/>.
2. Upload the first FASTA file from the selected location (e.g., the avian).
3. From the “Direction of nucleotide sequences” option, select the “Adjust direction according to the first sequence” button (*see Note 7*).
4. Adjust the alignment strategy to “Auto” or “FFT-NS-1” and submit.
5. After the alignment has completed, press “FASTA format” to view the results as text.
6. Copy the aligned sequences from the server, paste to a text editor file and save the data.
7. Align the remaining three sequence pools.

3.3 Alignment Validation

The first attempt to align a large-scale sequence data pool usually does not result in a high quality output directly usable for downward analyses. A sequence database, even an annotated one, is not in ideal state, i.e., it does not provide only full-length, high-quality sequences all in the right direction. Rather, as in the case of the IA

virus M segment, the data consists predominantly of fragments of variable lengths. In addition, low-quality sequences or misinterpreted data can be included in the database. Generally, the following categories of low-quality items can be recognized which are listed in decreasing significance (*see Note 8*):

- Single nucleotide insertions (Fig. 2a).
- Sequences in incorrect directions (Fig. 2b).
- Sequences longer than the standard length of the M segment, i.e., 1,027 nucleotides.
- Low-quality data with longer stretches of Ns or abundant Y, R, or other characters indicating uncertainty (Fig. 2c).
- Extremely short sequences with less than 100 nucleotides in length.

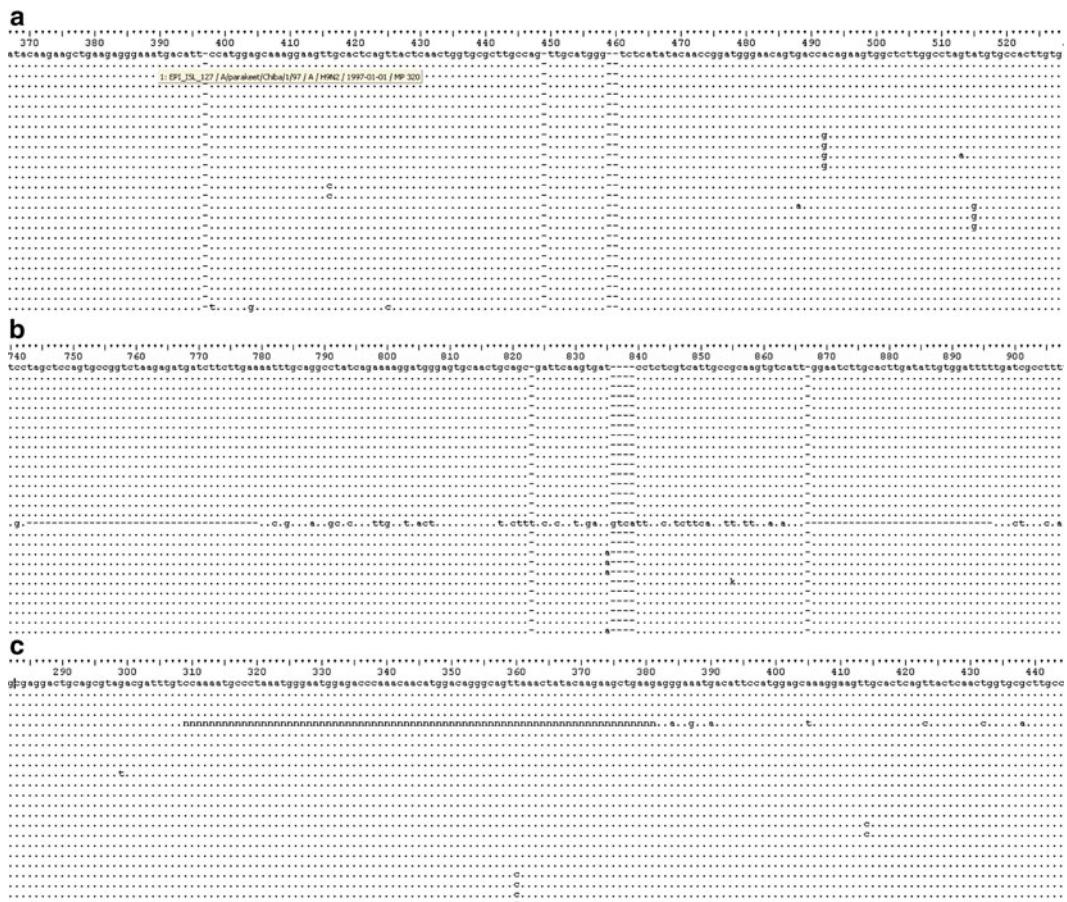


Fig. 2 Low-quality sequences and their effects on a sequence alignment. (a) Short nucleotide insertions; (b) a sequence in incorrect direction; (c) low-quality data with a long stretch of Ns. The alignment screenshots were made by BioEdit. The “empty” columns are marked with “–” (minus)

To eliminate these adverse effects, alignment validation is required which generally includes the following steps:

- Restoration of alignment continuity.
- Alignment trimming.
- Removal of the low-quality data.

3.3.1 Restoration of Alignment Continuity

As shown in Fig. 2a, b, sequences with incorrect direction or short artificial insertions introduce “empty” columns of variable length into the alignment body which significantly interrupt its continuity. Especially the sequences in incorrect direction are the main contributors to this adverse effect. Therefore, first of all we must focus on restoring the continuity of the alignment.

From the first view, this approach requires sorting the alignment to find and cut out the incorrect sequences and then to realign the data. However, this step is quite time-consuming. Since one sequence in incorrect direction is enough to seriously interrupt the whole alignment, searching it out from the pool of thousands of sequences is like looking for a needle in a haystack. Fortunately, by selecting the “Direction of nucleotide sequences” option in the MAFFT, the problem with incorrect direction can be easily overcome (*see Note 7*).

On the other hand, sequences with single artificial insertions accounting for “narrow empty” columns still remain in the data set. These will be removed by the manual editing of the sequence alignment.

Several freely accessible software products to edit and manipulate the nucleic or amino acid sequences and their alignments have been developed. In our protocol, we use the BioEdit [17].

3.3.2 Alignment Trimming

As was mentioned above, the sequence pool analyzed may also contain longer sequences than the standard length of a given gene or genome segment. Such longer sequences contain redundant artificial stretches at one or both termini and were not removed during the database submission by the authors. Despite occurring sporadically, they unnecessarily extend the length of the alignment and bias the exact position numbering. For that reason, the alignment length has to be standardized by trimming both termini. The trimming was performed again with the BioEdit program. The length of the IAV M segment as a currently used example was standardized according to the conserved sequences at the 5' and 3' noncoding regions [18].

3.3.3 Removal of Low-Quality Data

The third class of sequences which reduce the alignment quality are those containing longer stretches of uncertain positions (Fig. 2c). Although the degenerations will not appear in the results of the calculation of positional nucleotide numerical summary (*see Note 9*), it is optional to review the data and identify

low-quality sequences. Identification of longer stretches of low-quality data is possible using either the BioEdit or text editors, and searching out for example the “NN,” “YY,” “RR,” or similar motives.

1. Download and install the BioEdit software through the following link: <http://www.mbio.ncsu.edu/bioedit/bioedit.html>.
2. Open the first aligned file.
3. Perform visual inspection of the entire alignment either in horizontal and vertical directions (*see Note 10*).
4. Select the entire empty columns by clicking on the column number.
5. Cut or delete the empty columns to restore the alignment continuity.
6. To standardize the alignment length, drag the mouse along the redundant terminal sequence positions.
7. Trim the alignment to standardize the sequence length.
8. Save the data as FASTA.
9. Open the data in a text editor.
10. Search for example the “NN,” “YY,” “RR” or similar motives to identify the low quality data.
11. Remove the sequences of low quality if necessary.
12. Validate the remaining three alignments.

3.4 Quantification of Sequence Variability

3.4.1 Position Coverage

In the previous step, the sequence alignment was validated to remove low-quality data. However, the alignments are still far from ideal status because of the heterogeneity in sequence lengths. Beside the full-length sequences, the data includes also sequence fragments of variable lengths the total number of which may significantly exceed the full-length ones. Hence discarding all of the fragments might significantly bias the identification of the conserved regions.

The variation in sequence lengths results in an unequal distribution of data throughout the alignment and in differences in true data content between the columns. To gain insight into the alignment heterogeneity, the information content of the columns was evaluated by determining the position coverage in percentages. The result is visualized graphically as a coverage plot (Fig. 5).

3.4.2 Entropy Calculation

The sequence alignment provides an insight into the variability of the sequence of interest. For a few entries, the variability of the data is clearly visible by visual inspection of the alignment. However, in our case, the alignment contains thousands of lines and hundreds of columns. In such situations, a more sophisticated approach is required for clear visualization of the sequence variability, preferably with a graphical output.

One suitable way of expressing the variability through a column in an alignment comes from the information theory. According to this theory, the amount of variability is quantified as the entropy [2, 19–25].

The information entropy is a measure of uncertainty. Applying this approach to a large-scale sequence alignment means quantifying of the amount of variability in each position according to the formula:

$$H(i) = -\sum f(x,i) \log_2 f(x,i) \quad (1)$$

$$H(i) = -\sum f(x,i) \ln f(x,i) \quad (2)$$

where $H(i)$ is the entropy at position i , $f(x, i)$ is the frequency of each base x at position i in a multiple sequence alignment [26]. Accordingly, the most variable positions have the highest entropy values (the lowest information content and therefore the highest uncertainty). Conversely, the conserved positions are those with the lowest entropy values (the highest information content) converging or equal to zero. The output of the entropy calculation can be visualized in the form of a column diagram, entropy plot, where the entropy values are plotted against the respective positions of the alignment.

In the presented procedure, the entropy was calculated from the exact count of informative positions (i.e., only A, T, C, and G nucleotides) in each column of the alignment. First of all, the positional nucleotide numerical summary was determined. This summary was then converted to entropy values by using the EntropyCalculator software developed in our laboratory and available as open source (*see Note 11*).

The position coverage and information entropy were calculated separately for each sequence pool and visualized graphically as position coverage and entropy plots, respectively. Both of the plots can be drawn from the EntropyCalculator output (Fig. 4). To reveal the overall nucleotide variation of the M segment, regardless of the sequence origin, the partial entropy plots were assembled and visualized as the main entropy plot.

1. Open the validated alignment in the BioEdit software.
2. Select all sequences in the left column.
3. Perform the “Positional nucleotide numerical summary calculation” (*see Note 9*, Fig. 3) and save the data as a text file.
4. Launch the EntropyCalculator program through the following link: <http://www.szu.cz/entropycalculator>.
5. Perform the entropy calculations separately for each data pool.
6. Open the EntropyCalculator output in the Excel (Fig. 4).
7. Draw a position coverage graph separately for each sequence pool by plotting the %Coverage column values against the respective positions (Fig. 4, columns A and V).

Summary of numbers of nucleotides at each position
 Alignment: M:\Alex\Publikace\Projekty\Kniha\Data\Avian\Avian.aln valid.txt

Position	A	G	C	U	GAP
1 1192	1	2	1	9182	
2 3	1199	2	0	9174	
3 3	5	1196	0	9174	
4 1039	174	1	2	9162	
5 1225	0	1	10	9142	
6 1250	2	1	1	9124	
7 1253	5	1	37	9082	
8 6	1332	8	0	9032	
9 7	11	1339	0	9021	
10	1339	7	17	0	9015
11	9	1369	2	2	8996
12	3	1412	0	9	8954
13	3	0	0	4363	6012
14	4380	4	0	3	5991
15	5	4392	4	2	5975
16	4704	8	1	3	5662
17	17	7	5	4893	5456
18	4082	1098	2	24	5170
19	30	16	3	5278	5051
20	7	3	3	5391	4974
21	11	5203	0	247	4917
22	5466	23	0	3	4886
23	5609	24	1	1	4743
24	5732	2	0	1	4643
25	68	5754	1	0	4555
26	9098	0	0	1	1279
27	1	1	1	9107	1268
28	0	9124	0	1	1253
29	9177	1	0	0	1200
30	2	9198	2	2	1174
31	1	2	446	8806	1123
32	1	1	9257	15	1100
33	1	1	2	9286	1088
34	0	0	5	9290	1083
35	0	0	9260	37	1079
36	0	2	4	9295	1077
37	8740	519	0	36	1081
38	9301	0	18	4	1055
39	2	0	9326	6	1044
40	10	22	9286	81	979

Fig. 3 Positional nucleotide numerical summary calculation output of the BioEdit software was shown for the first forty positions of the avian M sequence file as an example. The abundance of each of nucleotides A, G, C, U(T) in each position was given as the absolute value (the N, R, Y, or other uncertainties are not taken into account). The “GAP” means missing data

J	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
t	Position	A	G	C	U	GAP	SUM	Total	Afreq	Gfreq	Cfreq	Ufreq	InAfreq	InGfreq	InCfreq	InUfreq	H(A)	H(G)	H(C)	H(U)	H(Q)	%Coverage
2																						
3	1	1192	1	2	1	9182	1196	10378	0.996656	0.00836	0.001672	0.000836	-0.00335	-7.08674	-6.39359	-7.08674	0.003339	0.005925	0.010692	0.005925	0.025881	11,52
4	2	3	1199	2	0	9174	1204	10378	0.002492	0.995847	0.001661	0	-5.99479	-0.00416	-6.40026	0	0.014937	0.004144	0.010632	0	0.029713	11,60
5	3	3	5	1196	0	9174	1204	10378	0.002492	0.004153	0.993355	0	-5.99479	-5.48397	-0.00667	0	0.014937	0.022774	0.006622	0	0.044334	11,60
6	4	1039	174	1	2	9162	1216	10378	0.854441	0.143093	0.000822	0.001645	-0.15731	-1.94427	-7.10332	-6.41017	0.13441	0.278209	0.005842	0.010543	0.429004	11,72
7	5	1225	0	1	10	9142	1236	10378	0.9911	0	0.000809	-0.00891	0	-7.11964	-4.81705	0.00808	0	0.00576	0.038973	0.053593	0.053593	11,91
8	6	1250	2	1	1	9124	1254	10378	0.998811	0.001465	0.000797	-0.00319	-6.45585	-7.13409	-7.13409	0.003185	0.010273	0.005669	0.005669	0.005835	12,08	
9	7	1253	5	1	37	9062	1278	10378	0.998368	0.000772	0.026549	-4.10217	-5.55716	-7.13409	-3.55612	0.003185	0.010273	0.005669	0.005669	0.011119	12,40	
10	8	6	1332	8	0	9032	1346	10378	0.004458	0.989599	0.005944	0	-5.41313	-0.1046	-6.25645	0	0.02413	0.010347	0.030463	0	0.06421	12,97
11	9	7	11	1339	0	9021	1357	10378	0.005158	0.008106	0.986735	0	-5.26712	-4.81514	-0.01335	0	0.02171	0.039032	0.013176	0	0.079378	13,08
12	10	1339	7	17	0	9015	1363	10378	0.982392	0.005133	0.012472	0	-0.01777	-5.27153	-4.38423	0	0.017452	0.027073	0.054682	0	0.099208	13,13

Fig. 4 EntropyCalculator output. Columns A–F represents “Positional nucleotide numerical summary calculation” (Fig. 3). The “SUM” column gives the number of true data (A, G, C, and U/T) and the “Total” column summarizes the true and missing data (F+G) per position, respectively. Columns I–T show the calculation procedure and the entropy “H(i)” and “%Coverage” values as the final results were visualized in columns U and V. The output is shown in an Excel spreadsheet

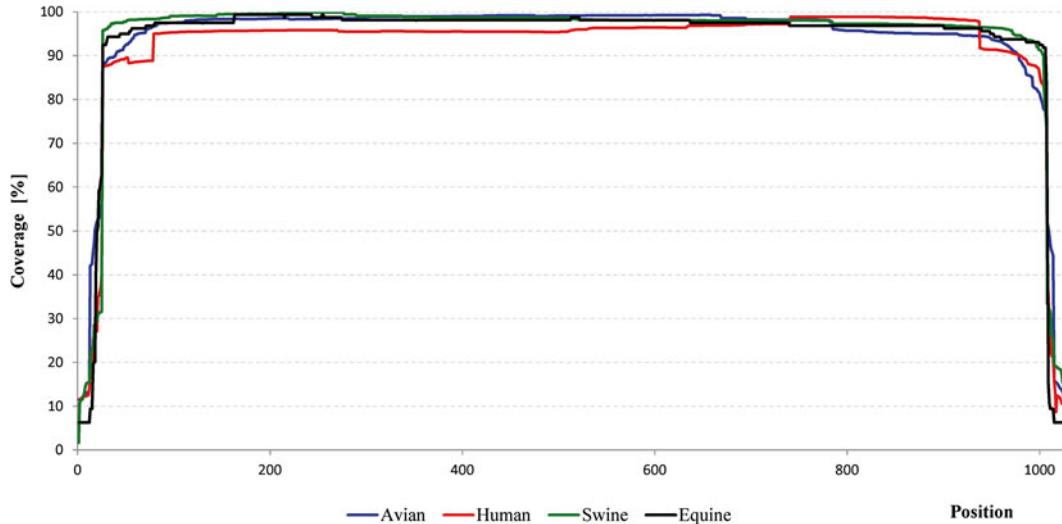


Fig. 5 The coverage plot. The %Coverage values for each position of the M segment (column V in Fig. 4) were calculated separately for four sequence groups: avian (blue), human (red), swine (green), and equine (black). Then, the obtained values were joined into a new Excel sheet and visualized by plotting the %Coverage values against the position number

8. Draw an entropy graph separately for each sequence pool by plotting the $H(i)$ values against the respective positions (Fig. 4, columns A and U).
9. Copy the %Coverage columns from all four data sets to a new Excel sheet.
10. Draw the main coverage plot (Fig. 5).
11. Copy the $H(i)$ columns from all four data sets to a new Excel sheet.
12. Draw the main entropy plot (Fig. 6).
13. Evaluate the sequence variability and identify the presence and distribution of the conserved sequence stretches (*see Note 12*).

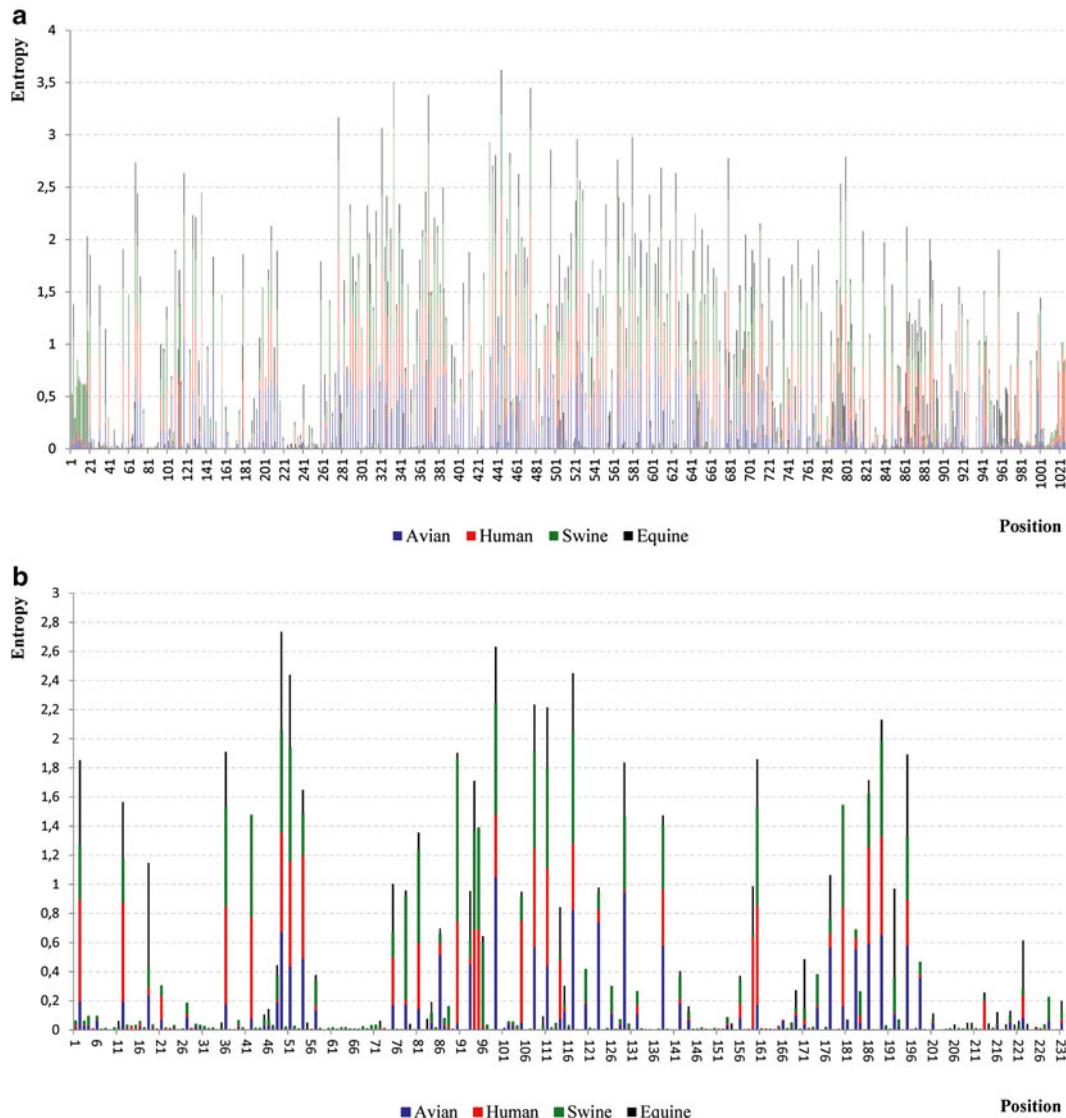


Fig. 6 The entropy plot. The entropy values for each position of the M segment (column U in Fig. 4) were calculated separately for four sequence groups: avian (blue), human (red), swine (green), and equine (black) Table 1. The obtained values were joined into a new Excel sheet and visualized as an assembled column diagram by plotting the entropy values against the M segment positions. The partial column heights are proportional to the entropy values observed within a given sequence group. The overall variation per position is expressed by the total column heights. The entropy plot for the entire M segment is shown in (a) and the conserved stretches within the 5' terminus are visualized in (b)

4 Results

Integration of the entropy values calculated separately for each M sequence group into the main entropy plot revealed considerable variation in the IA virus M segment at the nucleotide level (Fig. 6a).

The main entropy plot further indicated that the positions with elevated nucleotide variation were scattered throughout the entire segment, with the main stretch of variable sites located within the central part of the molecule. This variable domain was interrupted by short, up to eight-nucleotide-long conserved sequences. Longer conserved stretches with the $H(x)$ values very close or equal to zero were only found within the 5' terminus of the segment (Fig. 6a). The conserved motives in the 5' termini could serve as starting points for primer and probe selection and subsequent in silico and experimental evaluation of their potential use for universal PCR or qPCR assay design [13].

5 Notes

1. As a measure of the universality of a diagnostic assay, the inclusivity parameter was introduced. We defined the inclusivity as a quantitative expression of the proportion of the sequences within the nucleotide sequence collection known to date that can theoretically be detected by the assay. The inclusivity parameter has the theoretical and experimental aspects. The theoretical inclusivity expresses the percentages of sequences showing 100 % identity within the primer and probe binding regions. For further details, see ref. [13].
2. It is also possible to use an alternative data set and proceed from Chap. 3.2. However, if you wish to analyze the IA virus M segment but do not intend to register to the GISAID database, please refer to the Influenza Virus Database [8] through the following link: <http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database>. As part of the GenBank, the IVR database is an annotated collection of publicly available IA virus sequences. However, the IVR information content is smaller compared to the EpiFlu™ database.
3. Although different databases vary in their infrastructure, all should provide a common option for sequence download, i.e., “save, or export as FASTA.” FASTA [27] is one of the most often used formats in bioinformatics. It is a text-based format either for nucleotide or amino acid sequences represented by single-letter IUBMB (International Union of Biochemistry and Molecular Biology) codes. The FASTA format is easy to manipulate (copy, paste, edit) using text editors. A typical FASTA format (Fig. 1) begins with a greater-than symbol “>” followed by a sequence description line or header line which contains sequence-specific descriptors (accession codes, sequence name, origin, collection data, etc.). The header line content is usually customizable. Below the header the nucleotide or amino acid sequence is listed in columns of 70–80 characters per line.

4. It is possible to download only the full-length sequences (select the “full length only” option). However, to analyze an as broad as possible range of sequences and illustrate all the difficulties, the entire database content was analyzed.
5. We used the following header format: Isolate ID|Isolate name|Type|Collection date|Segment (Fig. 1).
6. If the selected data exceeds the database download limit, download the data *per partes* and assemble (copy and paste) the partial FASTA files in a text editor software into a single “ready for processing” file.
7. In a large dataset, some input sequences may be in the incorrect direction which results in alignment disruption (Chapter 3.3). The MAFFT algorithm (Chapter 3.2) can automatically adjust the sequence direction relative to the first sequence in the data set used as a reference. The direction adjustment is one of the latest improvements to the MAFFT [29] and significantly facilitates the alignment validation process.
8. Beside these obvious variations in sequence quality the sequence content of the IVSD is biased or affected by much complicated underlying anomalies, e.g., those listed in ref. [28]. The presence of such anomalies can be apparently implicated for other sequence databases. Nevertheless, they have negligible impact on the identification of the conserved positions. The second category of bias is represented by laboratory variants, i.e., sequences resulted from in vitro mutagenesis, recombination experiments, multiple passaged strains, cell line-adapted variants, etc.
9. The BioEdit’s “Positional nucleotide numerical summary calculation output” function lists the true number of nucleotide occurrences (only A, T, C, G, and Gaps) at each position of the alignment. In our case, the gaps mean missing data. N, R, Y, or other uncertainties are not taken into account. A typical positional nucleotide numerical summary calculation output is shown in Fig. 3.
10. Recently, a novel alignment viewer and editor software for large data sets, called AliView, was developed [30]. This intuitive software allows more swift alignment handling than the BioEdit. Therefore, it is more suitable for visual inspection of large alignments. The software is freely available through the following link: <http://ormbunkar.se/aliview>.
11. The entropy calculation is one of the inherent functions of the BioEdit software. However, based on our experience, the BioEdit only calculates the entropy for an ideal alignment. In the case of fragments, the absent data in columns were considered as true gaps. Hence, this approach does not discriminate between the true gaps and data absence.

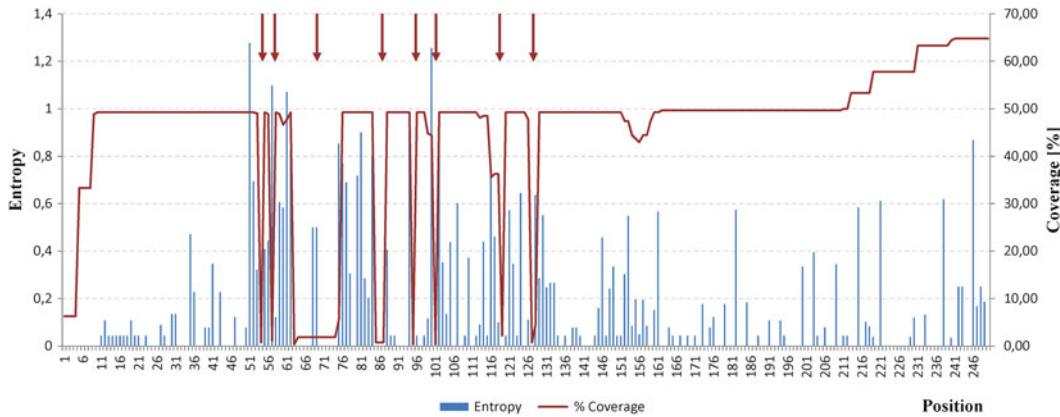


Fig. 7 The integrated entropy and % coverage plots. The entropy was drawn as column diagram related to the main y axis while the % coverage was shown as line graph related to the secondary y axis. The vertical arrows (brown) indicate insertions or low quality regions

12. To obtain more comprehensive results the entropy and % coverage plots can be drawn into a single graph. An example of such integrated plot is shown in Fig. 7. The % coverage plot also allows identifying the abundance, length and frequency of insertions or low quality regions in the alignments.

Acknowledgements

This work was supported by institutional support of the Ministry of Health of the Czech Republic no.1RVO-SZÚ/2014. We acknowledge the authors and originating and submitting laboratories of the sequences from GISAID's EpiFlu Database and NCBI's Influenza Virus Resource database.

References

1. Spackman E, Senne DA, Myers TJ et al (2002) Development of a real-time reverse transcriptase PCR assay for type A influenza virus and the avian H5 and H7 hemagglutinin subtypes. *J Clin Microbiol* 40:3256–3260
2. Van Boheemen S, Bestebroer TM, Verhagen JH et al (2013) A family-wide RT-PCR assay for detection of paramyxoviruses and application to a large-scale surveillance study. *PLoS One* 7:1–9
3. Escutenaire S, Mohamed N, Isaksson M et al (2007) SYBR green assay real-time reverse transcription-polymerase chain reaction assay for the generic detection of coronaviruses. *Arch Virol* 152:41–58
4. Toussaint JF, Sailleau C, Breard E et al (2007) Bluetongue virus detection by two real-time RT-qPCRs targeting two different genomic segments. *J Virol Methods* 140:115–123
5. Reid SM, Ferris NP, Hutchings GH et al (2002) Detection of all seven serotypes of foot-and-mouth disease virus by real-time, fluorogenic reverse transcription polymerase chain reaction assay. *J Virol Methods* 105:67–80
6. GenBank. <http://www.ncbi.nlm.nih.gov/>
7. VSD, Virus Sequence Database. <http://kcdc.labkm.net/vsd/>
8. IVSD, Bao Y, Bolotov P, Dernovoy D et al (2008) The influenza virus resource at the National Center for Biotechnology Information.

- J Virol 82:596–601 <http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html>
9. GISAID, Global Initiative on Sharing All Influenza Data. <http://platform.gisaid.org/epi3/frontend#23443a>
 10. HCV, Hepatitis C Virus Sequence Database. <http://hcv.lanl.gov/content/index>
 11. HIV, Human Immunodeficiency Virus Sequence Database. <http://hiv.lanl.gov/content/index>
 12. HPV, Human Papillomavirus Sequence Database. <http://ncv.unl.edu/Angelettilab/HPV/Database.html>
 13. Nagy A, Vostinakova V, Pirchanova Z et al (2010) Development and evaluation of one step real-time RT-PCR assay for universal detection of influenza A viruses from avian and mammal species. Arch Virol 155:665–673
 14. Ito T, Gorman OT, Kawaoka Y et al (1991) Evolutionary analysis of the influenza A virus M gene with comparison of the M1 and M2 proteins. J Virol 65:5491–5498
 15. Widjaja L, Krauss SL, Webby RJ et al (2004) Matrix gene of influenza A viruses from wild aquatic birds: ecology and emergence of influenza A viruses. J Virol 78:8771–8779
 16. Katoh K, Misawa K, Kuma K et al (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acid Res 30:3059–3066
 17. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/ NT. Nucl Acids Symp Ser 41:95–98
 18. Hoffmann E, Stech J, Guan Y et al (2001) Universal primer set for the full-length amplification of all influenza A viruses. Arch Virol 146:2275–2289
 19. Fouchier RAM, Bestebroer TM, Herfst S et al (2000) Detection of influenza A viruses from different species by PCR amplification of conserved sequences in the matrix gene. J Clin Microbiol 38:4096–4101
 20. Purohit HJ, Raje DV, Kapley A (2003) Identification of signature and primers specific to genus *Pseudomonas* using mismatched patterns of 16S rDNA sequences. BMC Bioinformatics 4:1–9
 21. Cao Y, Wang L, Xu K et al (2005) Information theory-based algorithm for in silico prediction of PCR with whole genomic sequences as templates. BMC Bioinformatics 5:1–5
 22. Batista MVA, Freitas AC, Balbino VQ (2013) Entropy-based approach for selecting informative regions in the L1 gene of bovine papillomavirus for phylogenetic interference and primer design. Genet Mol Res 12: 400–407
 23. Linhart C, Samir R (2002) The degenerate primer design problem. Bioinformatics 18: S172–S180
 24. Hysom DA, Naraghi-Arani P, Elseikh M et al (2012) Skip the alignment: degenerate, multiplex primer and probe design using k-mer matching instead of alignments. PLoS One 7:1–12
 25. Kruger D, Kapturska D, Fischer C et al (2012) Diversity measures in environmental sequences are highly dependent on alignment quality—data from ITS and new LSU primers targeting Basidiomycetes. PLoS One 7:1–21
 26. Hall TA (1999) BioEdit user's manual. See ref. 17
 27. Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441
 28. Krasnitz M, Levine AJ, Rabadian R (2008) Anomalies in the influenza virus genome database: new biology or laboratory errors? J Virol 82:8947–8950
 29. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780
 30. Larsson A (2014) AliView: a fast and lightweight alignment viewer and editor for large data sets. Bioinformatics 15:3276–32780

Chapter 5

Low-Concentration Initiator Primers Improve the Amplification of Gene Targets with High Sequence Variability

Kenneth E. Pierce and Lawrence J. Wangh

Abstract

The amplification and detection of diverse strains of an infectious virus or bacteria, or variants within a gene family is important for both clinical and basic research but can be difficult using conventional PCR. This report describes and illustrates a novel closed-tube method for amplifying and characterizing heterogeneous target sequences using members of the CTX-M beta-lactamase gene family. Different sub-groups of CTX-M genes exhibit low sequence identity, but accurate and efficient detection of these variants is critical because they all confer resistance to penicillin, cefotaxime, and other antibiotics of the beta-lactam class. The method combines a single pair of “thermodynamic consensus primers” (*tc*Primers) with one or more “initiator primers” (*i*Primers), added at low concentration (5–10 nM). Each *i*Primer improves the initial amplification of one or more variants because it has fewer mismatches to its intended target than the more abundant *tc*Primers. As a result of initial amplification, each heterogeneous sequence is shifted stepwise toward a better match with the *tc*Primers. As soon as the *tc*Primer hybridization takes place, amplification proceeds with high efficiency. The *tc*Primer pairs can be designed for symmetric PCR or for Linear-After-The-Exponential (LATE)-PCR. LATE-PCR offers the advantage of generating single-stranded DNA that can be characterized for different gene variants in the same closed tube, using low-temperature mismatch-tolerant fluorescent probes.

Key words Asymmetric PCR, Consensus primers, CTX-M beta-lactamase, Degenerate primers, Hybridization probes, Initiator primers, LATE-PCR, Molecular diagnostics, Nucleic acid sequence diversity

1 Introduction

Amplifying and distinguishing different gene sequences that vary among bacterial or viral strains can be challenging. Indeed, in some cases the homology between primers and a variant target sequence is so low that amplification either fails entirely or is significantly delayed [1, 2]. One potential solution is the use of degenerate primers, in which primers have different possible bases in one or more positions [3, 4]. Results using this method are inconsistent

due to the decreased concentration of effective primers and associated reduction in melting temperature (T_m) and to increased nonspecific amplification by unutilized primers [4]. An alternative approach is to use consensus primers for amplifying related gene sequences in bacteria or viruses [5–7]. The most frequent nucleotide in the gene family at each position of the variable target sequence is typically chosen for the primer sequence. While this strategy works well for many targets in the group, some targets are likely to amplify inefficiently due to a high number of destabilizing mismatches. Consensus primers used in clinical tests for human papillomaviruses can vary in sensitivity for tested strains by five orders of magnitude, which raises the possibility of false negatives [1]. Pooled consensus primers have provided improvement [2], but this method still risks lower efficiency due to reduced concentration of individual primers and increased nonspecific amplification. The approach described here uses consensus primers in which the nucleotide at each variable position is chosen according to criteria of base-pair stability, rather than nucleotide frequency. In general, C-C pairings are the most destabilizing and therefore decrease primer-target T_m the most. In contrast, G-T and G-G pairings are the least destabilizing mismatches [8].

By avoiding large numbers of highly destabilizing mismatches, *tc*Primers minimize the T_m difference between all of the known variants. They thereby minimize the probability that a highly divergent target will fail to amplify. Nevertheless, some highly mismatched targets sequences do fail to amplify, particularly at low initial copy numbers (see Subheading 3.6). In order to reduce the chances of this happening still further, several *i*Primers are also included in the reaction mixture. These primers are designed to have fewer mismatches to those targets that are most mismatched to the *tc*Primers. Each *i*Primer is added at a very low concentration and therefore only serves to initiate amplification before it is depleted. Several *i*Primers can carry amplification forward step-by-step until the resulting amplicon has a sequence which will hybridize readily to the *tc*Primer (Fig. 1). As soon as the *tc*Primer is extended on the modified target, the resulting complementary strand is fully matched to the *tc*Primer and subsequent amplification proceeds with high efficiency. This approach guarantees that even low numbers of highly mismatched targets are amplified.

The initial melting temperatures and concentrations of *tc*Primer pairs can be chosen for conventional symmetric PCR that generates double-stranded DNA or for LATE-PCR that generates abundant single-stranded DNA. LATE-PCR has the advantage that the single-stranded DNA can be further analyzed in the same closed tube using low-temperature probes over a wide range of temperatures (e.g., from 25 to 60 °C) [9–11]. This allows PCR end-point hybridization of “mismatch-tolerant” probes to any amplified gene variant regardless of nucleotide differences. Multiple probes labeled

Target	Sequence	Primer hybridized	T _m
CTX-M-1 gene	3' ... T G C A C T C T T A G T C G A A T A A G T A G C G G G T G C A A T A G C G A C A T G A ... 5'	<i>i</i> CTX-M-1a	70
<i>i</i> -1a complement	3' T G C A C T C G T T A G T C G A A T A A G T A G C G G G C A C A A T A G C G A C A T G A ... 5'	<i>i</i> CTX-M-1b	70
<i>i</i> -1b complement	3' A C T C G T T A G T C A A A C A A G T A G C G G G C A C A A C A G C G A C A T G A ... 5'	<i>i</i> CTX-M-1c	73
<i>i</i> -1c complement	3' A C T C G T T A G T C A A A C A A G T A G C G C C A C A A C A G C G A C A T G A ... 5'	CTX-M <i>tc</i> Excess	73
<i>tc</i> Excess complement	3' T C A A A C A A G T A C C G G C C A C A A C A G C G A C A T G A ... 5'	CTX-M <i>tc</i> Excess	76

Fig. 1 Changes in the CTX-M-1 sequence during amplification. The CTX-M-1 gene sequence shown at the *top* of the figure has several nucleotides (*shaded*) that are mismatched to the excess *tc*Primer. A first *i*Primer, *i*CTX-M-1a, mismatches only two of those nucleotides (15th and 16th from the *right*) and hybridizes to the gene target with a predicted *T_m* of 70 °C. A complementary target (*i*-1a complement, *second line*) is generated during the subsequent cycle by extension of the paired primer, effectively replacing two of the mismatches to the excess *tc*Primer. Extension of *i*CTX-M-1b on that target and subsequent extension of *i*CTX-M-1c on the *i*-1b complement further modifies the amplicon sequence, replacing all but one of the mismatches to the excess *tc*Primer. Once a target fully complementary to that primer (*bottom line*) is generated, additional amplification proceeds with high efficiency

with the same fluorophore can be hybridized to different regions of a long amplicon, generating a unique fluorescent signature for each target variant [12–14]. Thus, highly divergent gene variants are not only amplified but also distinguished in a single closed-tube reaction.

This novel approach to variable target detection is illustrated below using sequence variants that belong to each of the five different groups of bacterial CTX-M genes. These genes encode an extended spectrum beta-lactamase (ESBL) that provides increased resistance to several penicillin-like antibiotics, including cefotaxime. Over 140 different amino acid sequence variants have been identified (<http://www.lahey.org/Studies/other.asp#table1>). Amino acid identity is below 90 % for genes of the different groups [15]. It is important that the detection assay rapidly identifies the presence of any of these variants in the CTX-M family, so that appropriate medical treatment can be delivered. The scarcity of well-conserved regions within this gene provides an excellent test for *tc*Primers and *i*Primers.

2 Materials

1. DNA targets, purified from tissue or cell culture, or reverse transcribed from RNA. Custom gene synthesis can also be used to test variants that are not readily available (e.g., Integrated DNA Technologies, www.idtdna.com).
2. Sequence alignment program (e.g., JalView, www.jalview.org/ or ClustalW2, www.ebi.ac.uk/Tools/msa/clustalw2/).

3. PCR primer design software (e.g., Visual OMP™, DNA Software, Ann Arbor, MI).
4. Thermal Cycler with fluorescence detection capability (e.g., Bio-Rad CFX96™, Stratagene Mx3005P, or Cepheid SmartCycler®).
5. Optical sample tubes appropriate to the thermal cycler.
6. Racks for placing sample tubes on ice (e.g., ABI MicroAmp® Bases).
7. PCR reagents:
 - (a) Taq polymerase with hot start, either with anti-Taq antibodies, e.g., Platinum Taq, or modified enzyme, e.g., AmpliTaq Gold® (Life Technologies).
 - (b) Buffer containing Tris and potassium chloride (usually supplied with commercial Taq polymerases).
 - (c) Magnesium chloride stock solution at 25 or 50 mM.
 - (d) Custom oligonucleotide primers and probes.
 - (e) Deoxynucleotide triphosphates (dNTPs), PCR grade.
 - (f) Water, molecular biology grade.
 - (g) SYBR Green I (Molecular Probes) (optional).

3 Methods

Accurate estimates of oligonucleotide melting temperature (T_m) are critical at all steps of primer and probe design. Software programs should use an accurate nearest neighbor formula ([16–18], *see Note 1*), provide estimates for partially complementary strands having multiple mismatches, and should be able to identify extendible primer dimers (*see Note 2*). T_m estimates provided here are from Visual OMP™ software, version 7.6 (DNA Software®, Ann Arbor, MI) using 50 mM sodium salt and 3 mM magnesium salt (*see Note 3*).

3.1 Source and Preparation of Gene Targets

DNA or RNA targets can be obtained from tissues or cultured cells or virus plaques using standard purification techniques. mRNA can be reverse transcribed using appropriate priming methods and commercially obtained enzymes. We previously described the use of gene-specific LATE-PCR primers for a “one-step” reverse transcription and amplification of Foot and Mouth Disease Virus (FMDV) sequences [19, 20]. It should be possible to design thermodynamic consensus primers that can be similarly used for reverse transcription of moderately variable gene sequences.

Custom gene synthesis can be a useful tool in cases where the variants desired for testing are not readily available or where infectious agents would require special handling. For the study

described here, custom gene synthesis of nucleotides 101–500 of the coding sequences of CTX-M-1, CTX-M-2, CTX-M-8, CTX-M-9, and CTX-M-25 was done by Integrated DNA Technologies (IDT) (Coralville, Iowa) using their custom pIDTsmart vector with Kanamycin resistance for selection. Plasmids were resuspended at the desired concentrations based on yield estimates provided by IDT.

3.2 Thermodynamic Consensus Primer Design

3.2.1 Identifying Sites Within the Gene for Primer Hybridization

Gene variants are aligned using programs such as JalView (www.jalview.org) or ClustalW2 (www.ebi.ac.uk/Tools/msa/clustalw2/). This enables the rapid identification of the relatively well-conserved regions as potential targets for primers. All targets should be complementary to at least two, preferably eight or more nucleotides at the 3' end of the primer, since Taq polymerase needs a double-stranded DNA at least that long for extension [21]. The only mismatch that should be allowed in that region is a single G-T or G-G pairing with any target. The ten nucleotides at the 3' end of the CTX-M excess *tc*Primer are fully complementary to most gene variants, although a single G-T mismatch to the CTX-M-2 gene is present in that region (Fig. 2a). Similarly, all but 1 of the 11 nucleotides at the 3' end of the CTX-M limiting *tc*Primer are complementary to all gene targets. That G-T mismatch is 3 nucleotides from the 3' end of the primer most gene targets (Fig. 2b). Previous results have shown that a single G-T mismatch at this position

a Target	Excess <i>tc</i> Primer and gene target sequences	Predicted T _m
	5' - A G T T T G T T C A T G G C G G T G T T G T C G C T G T A C T - 3'	
CTX-M-1	3' ... • G C • • T • T • • • C [G] • • T • • • G • • G T G C • • T • • • C • • • A • ... 5'	47.7
CTX-M-2	• C T • • C • G • • • C [G] • • T • • • C • • T C A T • • C • • • T • • • G •	60.1
CTX-M-8	• C T • • C • G • • • C [G] • • C • • C • • G C A T • • C • • • C • • • G •	65.8
CTX-M-9	• C G • • C • G • • • [T] A • • C • • • C • • C C A T • • C • • • C • • • G •	71.7
CTX-M-25	• C T • • C • G • • • C [G] • • T • • • C • • T C A T • • T • • • C • • • A •	60.1

b Target	Limiting <i>tc</i> Primer and gene target sequences	Predicted T _m
	5' - T T G C T G A T G A G G C G C T T T G C G A T G T G C A G T A C C A G T A A G G T - 3'	
CTX-M-1	3' ... • • A • [C] A • G A • • A • T C • • G • • C • C • • • G • • • • T • • • ... 5'	75.4
CTX-M-2	• • G • [C] A • G G • • A • T [T] • • A • • C • C • • • A • • • • C • • •	70.6
CTX-M-8	• • G • [C] G • G G • • A • T C • • G • • C • [G] • • • G • • • • C • • •	73.5
CTX-M-9	• • A • [C] G • [C] A • • A • T [T] • • G • • G • T • • • A • • • • T • • •	64.1
CTX-M-25	• • G • [C] G • G G • • G • T C • • A • • C • G • • • G • • • • T • • •	63.2

Fig. 2 CTX-M gene variation at the sites of primer hybridization. (a) Excess *tc*Primer and gene targets. (b) Limiting *tc*Primer and gene targets. Letters are shown for nucleotides that vary among the five genes; dots represent nucleotides shared among those genes. Nucleotide mismatches to the primer that are moderately destabilizing (G-G or G-T) are shaded in light grey. More highly destabilizing mismatches are shaded in medium grey. The predicted T_m of the mismatched hybrids is shown at right

causes very little delay in amplification of a target that is otherwise complementary to the primer (data not shown). As with typical primers, the 3' end of each *tcPrimer* is designed with the percentage G+C between about 40 and 60 % and stable 3' primer dimers are avoided (*see Note 4*). Although neither CTX-M primer forms a 3' homodimer of more than three base-pairs, the five nucleotides at the 3' end of the CTX-M excess *tcPrimer* are capable of forming a heterodimer with the CTX-M limiting *tcPrimer*. This primer pair was deemed acceptable because of the low stability of the dimer and the lack of alternative conserved sites for primers.

3.2.2 Choosing Primer Nucleotides at Other Sites of Variation

Nucleotides throughout the rest of the primer are chosen, first, to minimize the destabilization of individual mismatches and, second, to minimize the number or location of mismatches to any given gene target that would substantially lower the hybrid T_m . To accomplish the second goal, some individual mismatches with greater destabilization can be tolerated. For example, a destabilizing C-C mismatch to the CTX-M-9 target was chosen near the 5' end of the limiting *tcPrimer*, because alternative choices would have prevented hybridization of that end of the primer to other gene variants (Fig. 2b). It is also preferable to avoid a *tcPrimer* that is highly mismatched to any individual target, resulting in a substantially lower T_m . Unfortunately, that is not always possible, as was found for the excess *tcPrimer* and the CTX-M-1 target (Fig. 2b). In that case, additional matched nucleotides for that target would have lowered the T_m with other variants by several degrees.

3.2.3 Adjusting the Length of the Primer to Obtain the Desired T_m with the Fully Complementary Target

Once a *tcPrimer* is extended, that product will serve as a template for the paired primer and a fully complementary target will be synthesized. Efficient and specific amplification depends on the primer having the desired T_m with that target, approximately 4° above the annealing temperature for the excess primer and 4–8° above the annealing temperature for the limiting primer. (Having the limiting primer T_m a few degrees above the excess primer T_m improves amplification efficiency [10].) The CTX-M limiting *tcPrimer* also includes 2T nucleotides added at the 5' end that do not hybridize to the initial gene targets. We have found that having 2 or 3 A or T nucleotides at the 5' end of the limiting primer reduces nonspecific amplification (*see Note 5*). The predicted T_m of the CTX-M excess *tcPrimer* and limiting *tcPrimer* to their fully complementary targets are 76 and 79 °C, respectively.

3.3 Initiator Primer Design

As can be seen from Fig. 2, the predicted T_m of the excess *tcPrimer* to an initial target sequence can be as much as 28 °C below its T_m to a perfectly complementary target, and the predicted T_m 's of the

Table 1
Primer sequences

Sequence name	Sequence (5' to 3')	Intended targets
Antisense primers:		
CTX-M-1-specific excess	CAGCTTGTTCATGCCACGTTATCGCTGTACT	CTX-M-1
CTX-M-2-specific excess	CAATCAGCTTATTCATGGCGGTATTGTCGCTGTACT	CTX-M-2
CTX-M-excess <i>tc</i> Primer	AGTTTGTTCATGGCGGTGTTGTCGCTGTACT	All CTX-M
CTX-M-1a <i>i</i> Primer	ACGTGAGCAATCAGCTTATTCATGCCGTGTTATCGCTGTACT	CTX-M-1
CTX-M-1b <i>i</i> Primer	TGAGCAATCAGTTGTTCATGCCGTGTTGTCGCTGTACT	1a- <i>i</i> complement
CTX-M-1c <i>i</i> Primer	TGAGCAATCAGTTGTTCATGCCGTGTTGTCGCTGTACT	1b- <i>i</i> complement
CTX-M-multi gene <i>i</i> Primer	TGGGCAATCAGCTTGTTCATGGCGGTATTGTCGCTGTACT	CTX-M-2, 8, 9, 25
Sense primers:		
CTX-M-limiting <i>tc</i> Primer	TTGCTGATGAGCGCTTGCGATGTGCAGTACCAAGTAAGGT	All CTX-M
CTX-M-9a <i>i</i> Primer	TATCGCGGTGATGAGCGCTTCCAATGTGCAGTACCAAGTAAGGT	CTX-M-9
CTX-M-9b <i>i</i> Primer	TATCGCGCTGATGAGCGCTTCCATGTGCAGTACCAAGTAAGGT	9a- <i>i</i> complement
CTX-M-25a <i>i</i> Primer	TACCGCGCTGATGAGCGTTTGCCATGTGCAGTACCAAGTAAGGT	CTX-M-25
CTX-M-25b <i>i</i> Primer	TACCGTGCTGATGAGCGTTTGCGATGTGCAGTACCAAGTAAGGT	25a- <i>i</i> comp., CTX-M-2

limiting primer to an initial target sequence can be as much as 16 °C below the T_m to its fully complementary target. Although amplification is delayed and variable when primers are so highly mismatched (*see* Subheading 3.6 below), most variants are amplified when 10,000 starting copies are used. Initial amplification becomes more reliable when *i*Primers are added to the reaction for each gene target having a T_m below 70 °C to the excess *tc*Primer, or a T_m below 74 °C to the limiting *tc*Primer. Sequences of all CTX-M *i*Primers are shown in Table 1.

The *i*Primers for enhancement of the excess *tc*Primer were used at a concentration of 10 nM, one hundredth that of the *tc*Primer (*see* Note 6). At that concentration, the *i*Primers needed to be considerably longer than the excess *tc*Primer to obtain the desired T_m . The *i*Primer to each target had 1–3 mismatches and had a T_m no more than 2° below the 72 °C annealing temperature, enabling at least a fraction of the targets to be hybridized and extended during the first cycle. The CTX-M-1 gene sequence was targeted by series of three *i*Primers to step the sequence toward that of the *tc*Primer (Fig. 1). A single *i*Primer (*i*-CTX-M-multi-gene) hybridized each of the other four gene targets with a T_m no lower than 70 °C. The fully complementary targets to each *i*Primer had a T_m no higher than 78 °C. Keeping the T_m similar to that of

the *tc*Primers with their complementary targets reduces the possibility of mispriming (*see Note 7*).

The *i*Primers for enhancement of the limiting *tc*Primer were used at a concentration of 5 nM, one tenth that of the *tc*Primer. A single CTX-M-2 *i*Primer, two CTX-M-9 *i*Primers, and two CTX-M-25 *i*Primers were designed to step those gene sequences toward that of the limiting *tc*Primer (Table 1). The T_m of each of these *i*Primers and the limiting *tc*Primer to its intended target was at least 73 °C.

3.4 Probe Design

While almost any probe design can be used, we prefer to use fluorescently labeled probes that are mismatch tolerant, meaning that a single probe can be hybridized to any gene variant in a temperature-dependent manner [12–14]. This is possible even for highly mismatched targets, since the single-stranded product of LATE-PCR is freely available to hybridize to probes at temperatures well below the PCR annealing temperature (*see Note 8*). In contrast with symmetric PCR, the T_m of the probe with any target must be at least 5° below the T_m of the limiting primer (to its fully complementary target), to insure efficient extension of the limiting primer and to minimize probe hydrolysis during PCR. The probes are dual labeled, having a fluorophore at one end and a quencher at the other. The probe includes two or three nucleotides at each end that can form complementary pairs. Some or all of those nucleotides may be complementary to targets, but nucleotides that do not hybridize with the targets can be added to the probe to provide this “mini stem,” enhancing contact quenching while maintaining mismatch tolerance not possible with typical molecular beacons. Contact quenching keeps fluorescence low when the probe is not hybridized to target, although affinities and background fluorescence vary with different fluorophore–quencher combinations (*see Note 9*).

Probe-target T_m can be measured prior to PCR experiments by melting analysis with synthetic single-stranded DNA targets using the PCR buffer with magnesium. Those targets should include nucleotides beyond the region hybridized by the probe but should not include regions targeted by the primers. Multiple probes can be used to target different regions of a single amplification product. If these hybridize in close proximity to one another on the target, the fluorophores and quenchers on separate probes can interact via contact quenching (*see Note 10*). The interaction between fluorophore and quencher or between juxtaposed quenchers can increase hybrid stability, increasing the T_m several degrees above the predicted values for individual probes.

Target	Probe and gene target sequences	Predicted T _m	Measured T _m
	5' - Cal Or - A A T C C G A T T G C T G A A A A A C A C G T T - BHQ2 - 3'		
CTX-M-1	3' ... • • A • • C • • C • C • • T • • C • • • • G • • ... 5'	53.4	53
CTX-M-2	• • G • • A • • G • • C • C • • C • • T • • • • A • •	47.0	46
CTX-M-8	• • G • • A • • G • • T • A • • T • • T • • • • G • •	53.3	53
CTX-M-9	• • G • • A • • C • • C • G • • T • • T • • • • G • •	62.2	63
CTX-M-25	• • G • • G • • G • • C • C • • T • • T • • • • G • •	49.0	44

Fig. 3 CTX-M gene variation at the site of probe hybridization. Nucleotide sequence and modifications of the probe are shown at *top*. Letters are shown for nucleotides that vary among the five genes; *dots* represent nucleotides shared among those genes. Nucleotide mismatches to the probe that are moderately destabilizing (G-G or G-T) are shaded in *light grey*. More highly destabilizing mismatches are shaded in *medium grey*. The predicted T_m and the measured T_m following PCR are shown at *right*

The CTX-M probe includes 2A nucleotides at the 5' end and 2T nucleotides at the 3' end, three of which hybridize with most gene targets (Fig. 3). This sequence yielded sufficiently low background fluorescence with the Cal Fluor® Orange 560, Black Hole Quencher® 2 dual-labeled probe. The probe is not fully complementary to any of the five gene targets but hybridizes to each over a predicted T_m range of 47–62 °C.

3.5 Amplification Conditions

Samples with *tc*Primers and/or *i*Primers generally can be amplified using standard reagents and concentrations. Taq polymerase, or other polymerases, should include a hot-start method. LATE-PCR *tc*Primers can be used at 50 and 1,000 nM for the limiting and excess primers, respectively. Dual-labeled probes for LATE-PCR are used in the range of 100–500 nM. At the higher concentration, fluorescent signal after 50 cycles is usually proportional to the amount of starting target enabling end-point quantification [11, 19].

Stringent annealing temperatures should be used from the first cycle to minimize nonspecific amplification. Optimal annealing temperature is about 3–5° below the predicted T_m of the excess primer. Using lower annealing temperatures during the first few cycles could potentially increase the number of targets hybridized and extended directly with the *tc*Primers but is unnecessary if *i*Primers are included. If high sensitivity (e.g., consistent detection of fewer than 100 gene copies) is desired, increase the number of *i*Primers for each target, using only one or two mismatches between each primer and the intended targets; keeping the T_m above the annealing temperature in each instance. Annealing times should be short (e.g., 10–30 s). For amplicons longer than about 100

nucleotides, an extension step at the temperature of maximal enzyme activity (i.e., 72 °C for Taq polymerase) should be added rather than increasing the duration of a lower temperature annealing step.

The CTX-M assay used Platinum Taq (Life Technologies) in 1× buffer and 3 mM magnesium, 0.4 mM each dNTP, 50 nM limiting *tc*Primer, and 1,000 nM excess *tc*Primer. When included, *i*Primers analogous to limiting primers and to excess primers were used at 5 and 10 nM, respectively. An initial denaturation step at 95 °C for 2 min was followed by 60 cycles of 95 °C for 10 s and 72 °C for 60 s. The high annealing/extension temperature with high T_m primers has several advantages: (a) LATE-PCR probes can be used over a wider temperature range during melting analysis; (b) Taq polymerase activity is maximum at that temperature; (c) Primers designed for that temperature are longer than if designed for lower annealing temperatures, minimizing the impact of individual mismatches. The 60 s step at 72 °C used for the CTX-M assay was probably unnecessarily long for the 232 nucleotide product but was used since co-amplification with longer products from other genes is planned in a final assay.

Immediately following LATE-PCR, fluorescence is monitored over a wide range of temperatures. Fluorescence can be detected during annealing as temperature is lowered from 72 to 25 °C (or desired lowest temperature), or the temperature can be slowly reduced to the lowest temperature and fluorescence detected during melting as the temperature is raised (see Note 11). For the CTX-M assay, temperature was lowered in 1 °C steps from 72 to 25 °C with fluorescence detection at each step. Each step was 30 s to allow time for hybridization and for multiple fluorescence detection reads using the Stratagene Mx3005P.

3.6 Analysis of Amplification and Post-PCR Probe Hybridization

Although hybridization probes can be used to detect amplification during the generation of the single-stranded product, quantification requires designing a probe to a sequence that is conserved among the gene variants and may not be possible in many cases. Where that is possible, the T_m of the probe-target hybrid must be at least 5° below that of the limiting primer to insure high amplification efficiency and minimize probe hydrolysis.

Real-time detection can also be carried out using SYBR Green to detect the presence of the double-stranded product during the exponential phase of LATE-PCR. Properly designed LATE-PCR primers that are fully complementary with targets provide detection threshold cycle (C_T) values similar to those of symmetric PCR [10]. Mismatched targets to a single primer will yield higher C_T values, the increase at a given target concentration being roughly proportional to the decrease in T_m of the primer-target hybrid. This delay is presumably due to the reduced fraction of gene targets

that will be hybridized and extended during the early cycles of PCR. Detecting amplification using SYBR Green helps to evaluate designs where one or both *tc*Primers are mismatched to targets, as well as the improvement gained by adding one or more initiator primers. Different initiator primer designs can be evaluated for efficiency and sensitivity. Also, melting analysis following amplification can often confirm the presence of the specific product. The presence of nonspecific product, as determined by different melting temperatures, may indicate the need to increase reaction stringency (e.g., by increasing annealing temperature or reducing annealing duration) or to redesign primers.

The CTX-M samples contained SYBR Green to determine the C_T increase caused by mismatches to the *tc*Primers and the relative improvement afforded by the use of *i*Primers. Samples with CTX-M-1 or CTX-M-2 were also tested using an excess primer that was specific for each target (Table 1), although one or two G to T mismatches to those respective targets were present to provide the same 3' nucleotides and final T_m for all excess primers. All samples contained 10,000 copies of the gene target and included the limiting *tc*Primer as the paired primer, so that any difference could be attributed to additional mismatches and lowered T_m with the excess *tc*Primer. Mean C_T values were 12.1 cycles higher for CTX-M-1 and 8.0 cycles higher for CTX-M-2 using the *tc*Primer relative to the respective specific primer (Table 2), consistent with the lower T_m of the excess *tc*Primer to those targets. SYBR Green melting peaks (not shown) and signal from the hybridization probe (*see* below) confirmed the specific product amplification. The results demonstrate that even with the T_m of the excess

Table 2
SYBR Green real-time detection of CTX-M synthetic gene amplification in samples containing *tc*Primers or *tc*Primers plus *i*Primers

Mean C_T values of replicates amplified using:				
Gene target	<i>tc</i> Limiting + specific excess	<i>tc</i> Limiting + <i>tc</i> excess	<i>tc</i> Primers + <i>i</i>Primers	Improvement with <i>i</i>Primers
CTX-M-1	26.6	38.5	34.3	4.2
CTX-M-2	27.2	35.2	30.1	5.1
CTX-M-8	NT	27.3	26.3	1.0
CTX-M-9	NT	36.7	29.7	6.9
CTX-M-25	NT	45.4 ^a	32.2	13.2

NT not tested

^aOnly one of three replicates with gene-specific product

*tc*Primer–CTX-M-1 hybrid well below the annealing temperature, a small number of targets were amplified during the first few cycles in each replicate sample. Thus, unknown targets at this concentration should be detected with the *tc*Primers, even if one has several mismatches resulting in a similar low T_m .

Similarly, samples with 10,000 copies of CTX-M-9 with both *tc*Primers yielded a much higher mean C_T value than would have been expected from specific primers. In that case, the excess *tc*Primer was reasonably well matched, having only 2G-T mismatches and a T_m of 72° with the gene target (Fig. 2), but the limiting *tc*Primer had a higher number of mismatches and a T_m of 64°. The lower mean C_T value compared to the results from CTX-M-1 and CTX-M-2 suggests that a low T_m with a limiting *tc*Primer may have a greater effect on amplification.

Only one of three replicates with 10,000 copies of CTX-M-25 generated specific product and the C_T value was 45.4, much higher than observed with other targets. Both *tc*Primers are mismatched to that gene and T_m are well below the annealing temperature. Thus, amplification requires a first extension event from a fraction of total copies, then a second extension event on a fraction of those first-extended products using the paired *tc*Primer. It is not surprising that specific amplification in the absence of *i*Primers in examples like this is poor.

SYBR Green detection also demonstrated that the addition of *i*Primers can lower C_T values when used in combination with *tc*Primers (Table 2). Even CTX-M-25 samples were consistently detected and the mean C_T value was only a few cycles above what might be expected from sequence-specific primers for that target.

While SYBR Green detection provides a useful tool for evaluating primer design, hybridization probes provide better confirmation of specific amplification. Mismatch-tolerant probes used with LATE-PCR can provide sequence-specific fluorescence profiles and enable the identification of gene variants in detection assays [12–14, 22]. Therefore, the hybridization probe described in Subheading 3.4 was used in all CTX-M samples. Specific product was detected following 60 cycles of PCR by measuring fluorescence as the temperature was lowered from 72 to 30 °C. The fluorescence readings were normalized using the values at 70 °C, a temperature at which there is no detectable hybridization of the probe to any target. This was accomplished by dividing the average fluorescence by the well-specific fluorescence at 70 °C to obtain “well factors.” The values at all temperatures were multiplied by the specific well factor to adjust for variations in background and probe concentration. Average signal from no-template controls was then subtracted to yield the fluorescence curves shown in Fig. 4.

Each of the CTX-M targets generated a different fluorescence curve, consistent with differences in T_m with the mismatch-tolerant probe. Those T_m values are included in the right column of Fig. 3.

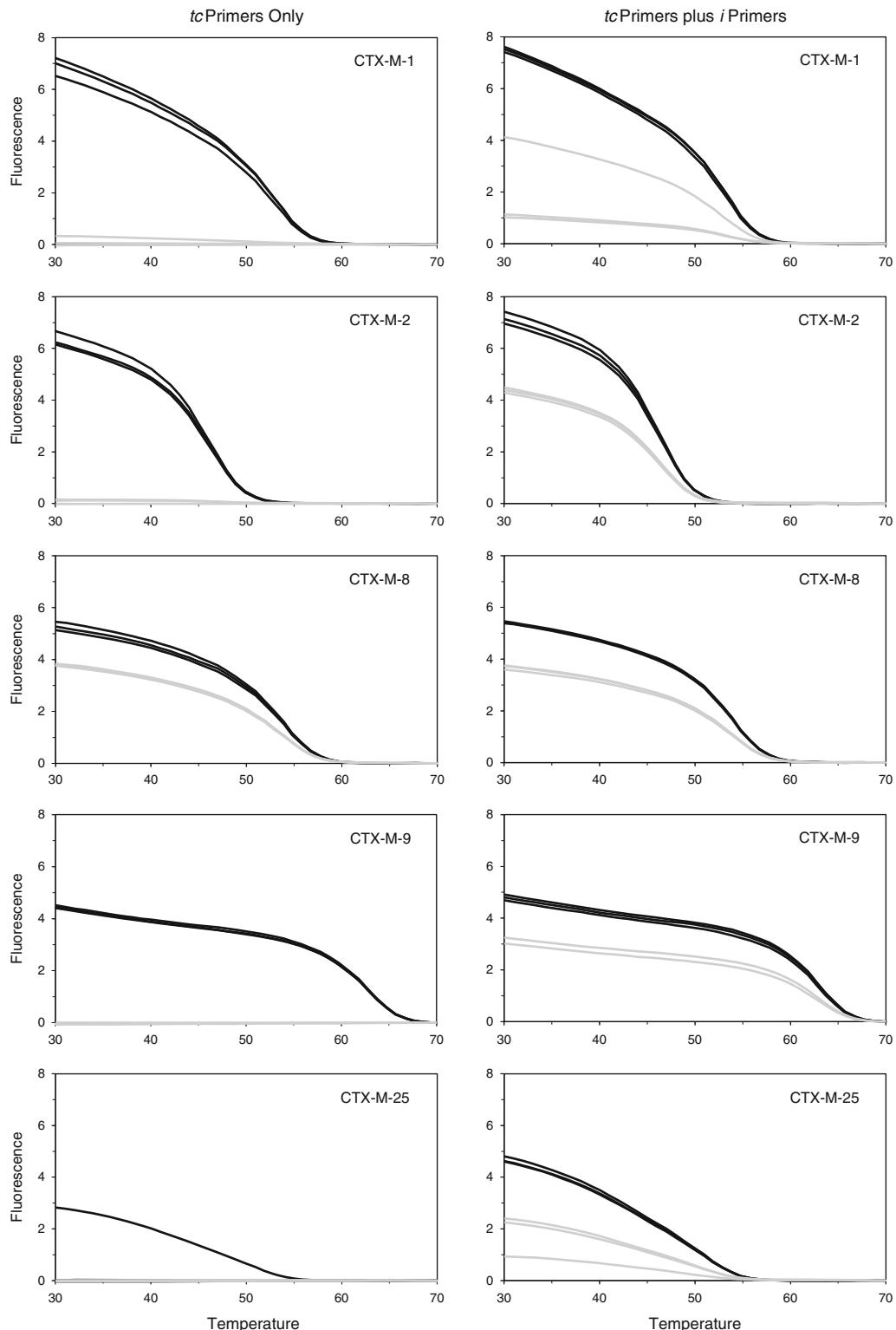


Fig. 4 Probe fluorescence signals obtained following PCR of CTX-M synthetic gene targets using *tc*Primers alone (*left panels*) or in combination with all *i*Primers (*right panels*). Each gene was tested in triplicate samples at 10,000 copies (*black lines*) and at 100 copies (*grey lines*). See Subheading 3.6 for a discussion of the results

In samples containing 10,000 initial copies, four of the five targets generated strong, consistent fluorescent signals (Fig. 4, left panels, black lines). The exception was CTX-M-25, where only one of three replicates generated detectable signal, consistent with real-time detection. In samples containing 100 initial copies of the target, only CTX-M-8 generated consistent fluorescent signal above background (Fig. 4, left panels, grey lines).

The addition of *i*Primers increased the fluorescent signal strength slightly in most samples with 10,000 copies (Fig. 4, right panels, black lines). The CTX-M-25 samples were dramatically improved; strong fluorescence at nearly identical levels was generated in the three replicate samples. Signal was also generated in all *i*Primer samples with 100 copies. That signal was lower than at the higher target concentration, consistent with end-point quantification using LATE-PCR. CTX-M-2, CTX-M-8, and CTX-M-9 targets generated consistent signal levels among replicates, while CTX-M-1 and CTX-M-25 targets generated more variable signals. It might be possible to improve amplification and obtain more reproducible signals at low target concentrations by including additional *i*Primer sequences for those targets, reducing the number of mismatches to one or two at each step toward the *t*₂Primer sequence.

4 Notes

1. Avoid the use of primer design software that uses the nearest neighbor estimates of Breslauer [23], as that can give inaccurate estimates of melting temperatures. Primer3 design software [24] can be used as long as the nearest neighbor and salt corrections of SantaLucia 1998 are chosen.
2. Four or more nucleotides at the 3' end of a primer that can hybridize with nucleotides of that primer (homodimer), the paired primer, or any other primer (heterodimers) in a multiplex amplification should be avoided if possible. However, the main factor in accepting a dimer at the 3' end should be the internal stability of the paired nucleotides. Positive ΔG° values generally do not cause problems; values below -1.0 should be avoided. Primers can be tested using SYBR Green for real-time detection. C_T values before about cycle 40 in the absence of template indicate primer dimer formation that could affect specific amplification.
3. Salt concentrations should be set to those of the PCR buffer. Template concentrations can be set at 1 nM or lower for primer design, so that only the primer concentration affects hybrid T_m , as it is the case at the beginning of PCR. Conversely, template concentration should be set at several hundred nM for probe design, the concentration that will be present at the

end of PCR. We have found that it is best to uncheck the box for “Duplex Polymer Salt Correction” in Visual OMP™, as that adjustment has generated inaccurate estimates.

4. Very high G+C percentages at the primer 3' end are more likely to generate stable primer dimers or other nonspecific amplification. Very low G+C percentages result in relatively low amplification efficiency.
5. The 3' ends of the amplicon strands have nucleotides that are the reverse complements of the primers. These strands can in some cases form partial hybrids and extend on each other to generate progressively longer products, which are observed as a high molecular weight smear or laddering upon gel electrophoresis. This nonspecific amplification is known as product evolution. For asymmetric PCR, including LATE-PCR, the 3' end of the abundant single-stranded product is the reverse complement of the limiting primer. Adding A and/or T nucleotides to the limiting primer 5' end therefore results in a lower G+C percentage at the 3' end of the single-stranded product and decreases the likelihood of product evolution.
6. It should be possible to design *i*Primers at other concentrations. However, it may be difficult to achieve the desired T_m , especially for A+T rich targets using primer concentrations below 1 nM, and may reduce the likelihood of target extension due to a reduced frequency of molecular interactions. High *i*Primer concentrations (e.g., 50 nM) may increase nonspecific amplification.
7. Although it might be possible to design a single *i*Primer for a given target that includes the complete *t*Primer sequence simply by adding sufficient nucleotides to the 5' end to obtain the desired T_m , such a primer might be extremely long and prone to mispriming.
8. LATE-PCR probe sequences must be derived from the same strand as the limiting primer, as both hybridize with the extension product of the excess primer. If there is a need to use the probe sequence from a particular strand, the probe should be designed first and then the limiting primer can be designed from the same strand.
9. Biosearch dyes including Cal Fluor® Orange 570, Cal Fluor® Red 610, and Quasar® 670 have strong contact quenching with Black Hole Quenchers® and a two base-pair stem with A or T residues is usually sufficient for probe design. Fluorophores such as FAM have lower affinity for the Black Hole Quencher and including a G-C pair and/or a third base-pair can help reduce background fluorescence. Useful information about specific fluor-quencher interactions has been published [25].

10. A lights-on/lights-off probing method has recently been described [12, 14]. Hybridizing several pairs of “Off Probes” and “On Probes” to juxtaposed positions on an amplicon generates a fluorescence signature that is unique for each sequence variant.
11. The rate of hybridization may vary for different probes. Probes with low T_m may hybridize more slowly due to the presence of secondary structure in the single-stranded amplicon. In that case, it is generally better to lower the temperature gradually than to drop the temperature rapidly below the T_m of the probe. It may be necessary to test different protocols.

Acknowledgements

This work was supported by Brandeis University and Smiths Detection Diagnostics. The authors thank Dr. Harald Peter and Dr. Till Bachmann for providing alignments of the CTX-M genes.

References

1. de Roda Husman AM, Walboomers JM, van den Brule AJ, Meijer CJ, Snijders PJ (1995) The use of general primers GP5 and GP6 elongated at their 3' ends with adjacent highly conserved sequences improves human papillomavirus detection by PCR. *J Gen Virol* 76:1057–1062
2. Gravitt PE, Peyton CL, Alessi TQ, Wheeler CM, Coutlée F, Hildesheim A, Schiffman MH, Scott DR, Apple RJ (2000) Improved amplification of genital human papillomaviruses. *J Clin Microbiol* 38:357–361
3. Mack DH, Sninsky JJ (1988) A sensitive method for the identification of uncharacterized viruses related to known virus groups: hepadnavirus model system. *Proc Natl Acad Sci U S A* 85:6977–6981
4. Souvenir R, Buhler J, Storto G, Zhang W (2007) An iterative method for selecting degenerate multiplex PCR primers. *Methods Mol Biol* 402:245–268
5. Hubbard RA (2003) Human papillomavirus testing methods. *Arch Pathol Lab Med* 127:940–945
6. Saladin M, Cao VT, Lambert T, Donay J-L, Herrmann J-L, Ould-Hocine Z, Verdet C, Delisle F, Philippon A, Arlet G (2002) Diversity of CTX-M β -lactamases and their promoter regions from *Enterobacteriaceae* isolated in three Parisian hospitals. *FEMS Microbiol Lett* 209:161–168
7. Ayers M, Adachi D, Johnson G, Andonova M, Drebot M, Tellier R (2006) A single tube RT-PCR assay for the detection of mosquito-borne flaviviruses. *J Virol Methods* 135: 235–239
8. SantaLucia J Jr, Hicks D (2004) The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* 33:415–440
9. Sanchez JA, Pierce KE, Rice JE, Wangh LJ (2004) Linear-after-the-exponential (LATE)-PCR: an advanced method of asymmetric PCR and its uses in quantitative real-time analysis. *Proc Natl Acad Sci U S A* 101:1933–1938
10. Pierce KE, Sanchez JA, Rice JE, Wangh LJ (2005) Linear-after-the-exponential (LATE)-PCR: optimizing primer design for high yields of specific single-stranded DNA and improved real-time detection. *Proc Natl Acad Sci U S A* 102:8609–8614
11. Rice JE, Sanchez JA, Pierce KE, Reis AH Jr, Osborne A, Wangh LJ (2007) Monoplex/multiplex linear-after-the-exponential-PCR assays combined with PrimeSafe and Dilute-'N'-Go sequencing. *Nat Protoc* 2:2429–2438
12. Rice JE, Reis AH Jr, Rice LM, Carver-Brown RK, Wangh LJ (2012) Fluorescent signatures for variable DNA sequences. *Nucleic Acids Res* 40:e164
13. Pierce KE, Khan H, Mistry R, Goldenberg SD, French GL, Wangh LJ (2012) Rapid detection of sequence variation in *Clostridium*

- difficile genes using LATE-PCR with multiple mismatch-tolerant hybridization probes. *J Microbiol Methods* 91:269–275
14. Pierce KE, Peter H, Bachmann TT, Volpe C, Mistry R, Rice JE, Wangh LJ (2013) Rapid detection of TEM-type extended-spectrum β -lactamase (ESBL) mutations using lights-on/lights-off probes with single-stranded DNA amplification. *J Mol Diagn* 15:291–298
 15. Bonnet R (2004) Growing group of extended-spectrum β -lactamases: the CTX-M enzymes. *Antimicrob Agents Chemother* 48:1–14
 16. Allawi HT, SantaLucia J (1997) Thermodynamics and NMR of internal G-T mismatches in DNA. *Biochemistry* 36:10581–10594
 17. SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* 95:1460–1465
 18. Owczarzy R, Vallone PM, Gallo FJ, Paner TM, Lane MJ, Benight AS (1998) Predicting sequence-dependent melting stability of short duplex DNA oligomers. *Biopolymers* 44: 217–239
 19. Pierce KE, Mistry R, Reid SM, Bharya S, Dukes JP, Hartshorn C, King DP, Wangh LJ (2010) Design and optimization of a novel reverse transcription linear-after-the-exponential PCR for the detection of foot-and-mouth disease virus. *J Appl Microbiol* 109:180–189
 20. Reid SM, Pierce KE, Mistry R, Bharya S, Dukes JP, Volpe C, Wangh LJ, King DP (2010) Pan-serotypic detection of foot-and-mouth disease virus by RT linear-after-the-exponential PCR. *Mol Cell Probes* 24: 250–255
 21. Zhao G, Guan Y (2010) Polymerization behavior of Klenow fragment and Taq DNA polymerase in short primer extension reactions. *Acta Biochim Biophys Sin* 42:722–728
 22. Pierce KE, Wangh LJ (2013) Rapid detection and identification of Hepatitis C Virus sequences using a molecular assay with mismatch-tolerant hybridization probes: a general method for analysis of sequence variation. *Biotechniques* 55(3):125–132
 23. Breslauer KJ (1986) Methods for obtaining thermodynamic data on oligonucleotide transitions. In: Hinz H (ed) *Thermodynamic data for biochemistry and biotechnology*. Springer, New York, pp 402–427
 24. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3 – new capabilities and interfaces. *Nucleic Acids Res* 40:e115
 25. Marras SA, Kramer FR, Tyagi S (2002) Efficiencies of fluorescence resonance energy transfer and contact-mediated quenching in oligonucleotide probes. *Nucleic Acids Res* 30:e122

Chapter 6

Multiplex PCR Primer Design for Simultaneous Detection of Multiple Pathogens

Wenchao Yan

Abstract

Multiplex PCR provides a powerful tool for simultaneous detection and discrimination of multiple pathogens or different subtypes of a causative agent from humans, animals, and plants in a single reaction, and saves time and cost in the clinical diagnostic laboratory. Here, we describe the specific protocol of multiplex PCR primer design for simultaneous identification of more than one target from a same specimen. Different sizes of amplicons and similar T_m values of primer sets are essential to successfully develop a feasible multiplex PCR assay.

Key words Multiplex PCR, Primer design, Simultaneous detection, Multiple pathogens, Different subtypes of one infectious agent, Oligo primer analysis software, Primer-BLAST tool, Sizes of amplicons, T_m values of primer sets

1 Introduction

Polymerase chain reaction (PCR) technique is widely used to amplify DNA of interest in biomedical research field and the clinical diagnostic laboratories [1, 2]. In contrast with conventional PCR, multiplex PCR has more significant advantages in the clinical diagnostic labs [3–7], and is able to simultaneously identify multiple pathogens or more than one subtypes of a same infectious agent in a single specimen such as bloods, feces, urines, and other tissues with lesions. In addition to its high specificity and sensitivity, multiplex PCR can save time and cost in detection of infectious organisms in patients or the sick animals.

Choosing suitable primers is one of the most important factors affecting multiplex PCR [2]. Oligo primer analysis software programs are commonly used to design and analyze PCR primers, and it has powerful and flexible functions in primer analysis particularly [8]. Primer-BLAST software publicly available on website [9] is a robust tool to design target-specific PCR primers, and more importantly combines BLAST with a global alignment algorithm to

check specificity of primers within all GenBank databases with broad organism coverage. Primer-BLAST offers some flexible options to adjust the specificity threshold and other primer properties [2]. Because there are more than two oligonucleotides in a reaction mixture, it is more complex that all primers in multiplex PCR could maybe form secondary structures than that of single primer pair. We should carefully analyze intramolecular and intermolecular interactions for all primers in a reaction mixture to avoid interference of duplex, hairpin structures on efficiency of multiplex PCR amplification [4]. MFold, an open source software, is available to help in evaluating the secondary structure of individual oligonucleotides in a reaction mixture [10].

In this chapter, we chose four highly pathogenic coccidia in rabbits (*Eimeria stiedae*, *E. intestinalis*, *E. flavescens*, and *E. magna*) as one example of target pathogens to illustrate the exact procedure of multiple PCR primer design for simultaneous detection of multiple pathogens (see Note 18). The specific primer pairs for four pathogenic rabbit coccidia were designed using Primer-BLAST, and then analyzed mispriming and secondary structure potential with Oligo 6 and MFold bioinformatic software to develop a rapid, specific, and sensitive multiplex PCR differential assay [6].

2 Materials

All of procedures for multiplex PCR primer design will be carried out on a computer installed with Windows XP, or Windows 7, or Linux, or Unix operating systems. Some jobs need to be finished online.

1. *Target sequences of pathogens* (see Note 1 and 18): The complete sequences of ITS1-5.8SrRNA-ITS2 of *E. intestinalis* (JX406874), *E. flavescens* (JX406873), and *E. magna* (JX406876), and the partial sequence of 18S rRNA of *E. stiedae* (HQ173837). These DNA sequences need to be searched through entering some corresponding keywords and be downloaded from GenBank [11].
2. *Primer-BLAST tool* (see Note 2): An open source and free software for PCR primer design on NCBI website [9].
3. *Oligo Primer Design and Analysis Software*: Oligo Primer Design and Analysis Software Version 6. The updated versions of it such as Oligo Primer Design and Analysis Software Version 7 are available on their official website [8].
4. *MFold software* (see Note 3): MFold version 3.6. This version of MFold can run in Unix, Linux and Mac OS X operating systems. If you want to run in Windows XP or 7, you can purchase an upgrade version from online, such as UNAFold version 3.8.exe [10].

3 Methods

3.1 Multiple Alignment of Target Sequences

1. Enter several sets of key words “*Eimeria* and ITS1-5.8S rRNA-ITS2”, “*Eimeria* and 18S rRNA and rabbit” in the blank of search engine on GenBank [11] (see Note 4), and select one of all options in drop-down menu “Nucleotide”, and finally click “Search” to obtain your target sequences including the complete sequences of ITS1-5.8SrRNA-ITS2 of six rabbit *Eimeria* species (JX406873-JX406877, JQ328190), and the partial sequences of 18S rRNA of 11 species of rabbit coccidian (HQ173828-HQ173838).
2. Select your target sequences, and then download them to your personal computer as the “FASTA” or “GenBank (full)” format (see Note 5).
3. Open the EditSeq program in the DNASTar software package, and Click “File” in the menu → “New” → “New DNA” to create some new documents with seq format.
4. Copy these sequences from the above downloaded files, and paste them to the above produced new documents of EditSeq program one by one, and respectively save them as the new Seq documents with specific names such as ITS_E_stiedae.seq, ITS_E_magna.seq, and 18S_E_stiedae.seq (see Note 6).
5. Open the MegAlign program in the DNASTar package, then click “File” in the menu → “Enter sequence...” → select the edited seq documents, and click “Add” → “Done” to enter ITS1-5.8SrRNA-ITS2 of six rabbit *Eimeria* species to the MegAlign program.
6. Click “Align” in the menu → “By Clustal W Method or By Clustal V Method” in the drop-down menu, and run for a few seconds, then show the multiple alignment result of ITS1-5.8SrRNA-ITS2 of six rabbit *Eimeria* species (seen in Fig. 1). The procedures of entering and aligning 18S rRNA sequences of 11 species of rabbit coccidia is same as described previously in the steps 5 and 6.

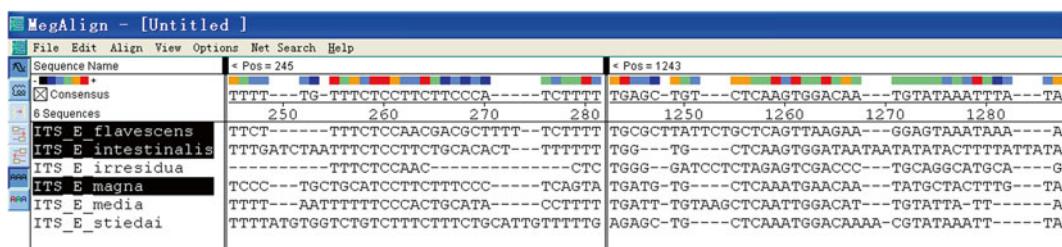


Fig. 1 Multiple sequence alignment of ITS1-5.8SrRNA-ITS2 of six rabbit *Eimeria* species. ITS1 and ITS2 regions of rabbit coccidia were species specific, and can be used as target sites for multiplex PCR primer design

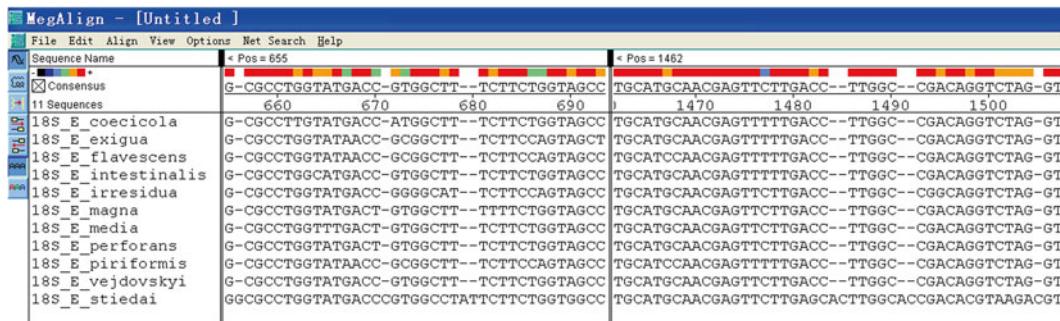


Fig. 2 Multiple sequence alignment of 18S rRNA of 11 rabbit *Eimeria* species. The 615–728 bp and 1,465–1,503 bp regions of *E. stiedae* were species specific, and can be used as target sites for multiplex PCR primer design

Table 1
The specific polymorphic sites of target sequences of four highly pathogenic rabbit coccidia

Species	Target sequences	Polymorphic regions in their own target sequences
<i>E. intestinalis</i>	ITS1-5.8SrRNA-ITS2	41–255 bp, 528–1,088 bp
<i>E. flavescens</i>	ITS1-5.8SrRNA-ITS2	32–355 bp, 518–1,041 bp
<i>E. magna</i>	ITS1-5.8SrRNA-ITS2	42–352 bp, 520–1,076 bp
<i>E. stiedae</i>	18S rRNA	615–728 bp, 1,465–1,503 bp

7. Check the sequence aligning results carefully, and find some polymorphic sites of ITS1, ITS2 (seen in Fig. 1) and 18S rRNA sequences (seen in Fig. 2), then mark these exact polymorphic regions (seen in Table 1) in your mind to provide a vital reference when you will design and analyze multiplex PCR primers for four highly pathogenic *Eimeria* species (see Note 7).

3.2 Designing Specific Primer Pairs for Individual Pathogens Using Primer-BLAST

In this section, you can use the Primer-BLAST tool to design primers for only one pathogen every time. So you have to design multiplex PCR primers for the four highly pathogenic rabbit *Eimeria* species one by one.

1. Connect to the home page of NCBI website [11], and click “BLAST” on the home page → “Primer-BLAST” on the page of BLAST, then you open the concise interface of Primer-BLAST software (seen in Fig. 3).
2. Copy each sequence of the four rabbit coccidia (including ITS1-5.8S rRNA-ITS2 of *E. intestinalis*, *E. flavescens*, *E. magna*, and 18S rRNA of *E. stiedae*) to a blank of “PCR template” on the Primer-BLAST page (Fig. 4).
3. According to the polymorphic regions (listed in Table 1), type specific ranges of forward and reverse primers in each target sequence into the blank places of “Range” item.

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred)

Range
Forward primer From To
Reverse primer

Or, upload FASTA file

Primer Parameters

Use my own forward primer (5'>3' on plus strand)

Use my own reverse primer (5'>3' on minus strand)

PCR product size Min Max

of primers to return 5

Primer melting temperatures (T_m) Min Opt Max Max T_m difference

Exon/intron selection

A refseq mRNA sequence as PCR template input is required for options in the section

Exon junction span No preference

Exon junction match Exon at 5' side Exon at 3' side 7 4 Minimal number of bases that must anneal to exons at the 5' or 3' side of the junction

Intron inclusion Primer pair must be separated by at least one intron on the corresponding genomic DNA

Intron length range Min Max 1000 1000000

Primer Pair Specificity Checking Parameters

Specificity check Enable search for primer pairs specific to the intended PCR template

Database Refseq mRNA

Organism Homo sapiens Enter an organism name, taxonomy id or select from the suggestion list as you type.
[Add more organisms](#)

Exclusion (optional) Exclude predicted Refseq transcripts (accession with XM, XR prefix) Exclude uncultured/environmental sample sequences

Entrez query (optional)

Primer specificity stringency Primer must have at least 2 total mismatches to unintended targets, including at least 2 mismatches within the last 5 bps at the 3' end.
 Ignore targets that have 6 or more mismatches to the primer.

Misprimed product size deviation 4000

Splice variant handling Allow primer to amplify mRNA splice variants (requires refseq mRNA sequence as PCR template input)

Get Primers Show results in a new window Use new graphic view

[Advanced parameters](#)

Fig. 3 The Primer-BLAST Web interface

4. In the item of “PCR parameters”, you can enter PCR product size for each rabbit coccidium in terms of the polymorphic sites obtained in Subheading 3.1. In this experiment, PCR product sizes of *E. intestinalis*, *E. flavescens*, *E. magna*, and *E. stiedae* are set as 350–450 bp, 200–300 bp, 500–650 bp, 700–900 bp, respectively (see Note 8).

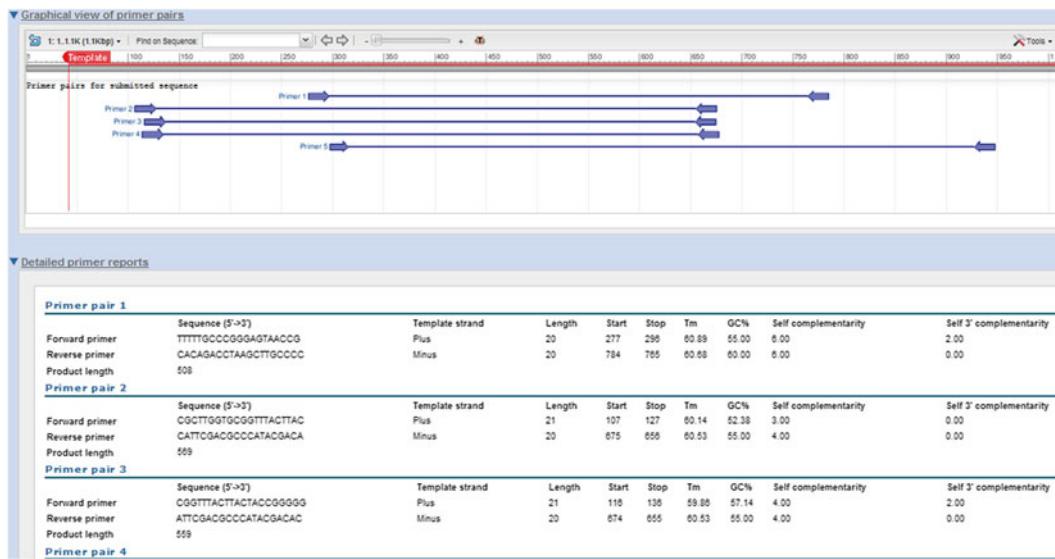


Fig. 4 Example results of designing target-specific primers for *E. magna* using Primer-BLAST

- Enter 58–63 °C in the blanks of T_m values of primers, max T_m difference is 2–3 °C [1] (see Note 9).
- For primer pair specificity checking parameters, choose the option “Enable search for primer pairs specific to the intended PCR template”, and select the “nr” database (see Note 10) to search the primers against the selected database and determine whether a primer pair can generate a PCR product on any targets in the database based on their matches to the targets and their orientations [2].
- Other parameters are set as the default values of the Primer-BLAST tool (see Note 11). Choose both options “Show results in a new window” and “Use new graphic view”, and finally click “Get Primers” button to automatically show your results of designing primers in a new window.
- Select appropriate primer pairs from the results, and develop multiple primer sets which consist of four primer pairs from the primer pairs designed by Primer-BLAST for multiplex PCR detection of four target rabbit coccidia according to sizes of PCR products and T_m values of primers.

3.3 Secondary Structure and Mispriming Analysis for the Designed Primer Pairs in Subheading 3.2 Using Oligo 6 Software

The single primer pair of multiple primer pair sets should be further analyzed with Oligo 6 program to determine secondary structure and mispriming (see Note 12) (Fig. 5).

- Open Oligo 6 program, and click “File” in the menu → “New sequence” in the drop-down menu to create a blank file.
- Copy (Ctrl+C) the target sequence of each rabbit *Eimeria* species from the previously created files with seq format, and

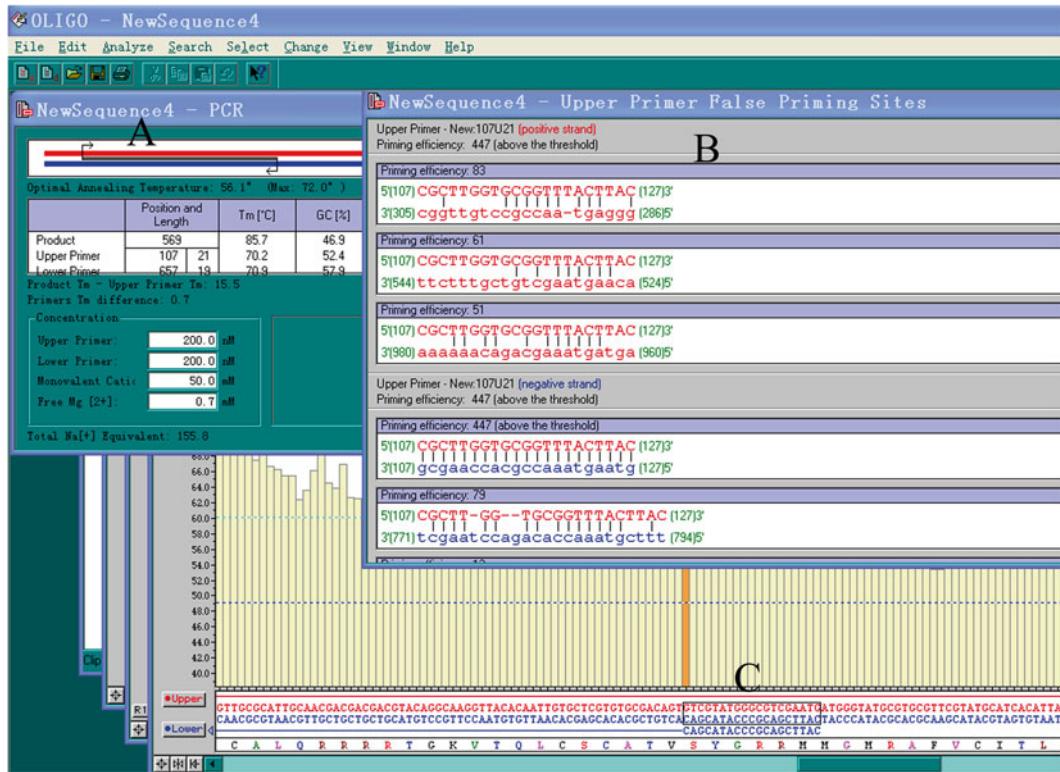


Fig. 5 The Oligo working interface. **(a)** Showing the general information on your single primer pair; **(b)** indicating all potential priming sites of your primer on both positive and negative strands; **(c)** showing the location of fixing lower primer on the template sequence in the subwindow of T_m

paste (Ctrl + V) it in the created blank file of Oligo 6, then click “Accept” button in the menu, show three new subwindows named as Melt Temperature (T_m), Internal Stability (ΔG), and Frequencies of 6 mers (Frq), respectively.

3. In the T_m subwindow, fix the exact positions for each selected primer pairs (including upper and lower primers) in the individual target sequence to evaluate mispriming and secondary structures of a single primer pair and edit them (see Note 13).
4. Click “Analyze” in the menu → “Duplex Formation” in the drop-down menu → “Upper Primer, Lower Primer, Upper/Lower” to analyze primer dimmer formation (see Note 14).
5. Click “Analyze” in the menu → “Hairpin Formation” in the drop-down menu → “Upper Primer, Lower Primer” to analyze hairpin structure formation (see Note 14).
6. Click “Analyze” in the menu → “False Priming Sites” in the drop-down menu → “Upper Primer, Lower Primer” to analyze specificities of upper and lower primers (see Note 15).
7. Click “Analyze” in the menu → “PCR” in the drop-down menu, show the general information including product size, T_m values of upper and lower primers, GC contents.

If the results show that the primers are not suitable for multiplex PCR assay, you can edit them by changing some parameters in the Oligo 6 program as follows:

8. Click “Change” in the menu → “Current Oligo Length...” in the drop-down menu → enter the oligonucleotide length as 18–22 bp to change the length of upper and lower primers.
9. Click “Edit” in the menu → “Upper Primer, Lower Primer” in the drop-down menu, and show “Edit Upper Primer, Edit Lower Primer” windows, you can add nucleotides at the 5' end of upper primer or lower primer to adjust their GC contents and T_m values.
10. After being edited, both upper and lower primers should be analyzed again according to the described protocol in the **steps 3–6** of Subheading 3.3 to obtain a pair of ideal primers for each pathogen.

After finishing primer analysis and modification, you can save the key data of PCR primers to your computer as follows:

11. Click “Analyze” in the menu → “PCR” in the drop-down menu, and then show the general information → “File” → “Save” → “Data AS” to create a txt file named as E_STIEDAE primers.txt.
12. Then click “Analyze” in the menu → “False Priming Sites” in the drop-down menu → “Upper Primer, Lower Primer”, then show information of upper/lower primer priming sites → “File” → “Save” → “Data” to save them to the same file named as E_STIEDAE primers.txt (*see Note 16*).

3.4 Compatibility Analysis of the Mixture Primers Using MFold Software

Oligo software and other primer design programs are usually employed to analyze duplex formation potential for single primer pair rather than mixture primer pairs [1]. Currently the MFold 3.6 software is able to evaluate individual primers in the multiple primer pair set to determine their compatibility for use as multiplex PCR primers [12]. Here is an example using PrimePair in the MFOLD 3.6 software to find compatible primer pairs among four forward primers and four reverse primers of the four rabbit coccidia that are typed in interactively:

1. Type “% primepair” to open the PrimePair program in Unix or Linux operating system (*see Note 17*).
2. Enter forward primers individually, one per line. End the list with a blank line as follows:

Primer 1: CGCTTGGTGC GGTTACTTAC

Primer 2: TATGAAGAACGGTTGTTG

Primer 3: ACGCTTTCGAAAGTATG

Primer 4: TGGTCATCCACCAGGTGTC

Primer 5:

3. Enter reverse primers individually, one per line. End the list with a blank line as follows:

Primer 1: CATTGACGCCATACGACA

Primer 2: CAGCAAGAACGGTGTACT

Primer 3: GGACGTGACACAGCTTACT

Primer 4: TGGTCATCCACCGGTGTC

Primer 5:

4. Select the default values of PrimePair parameters, and then enter the output file name you want such as “4Fprimepair.primepair4R”. After running for a few seconds, export the output file: 4Fprimepair.primepair4R.
5. In the output file, you will see which primers (including forward and reverse primers) are accepted, and which ones are rejected. According to your specific research objective, you can adjust parameters to relax the constraints.

Through designing primer using Primer-BLAST and analysis of secondary structures and mispriming with both Oligo and MFold software, we can obtain the suitable set of four primer pairs that have different product sizes and similar T_m values for multiplex PCR detection of *E. intestinalis*, *E. flavescens*, *E. magna*, and *E. stiedae*.

4 Notes

1. It is a key step for multiplex PCR assay to select the specific genes or sequences to your different target pathogens or various subtypes of a same organisms. These genes or sequences, which are probably from a same gene or different genes of various pathogens or subtypes, should be polymorphic and species specific [4–6].
2. The Primer-BLAST tool is an integrated primer design software with functions of BLAST and Primer 3. It can be used to not only design specific primers but also check primer specificity locally and globally using the GenBank database with a broad organism coverage [2]. The most important reason why we choose Primer-BLAST to design primers is its global alignment algorithm for specificity checking.
3. The MFold software is mainly used to analyze and predict RNA and DNA folding and hybridization. More importantly, PrimerPair in the MFold is capable of evaluating secondary structure formation of individual primers in the mixture primers to determine their compatibility for use as multiplex PCR primer pairs [12].

4. There are two routes to obtain your target sequences: one route is cloning them yourselves in lab, and other one is searching and downloading them from GenBank online [11].
5. In addition to your target sequences, you need to download more sequences of related species to your target pathogens to align and find accurately the polymorphic sites of your target sequences [6].
6. Only the sequences as a seq format is detected by MegAlign program in DNASTar package to align and analyze your target sequences.
7. Through aligning sequences, you should find and remember the polymorphic regions of each target sequence. These key results will provide an important reference for the following primer design.
8. The aim of this step is to control the length of PCR product. You have to consider that sizes of multiplex PCR products should be different enough to be distinguished in agarose gel electrophoresis [4].
9. The aim of this step is to control T_m values of each primer. Because a same annealing temperature is used in multiplex PCR reaction, T_m value of each primer of multiplex PCR should be similar when you design primers [1].
10. The traditional “nr” database, containing redundant entries, is available and is mostly recommended for organisms that are not covered by other databases or for sequence entries not covered by the RefSeq databases [2].
11. If you have special requirements for your primers, you can click the button “Advanced parameters” to reset specific parameters [9].
12. Oligo software can be used to evaluate and modify the single primer pair including upper and lower primers.
13. Including a G or C residue at the 3' end of primers increases the priming efficiency. However, primers ending with a thymidine residue tend to reduce specificity [1].
14. The presence of 3' dimmers and 3' hairpins will more seriously interfere with amplification than 5' dimmers and 5' hairpins. These secondary structures at the 3' end of primers will lead to the nonspecific amplification of sharp background products [1, 2].
15. This option “False Priming Sites” can be used to search and analyze the priming potential of each primer on both positive strand and negative one.
16. Saving these key data as a same file will provide an important reference for later PCR amplification.

17. Because the MFold version 3.6 is a Unix software, you have to open a command prompt, and then run an individual program [12].
18. If your target pathogens belong to eukaryotic organisms, or bacteria, or DNA viruses, you can directly use DNA sequences of these DNA pathogens as templates to design primers following the described procedure in this chapter. However, if your target pathogens belong to RNA viruses, you should translate RNA sequences of the RNA viruses into cDNA sequences using EditSeq program in the DNASTar package, and then use the cDNA sequences as templates to design primers for multiplex PCR assay. In addition, you should obtain cDNA sequences with RT-PCR before carrying out multiplex PCR amplification [7].

Acknowledgement

This work was supported by National Foundation of Natural and Science (31001058) and the Doctor Startup Foundation of Henan University of Science and Technology (09001350).

References

1. Apte A, Daniel S (2003) PCR primer design. In: Dieffenbach CW, Dveksler GS (eds) PCR primer: a laboratory manual, 2nd edn. Cold Spring Harbor Laboratory, New York, pp 61–74
2. Ye J, Coulouris G, Zaretskaya I et al (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 13:134
3. Henegariu O, Heerema NA, Dlouhy SR et al (1997) Multiplex PCR: critical parameters and step-by-step protocol. Biotechniques 23(3): 504–511
4. Exner MM (2012) Multiplex molecular reactions: design and troubleshooting. Clin Microbiol News 34(8):59–65
5. Fernandez S, Pagotto AH, Furtado MM et al (2003) A multiplex PCR assay for the simultaneous detection and discrimination of the seven *Eimeria* species that infect domestic fowl. Parasitology 127(4):317–325
6. Yan WC, Wang WL, Wang TQ et al (2013) Simultaneous identification of three highly pathogenic *Eimeria* species in rabbits using a multiplex PCR diagnostic assay based on ITS1-5.8S rRNA-ITS2 fragments. Vet Parasitol 193:284–288
7. Mishra B, Sharma M, Pujhari SK et al (2011) Utility of multiplex reverse transcriptase-polymerase chain reaction for diagnosis and serotypic characterization of dengue and chikungunya viruses in clinical samples. Diagn Microbiol Infect Dis 71:118–125
8. <http://www.oligo.net/>
9. http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi?LINK_LOC=BlastHome
10. <http://www.ibridgenetwork.org/rpi/unafold>
11. <http://www.ncbi.nlm.nih.gov/>
12. Markham N, Zuker M (2003) DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res 33:W577–W581

Chapter 7

Degenerate Primer Design for Highly Variable Genomes

Kelvin Li, Susmita Srivastava, and Timothy B. Stockwell

Abstract

The application of degenerate PCR primers towards target amplification and sequencing is a useful technique when a population of organisms under investigation is evolving rapidly, or is highly diverse. Degenerate bases in these primers are specified with ambiguity codes that represent alternative nucleotide configurations. Degenerate PCR primers allow the simultaneous amplification of a heterogeneous population by providing a mixture of PCR primers each of which anneal to an alternative genotype found in the isolated sample. However, as the number of degenerate bases specified in a pair of primers rises, the likelihood of amplifying unwanted alternative products also increases. These alternative products may confound downstream data analyses if their levels begin to obfuscate the desired PCR products. This chapter describes a set of computational methodologies that may be used to minimize the degeneracy of designed primers, while still maximizing the proportion of genotypes assayed in the targeted population.

Key words Degenerate PCR primers, Primer design, Mixed population PCR, Targeted resequencing, Viral genome amplification, Hierarchical clustering, Population stratification

1 Introduction

In an ideal situation, selecting PCR primers for a heterogeneous population under investigation would simply involve identifying highly conserved regions that flank the variable regions, and then choosing non-degenerate, i.e., standard, primers from those conserved regions. Standard PCR primers are less costly to synthesize, more predictable in their effectiveness, and when selected with sufficient computational scrutiny, are capable of isolating the targeted DNA region of interest very specifically and efficiently [1].

More difficulty arises when the variable regions across the heterogeneous population are distributed across the entire genome or if the variable regions are larger in size than the desired amplicon size. These constraints on amplicon size variability are especially enforced in a high-throughput sequencing center, where a single PCR protocol needs to be effective across thousands of different amplicons. High rates of PCR success are critical in order to

minimize the labor costs associated with reassessing genomic regions that could not be assayed due to prior PCR amplification failure.

The methodology described in this chapter provides the detailed steps that could be performed to maximize the success rate of degenerate primers used in the high-throughput amplification and sequencing of viral isolates [2]. The goal of these methodologies is to systematically minimize both the number of primer pairs and the degeneracy of the primers selected by careful construction of the consensus sequence used as the primer design template.

The inputs into the series of steps described below are a set of sequences that represents the targeted population of interest. The number of sequences to be considered sufficiently large depends on the amount of variation that exists in the population. Sequencing error also varies from one sequencing technology to the next, so a large enough sample of sequences should be selected so that sequencing errors can be distinguished from true population variation. For example, to filter out spurious base substitutions due to a sequencing error rate of up to 5 %, would require at least 20 sequences.

The product of the steps described is one or more degenerate consensus sequences that can be used as input into a computational primer design system. A few degenerate primer design systems are available, for example the JCVI Primer Designer [1] or NCBI's Primer-BLAST [3]. Since software systems evolve or may become obsolete over time, it is the expectation that the described methodology and the rationale behind the technique in general will endure beyond the exact implementation of any specific step.

2 Materials

2.1 Hardware/OS

A Unix operating system that can run the software described in Subheading 2.3 will be necessary. Much of the software has been developed under Linux; however, with the possible exception of some minor system differences the user may encounter, there is no known reason why Mac OS may not be used, as well.

2.2 Internet Access (Optional)

Internet access will be necessary to download sequences for the population you wish to design primers for and to download the software that will be demonstrated. However, once downloaded, the methods may be performed without any Internet access. If you choose to design degenerate primers on one of the publically available online servers, you will need Internet access after you have the degenerate consensus sequence(s) designed.

2.3 Software

1. Perl is a powerful scripting language that is generally cross-computing-platform-compatible and is commonly used by bioinformaticians. It should already be preinstalled on your operating system. You can confirm this by using the command ‘which perl’.

2. R is a powerful statistical computing package that complies with and runs on a wide variety of operating systems, including Linux, Mac OS, and Windows. It is free and you can download it from <http://www.r-project.org>.
3. ANDES is a suite of Perl and R scripts that is used to analyze multiple sequence alignments and to help design the consensus sequences [4]. It is short for “statistical tools for the ANalyses of DEep Sequencing”. It can be downloaded from Sourceforge at <http://sourceforge.net/projects/andestools>.
4. Clustalw is a multiple sequence alignment (MSA) tool (<http://www.clustal.org/clustal2/>). The alignments can be generated using clustalw [5]; however, other MSA tools are also available, including Muscle (<http://www.drive5.com/muscle/>) [6], MAFFT (<http://mafft.cbrc.jp/alignment/software/>) [7], etc. A high quality alignment will reduce spurious substitutions from being introduced into your alignment. Although it is recommended that you use the best alignment tool for your sequences, you should also be familiar with how to prepare the inputs and outputs, so they are consistent with the file formats that precede or follow, respectively.

2.4 DNA Sequences

Sequences representing the target population of interest can be based on your own collection or be downloaded from NCBI at <http://www.ncbi.nlm.nih.gov>. Sequences should be in a format that is supported by the MSA tool you wish to use. This is likely to be the fasta format.

3 Methods

The key to successful degenerate primer design is generating a consensus reference sequence and a consensus template sequence with ambiguous bases that represents the genotypic variation in the targeted population. In general, a template sequence, from which degenerate primers will be selected, will be required for all degenerate primer design. Some primer design systems, which also check the candidate degenerate primer pairs against a reference genome to perform a global alternative product check, will also require a reference genome to be specified. This reference genome can also represent the other nucleotide sequences that will be present in the biological isolate, such as host DNA, which may be unintentionally amplified. The methods that follow refer to the generation of the degenerate consensus sequence with the systematic selection of which nucleotides to make ambiguous. Figure 1 provides an overview of the key steps and decisions that will take place during degenerate primer design.

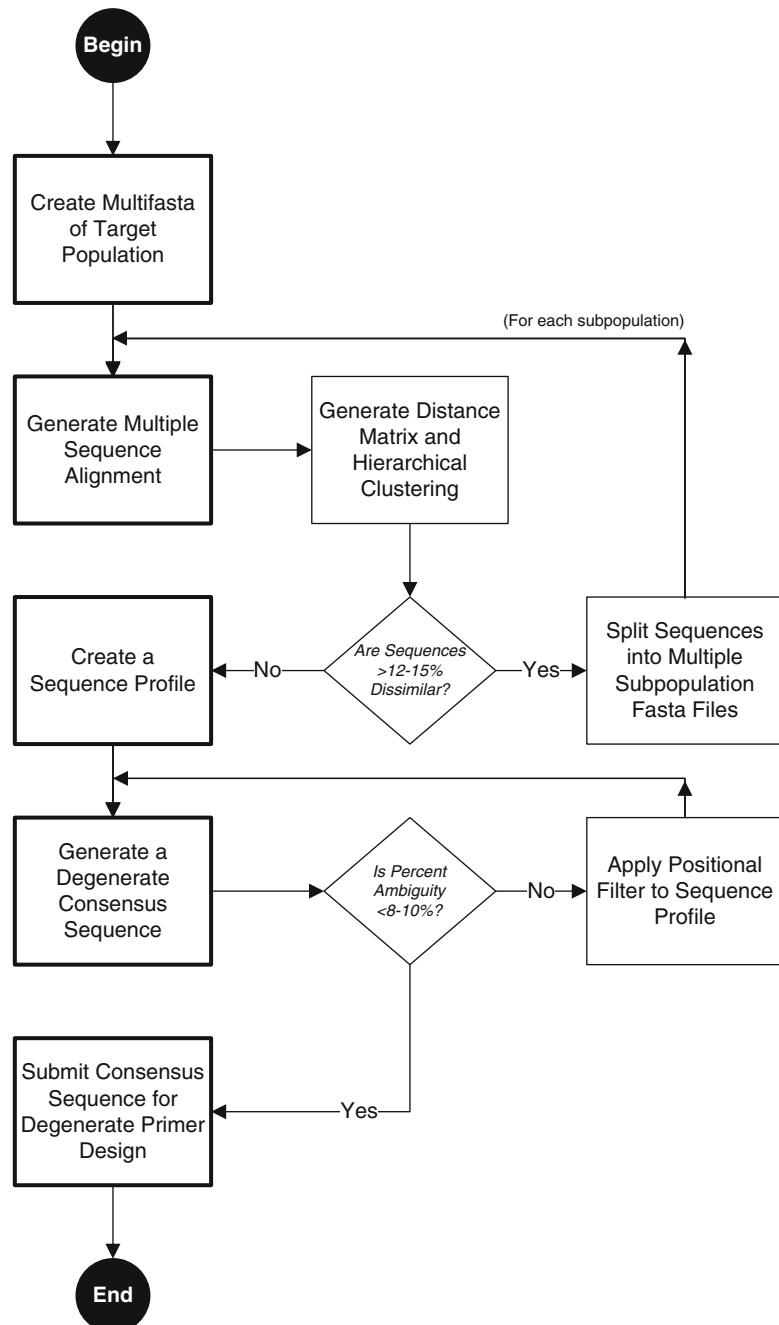


Fig. 1 Flowchart of degenerate primer design steps. Processes outlined in **bold** are required

3.1 Generating a Multifasta for the Target Population

1. Collect a representative sample of sequences for your target population. Go to a publically available database, such as NCBI to download the representative sequences. (<http://www.ncbi.nlm.nih.gov/nuccore/>).

2. Select the targeted sequences, preferably full length or complete genomes, through their interface, and download a multifasta file (*see Note 1*). Do *not* remove redundant sequences. For best results, the sequences in the multifasta should represent a sample from the targeted population.

3.2 Generating a Multiple Sequence Alignment

Generate a multiple sequence alignment (MSA) file based on the multifasta file that you created for your target population. Various MSA tools are available, such as Clustalw2 [5], Muscle [6], or MAFFT [7] (*see Note 2*). For this example, we demonstrate this step with Clustalw2.

1. To generate a multiple sequence alignment with Clustalw2, run the following command:

```
clustalw2 -infile=<name>.fasta -QUICKTREE
```

This will produce:

<name>.aln

An example of a MSA is shown in Fig. 2.

3.3 Generating a Sequence Profile from the MSA

The profile for a MSA represents the distribution of the four nucleotides and gap, for each position across the entire alignment. *See Fig. 2*. Leading and trailing gaps are not included in the profile since they are considered missing sequence.

1. To generate a profile using ANDES run the following command:

```
ClustalALN_to_PositionProfile.pl -a <name>.aln
```

This will produce the profile:

<name>.prof

An example of the contents of a sequence profile is shown in Fig. 3. The values in Fig. 3 are derived from the MSA in Fig. 2. Note that this is only a representation of the data, the actual file format may change depending on version or residue type.

3.4 Generating the Consensus Sequence from the Profile

The degenerate consensus sequence is generated based on the sequence profile by assigning the appropriate IUPAC ambiguity code when there is more than one base represented at a particular position (*see Note 3*).

<i>Sequence1</i>	-TCAGGA-TGAAC---
<i>Sequence2</i>	ATCACGA-TGAACC---
<i>Sequence3</i>	ATCAGGAATGAATCC--
<i>Sequence4</i>	-TCACGATTGAATCGC-
<i>Sequence5</i>	-TCAGGAATGAATCGCM

Fig. 2 Multiple sequence alignment

<i>Major Allele</i>	<i>A</i>	<i>T</i>	<i>G</i>	<i>C</i>	-
A	2	0	0	0	0
T	0	5	0	0	0
C	0	0	0	5	0
A	5	0	0	0	0
G	0	0	3	2	0
G	0	0	5	0	0
A	5	0	0	0	0
A	2	1	0	0	2
T	0	5	0	0	0
G	0	0	5	0	0
A	5	0	0	0	0
A	5	0	0	0	0
T	0	3	0	2	0
C	0	0	0	4	0
G	0	0	2	1	0
C	0	0	0	2	0
A	0.5	0	0	0.5	0

Fig. 3 Profile generated from MSA

1. To produce a consensus sequence with ambiguity codes, run the following ANDES command:

```
Profile_To_ConsensusFASTA.pl -p <name>.prof -c
<name>.consensus.fasta
```

This will produce the consensus fasta file:

<name>.consensus.fasta

For the profile demonstrated in Fig. 3, the degenerate consensus sequence would be:

ATCAGAWTGAAYCSCM

These are the basic steps necessary to start from a set of sequences and arrive at a degenerate consensus sequence. If the amount of diversity that is represented in your target population is low or uniformly distributed across the length of the genome, then it is unlikely that degenerate primer design will be beneficial to the project at hand. However, if the diversity in your population is significant enough that the number of ambiguity codes in your consensus sequence would lead to more than three degenerate bases per primer in the designed primer pairs, then the application of some statistical tools will be necessary to systematically reduce the number of degenerate bases in the consensus sequence. The suggested methodology described below will prioritize which degenerate bases to retain based on their prevalence in the population, using the information in the sequence profile that had been generated in Subheading 3.3.

3.5 Positional Filtering to Remove Spurious Variations

After the consensus sequence has been generated, ANDES tools may be used to check for the percentage of degenerate bases that have been introduced. (Any tool that reports the relative composition of residues in a fasta file may also replace this step.) Ideally, the targeted percent of degenerate bases should be between 8 and 10 % in the consensus sequence, depending on how evenly the degenerate bases are distributed. If the degenerate bases are clumped tightly together, then it may be possible to design an amplicon with primers that completely flank these high diversity areas. If the degenerate bases are spread apart, then primer pairs may be identified that require less than three degenerate bases per primer.

1. Check the composition of degenerate bases in the consensus sequence. The following ANDES-based command may be of use:

```
cat <name>.consensus.fasta | Report_Base_Composition.pl
```

When the percentage of bases exceeds the recommended percent ambiguity that is recommended, there are several methods available to apply filtering of the low frequency alleles using ANDES. Each of these methods has their own strengths and weaknesses, so it is worthwhile to understand how each one works. The goal of filtering in all methodologies is to remove the representation of minor alleles in the profile so that only the dominant nucleotide is present.

3.5.1 Nominal Filtering

A nominal filter requires the specification of a single threshold value. For example, if the value of 5 is specified, then any position across the length of the sequence profile with fewer than five nucleotides will be set to 0. If there were six A's and four T's, at a particular position, prior to filtering, an ambiguous base W would have been summoned to represent that base position. After filtering, 6 A's and 0 T's would remain, resulting in the non-degenerate A representing that base position.

Nominal filtering is simple and predictable, assuming the user is aware of the number of sequences that have been included in the input. The disadvantage of nominal filtering is that it is not sensitive to the number of sequences used in the input, so high-throughput degenerate primer design cannot be automated.

The following is an example of how to invoke the filter script:

```
Filter_Profile_By_Threshold.pl -i <name>.prof -o <name>.filter.prof -t 5
```

3.5.2 Percentage Filtering

A percentage filter requires the specification of a single percent threshold value. For example, if the value of 5 is specified, then any position across the length of the sequence profile with fewer than 5 % nucleotides abundance will be set to 0. If there were 25 A's and 1 T's, at a particular position prior to filtering, an ambiguous base

W would have been summoned to represent that base position. After filtering, 25 A's and 0 T's would remain, resulting in the non-degenerate A representing that base position.

Percentage filtering is adaptive because it does not depend on a fixed number of input sequences and assumes that there may be a constant rate of randomness or sequencing error across the selected sequences. The disadvantage of percentage filtering is that when the number of sequences involved is very small, the proportion of random noise at a single position may exceed the percentage threshold specified, and thus not be removed.

The following is an example of how to invoke the filter script:

```
Filter_Profile_By_Threshold.pl -i <name>.prof -o
<name>.filter.prof -p 5
```

3.5.3 Statistical Binomial Filtering

In order to compensate for the shortcomings of both the nominal and percentage filter, a statistical filter based on the binomial distribution is recommended. This filter combines both count information and percentage information that is captured for each base position. The implementation of this binomial filter utilizes the binomial distribution in order to determine whether an allele would be redetected upon resampling to the same depth based on a user-specified confidence threshold, i.e., $(1-\alpha) \times 100\%$.

For example, by specifying a 95 % confidence of presence ($\alpha=0.05$), the filter will only retain alleles that are likely to be present greater than 95 % of the time, if the sequences that were selected were to be resampled from the population. With 20 sequences, if 2/20 were T's and the remaining 18/20 were A's, according to the binomial distribution, if another 20 sequences were resampled from the population, at least 1 A would be detected 87.8 % of the time. Since 87.8 % is less than 95 %, the filter would discard the T's. However, with 40 sequences where 4/40 were T's and the remaining 36/40 were A's, the probability of detecting at least 1 T would be 98.5 %, if another 40 sequences were resampled from the population. This exceeds the 95 % cutoff, so the W ambiguity code would be used. The statistical binomial filter captures the increasing confidence of an allele's presence in the population as additional sequences are included. See **Note 4** on how to estimate the binomial percentages using R.

The advantage of this methodology is that it is grounded in statistical and sampling theory. If the input sequences are partial and result in various depths of positional coverage, then this filter is the most adaptive. The key disadvantage is that the results are less intuitive, since they depend on both the number of alleles and their prevalence at each position.

The following is an example of how to invoke the filter script:

```
Filter_Profile_By_Threshold.pl -i <name>.prof -o
<name>.filter.prof -b 95
```

3.5.4 Optimization-Based Percentage Filtering

Lastly, an optimization-based filter is also available. The user specifies the maximum percentage of degenerate bases to be allowed in the consensus sequences, and the algorithm iteratively removes alleles starting with the ones with the lowest proportion across all the sequences until the specified percentage of degenerate bases has been achieved.

For example, assume that initially the percentage of ambiguous positions in the consensus sequence was 15 % and the targeted percent ambiguity was 9 %. The optimization-based filter would iteratively increase the percentage cutoff (using the filter described in Subheading 3.5.2), until the consensus sequence had <9 % ambiguity.

The advantage of this filter is that it will reduce the number of degenerate bases in the consensus sequence by brute force. As a quick and dirty method, it is the least elegant method of achieving a consensus sequence, but the consensus sequence generated will likely be successful against the majority of the diversity represented in the input sequences.

The following is an example of how to invoke the ANDES filter script, using a 9 % ambiguity threshold across the entire consensus sequence:

```
Filter_To_Percent_Ambiguity.pl -i <name>.prof -o  
<name>.filter.prof -p 9
```

From these various filtering methodologies, it should be very clear just how important acquiring a representative sample of the target population through proper sequence selection is. All these filtering methods assume that the proportion of alleles represented in the input multifasta file are the same proportions that would be found in the targeted population.

After filtering has been performed, the filtered sequence profile is ready to be transformed into consensus sequences, Subheading 3.4.

3.6 Clustering to Separate Divergent Subpopulations

When the application of filtering is insufficient to reduce the percentage of degenerate bases in the consensus sequence, it is likely that the prevalence of the variations at many positions is not at a low enough level to be considered spurious. This occurs when the sequences that have been selected actually represent more than one divergent subpopulation. When this occurs, forcing the removal of degenerate bases in the consensus sequence will have a detrimental impact on the success of the designed primer pairs because entire subpopulations of the targeted population will not be assayed.

A solution to this problem is to design multiple sets of primers, one set for each cluster of sequences. With each cluster of sequences, the steps described previously will need to be run: MSA, profile generation, application of filtering of spurious alleles, and consensus generation. Primer design would then need to be performed on each consensus sequence, and then redundant primer pairs

across the subpopulations would be removed (*see Note 5*). In general, when the sequences used in consensus generation are between 12 and 15 % apart by percent sequence similarity, clustering may need to be performed in order to reduce the total variance per consensus sequence. The more the alleles within a subpopulation are correlated, the more significantly this clustering methodology will reduce the number of degenerate bases necessary to amplify the majority of genotypes.

Cluster analysis is performed on the MSA using ANDES.

1. Generate a distance matrix using the following command using the MSA as input.

```
ClustalALN_to_RDistanceMatrix.pl -a <name>.aln
```

This produces the following file:

```
<name>.r_distmat
```

2. Visualize the distance matrix with hierarchical clustering in a dendrogram:

```
QuickPlot_DistanceMatrix.r -i <name>.r_distmat
```

This produces the following file:

```
<name>.r_distmat.dendrogram.pdf
```

The PDF file (*<name>.r_distmat.dendrogram.pdf*) depicts the distances among sequences in a dendrogram. This gives an idea of how variable the sequences are in the original fasta file. When there are subpopulations, branch lengths will connect groups of sequences further up in the height of the dendrogram. In Fig. 4, six sequences could be separated into two clusters, each consisting of three members. The left three

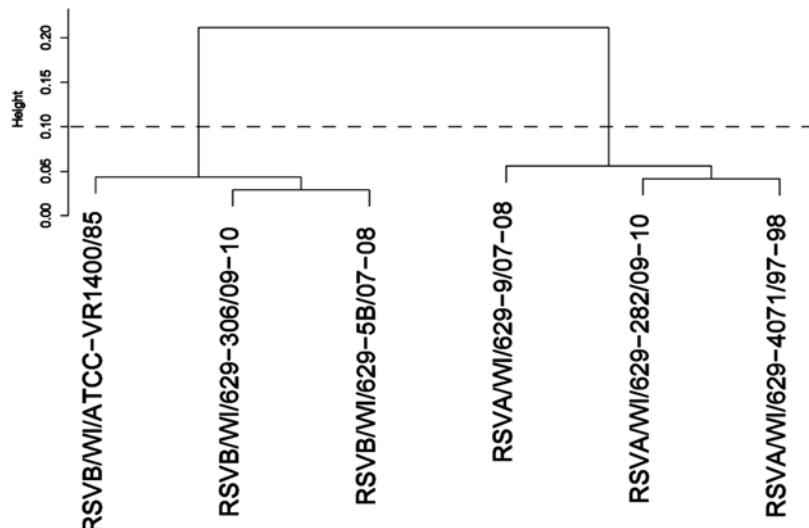


Fig. 4 Dendrogram representing percent dissimilarity between samples

and the right three members are different from each other by over 20 % identity. By cleaving the tree at a height of 10 % (0.1 in the figure), two clusters will be formed, each with an intracluster percentage difference of less than 5 %.

3. Cut the tree at a height of 10 % to generate clusters of sequences with greater than 90 % sequence similarity:

```
Partition_Members_byDistanceMatrix.r -d <name>.r_
distmat -h 0.1
```

This produces a (*<name>.r_distmat.##*) file for each cluster that was generated at the specified cutoff. Each file contains a list of sequence names which are the members of each cluster.

4. Use each of the *<name>.r_distmat.##* files to extract sequences from the original full multifasta file, to generate new multifasta files. For example:

```
Extract_Record_From_FASTA_By_List.pl -f <name>.fasta
-l <name>.r_distmat.01 > <name>.r_distmat.01.fasta
```

You will need to repeat this command for each *<name>.r_distmat.##* that was generated.

Once a new multifasta file has been generated for each cluster, Subheadings 3.2–3.4 can be repeated on each multifasta, to acquire the degenerate consensus sequences.

3.7 Designing Degenerate Primers

Once the degenerate consensus sequence has been finalized, options for degenerate primer design software include a few publicly available tools. As always, the tradeoff between ease of use and flexibility holds. An example of an online tool that is publically available is the NCBI's Primer-BLAST tool (<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>). It is simple to use because of its graphical user interface. For more advanced users with bioinformatics expertise who need additional design flexibility in a high-throughput environment, and also need a long term solution to design primers on the order of thousands of primers across many projects, the JCVI Primer Designer is recommended. (<http://sourceforge.net/projects/primerdesigner/>) [1, 8].

Both the NCBI site and the JCVI Primer Designer allows for selecting the minimum and maximum size of the amplicon, the minimum and maximum range of the primer melting temperatures, along with the desired temperature difference between the forward and the reverse primers. Additional automation features of the JCVI Primer Designer include allowing the user to specify the minimum/maximum amplicon dynamic tiling overlap and the coverage depth of amplicons desired across the template. The detection of hairpins loops and PCR stutter, and checks for both internal and end primer dimer when a sequencing adapter is utilized are also augmentations to improve overall amplification success rates.

4 Notes

1. Ideally, one would utilize all the genotype information that is available by using all sequences that have been sequenced. Partial genomes contain stretches of gaps (N's), or truncated 5' or 3' ends that may produce noisy alignment profiles, so these should only be included if there is an insufficient number of full length and complete genomes available. Maximizing the number of sequences utilized ensures capturing the heterogeneity of the population when the consensus sequence is generated.
2. Each MSA tool has its own strengths and weaknesses in terms of producing the alignment that you expect. The alignment should be checked to confirm that the edges are properly aligned, and that gaps and substitutions are introduced in the most parsimonious manner. A graphical MSA tool, that allows the user to make and save minor adjustments, may be useful if you cannot adjust the tool's alignment parameters to produce the results you want. When edges are not aligned well, they may be trimmed before moving onto the next step to avoid spurious introduction of ambiguous nucleotides in the consensus. When using an alternative MSA tool, you should invoke the run options so that it produces a clustal aln formatted output file.
3. The EMBOSS tool `cons` may also be used to generate a consensus sequence directly from a MSA. However, the usage of a sequence profile is recommended since it will allow extra flexibility in determining which positions should be allowed to be degenerate based on some additional statistical analyses.
4. To pre-compute the binomial thresholds in R, you would use the `pbinom` function. For example, if 2/20 bases are T's and 18/20 bases are A's, then the probability of seeing a T upon resampling another 20 sequences from the same population would be `1-pbinom(q=0, size=20, prob=2/20)`. This equals 0.8784233. Similarly, if 4/40 bases were T's and 36/40 were A's, then `1-pbinom(q=0, size=40, prob=4/40)` would equal 0.9852191.
5. To remove redundant primer pairs, the only criterion that needs to be considered is the exact sequence of the forward and reverse oligo. Other primer pair characteristics, such as genomic position or amplicon melting temperature will vary depending on which consensus sequence was used. The format of the output primer pairs may vary depending on your choice of primer design software. When using the JCVI primer design software, a script in the `PrimerDesignResultsTools` directory, named `Remove_Duplicates_From_PrimerCritiquorCSV.pl`, reads in a comma-separated-value file that contains the primer pairs concatenated from multiple primer design runs. With only one

primer pair record per line, redundant primer pairs are quickly removed by only checking the two columns containing the forward and reverse sequences.

Acknowledgement

The development of the degenerate primer design consensus sequence construction strategy was funded in whole or part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, and Department of Health and Human Services under contract number HHSN272200900007C. This work was supported by the Office of Biological and Environmental Research in the DOE Office of Science award DESC0006837.

References

1. Li K, Brownley A, Stockwell TB, Beeson K, McIntosh TC, Busam D et al (2008) Novel computational methods for increasing PCR primer design effectiveness in directed sequencing. *BMC Bioinformatics* 9(1):191
2. Li K, Shrivastava S, Brownley A, Katzel D, Bera J, Nguyen AT et al (2012) Automated degenerate PCR primer design for high-throughput sequencing improves efficiency of viral sequencing. *Virol J* 9(1):261
3. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden T (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13:134
4. Li K, Venter E, Yooseph S, Stockwell TB, Eckerle LD, Denison MR et al (2010) ANDES: statistical tools for the ANalyses of DEep Sequencing. *BMC Res Notes* 3(1):199
5. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H et al (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947–2948
6. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797
7. Katoh K, Kuma KI, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33(2):511–518
8. Li K, Brownley A (2010) Primer design for RT-PCR. In: RT-PCR protocols. Humana, Totowa, NJ, pp 271–299

Chapter 8

Allele-Specific Real-Time Polymerase Chain Reaction as a Tool for Urate Transporter 1 Mutation Detection

Juliet O. Makanga, Antonius Christianto, and Tetsuya Inazu

Abstract

Allele-specific polymerase chain reaction (ASPCR) method has long been applied for the detection of nucleotide variations and genotyping, which are detected by the presence or absence of DNA amplification PCR products. Recently, Real-Time PCR genotyping has fast developed and offered a rapid method of detecting mutations without the need of gel electrophoresis as with ASPCR. Here, we describe an easy and rapid touchdown real-time PCR method for the detection of nucleotide variations. Using our method we successfully detect two main mutations in human urate transporter 1 (*SLC22A12*), W258X and R90H, and validate the results. The method can potentially be applied to genotype of various other nucleotide variations.

Key words Allele-specific real-time polymerase chain reaction, Urate transporter, Genotyping

1 Introduction

The concept of allele-specific PCR (ASPCR) was first developed by Newton et al. [1]. Newton and his group came up with ARMS (Amplification Refractory Mutation System) that was based on 3' mismatch in oligonucleotides that caused primers not to anneal to template DNA under specific conditions. This enables distinguishing of different alleles and the detection of single nucleotide polymorphisms (SNPs). Although this method did not require restriction enzyme digestion or DNA sequence analysis of PCR products as was conventionally the case, it still required agarose gel electrophoresis as a means of genotyping. This is not only a tedious, time-consuming process but also limits the number of samples that can be examined at a time and may be affected by formation of spurious PCR amplification products. Nonetheless, ASPCR has become widely utilized and many approaches have been developed to ease post-PCR processing. Real-time detection with fluorescence-labeled oligonucleotides (e.g., TaqMan®, Invader assay) is one such method. Although these methods are effective, they are costly as

they make use of expensive fluorescence-labeled oligonucleotides. Employing the kinetics of real-time PCR has enabled the use of nonspecific DNA fluorescent dyes to efficiently detect mutations [2]. Here we describe a method that uses SYBR green®, a nonspecific DNA fluorescence dye, in combination with a touchdown thermal cycling protocol to detect nucleotide variations with high efficiency [3]. Our method does not require restriction enzyme cleavage or PCR products purification. The results are interpreted automatically based on ΔCt evaluation and hence do not require any specialized laboratory personnel. This method offers a great alternative to the TaqMan® and real-time PCR methods that use expensive fluorescence-labeled oligonucleotides.

2 Method

2.1 Principle of the Method

The principle for this method is illustrated in Fig. 1. As is with the conceptual ASPCR primer design, allele-specific primers identical in sequence except for 3' nucleotide matching each allele are designed. There is a difference in amplification efficiencies between two, matched and mismatched, reactions. In cooperating the touchdown PCR protocol increases stringency hence ensuring that only the specific primer match reaction is initially amplified. As cycling progresses, lower and more permissive annealing temperature allows amplification of the mismatched primer reaction. This produces a difference in Ct values of the respective reactions. Ct value refers to the fractional cycle number at which fluorescence passes threshold and is generated automatically by the StepOne (Applied Biosystems) software. This difference in Ct values is defined as ΔCt . Theoretically, low negative ΔCt represents homozygous allele A, high positive a homozygous allele B and close to 0 the heterozygous allele A/B. Using the quantitative discrimination achieved by ΔCt value offers an easy and rapid method for genotyping.

3 Protocol

3.1 Materials

- 10 ng genomic DNA.
- Allele-specific primers and respective opposite primer.
- SYBR® Premix Ex Taq™ (Takara JAPAN).
- PCR tubes.
- Sterile distilled water.
- StepOne (Applied Biosystems) Thermal cycler.

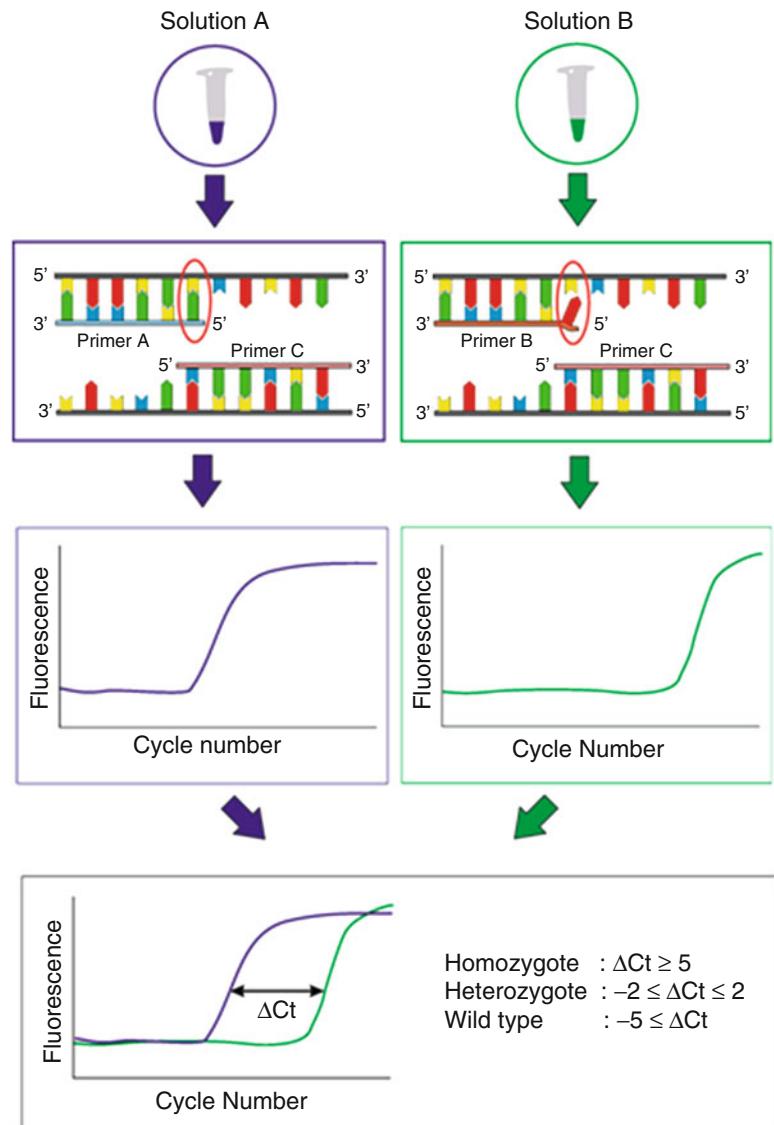


Fig. 1 For each sample and each SNP, two reactions are set up (Solution A and B). The reactions are of identical composition except for the allele-specific primers corresponding to each nucleotide variations (Primer A and primer B) and the corresponding reverse primer (Primer C). Representative amplification plots (Cycle number vs. Fluorescence) are shown. Each of the primer pairs amplify DNA template with different efficiency resulting in the difference in Ct value. The difference, ΔCt , is used to discriminate allele type. High positive ΔCt value represents homozygote, low negative ΔCt value represents wild, and close to zero represents heterozygote

Table 1
PCR setup

	Initial concentration	Final concentration	Final volume per 10 µl reaction
Genome DNA ^a	–	10 ng	X µl
Allele-specific primer ^b	5 µM	0.3 µM	0.6 µl
Common primer ^b	5 µM	0.3 µM	0.6 µl
SYBR Premix	2×	1×	5 µl
dH ₂ O			Up to 10 µl

^aTotal amount of template genome DNA added is dependent on starting sample concentration and will need to be determined empirically. As little as 10 ng is sufficient although higher concentrations can be used

^bWe recommend the use of freshly diluted primers as multiple freezing and thawing of primers may reduce ΔCt values (an observation in our laboratory)

Critical Notes

The length of the allele-specific primers must be optimized so as to have an estimated annealing temperature as close as possible as the respective opposite primer.

3.2 PCR Setup

For each reaction tube, set up a reaction as follows (Table 1 PCR setup) on ice. Be sure to mix all reagents well.

Critical Notes

Ct value is known to be affected by the quantity of template DNA used in the reaction. We tested detection sensitivity of PCR amplifications at various concentration of genomic DNA ranging from 100 to 3.125 ng/ml. Results did not vary within this range of initial genomic DNA material, and we obtained similar ΔCt values with the varying DNA concentrations. Needless to say, concentration should be constant in sister experiments.

We observed a reduction in ΔCt values following the use of primers that had been frozen and thawed multiple times. It is recommended that primers are stored in aliquots and used fresh for each assay.

SYBR Premix Ex TaqII manufacturer recommended a larger reaction volume. In our hands, we were able to reduce reaction volume down to 10 µl with no effect on results.

It is very important to mix all the reagents well for consistent and reproducible results.

3.3 Thermal Cycling

Thermal cycling is then performed in two phases as shown in Table 2 below at the appropriate annealing temperature. We used StepOne (Applied Biosystems) Thermal cycler, but a variety of other real-time thermal cyclers available can be used.

Table 2
Thermal cycling

Phase 1	Step	Temperature	Time
1	Initial denature	95 °C	30s
2	Denature	95 °C	3 s
3	Anneal	Tm + 10 °C ^a	20 s
Repeat steps 2–3 for 12 cycles			
Phase 2	Step	Temperature	Time
4	Denature	95 °C	30 s
5	Anneal	Tm	20 s
6	Hold	4 °C	∞
Repeat steps 4–5 for 32 cycles			

^aAnnealing temperature is decreased by 0.8 °C during each cycle until the estimated Tm temperature is reached after around 12 cycles

Critical Notes

Phase 1 annealing temperature is crucial in this method. The initial annealing temperature should be set 10 °C above the estimated optimum annealing temperature and decreased by 1–0.5 °C during each cycle until estimated Tm temperature is reached after around 12–15 cycles. Too low Phase 1 annealing temperature will result in unreliable ΔCt values.

3.4 Application

In humans, urate is the end product of purine metabolism as they lack uricase, which is responsible for the conversion of urate into allantoin. Serum urate levels are regulated by reabsorption in the kidney's proximal tubule. Urate transpoter1 (URAT1) encoded by *SLC22A12* [4, 5] is a major player in reabsorption of urate and evidence has shown defects in *SLC22A12* results in idiopathic renal hypouricemia whose prevalence is high in Japanese and non-Ashkenazi Jews populations. Below we describe the detection of two *SLC22A12* mutations, W258X (nucleotide change of *G774A*) and R90H (nucleotide change of *G269A*) alleles [3, 6, 7].

Table 3
Primer setting

Primer name	Sequence
W258X-Wt-Forward Primer (774G)	5'-TAC GGT GTG CGG GAC TGG-3'
W258X-Mt-Forward Primer (774A)	5'-TAC GGT GTG CGG GAC TGA-3'
W258X-Reverse Primer	5'-GCC AGC CAC CAG GAG TAC AA-3'
R90H-Forward Primer	5'-CCC TCC TGG ACA ACA GCA C-3'
R90H-Wt-Reverse Primer (269G)	5'-ACT GTG GCT GGC GGA AGC-3'
R90H-Mt-Reverse Primer (269A)	5'-ACT GTG GCT GGC GGA AGT-3'

3.5 Materials

- 10 ng genomic DNA extracted from human blood by Gentra Puregene Blood kit (QIAGEN) according to the manufacturer's instructions.
- Allele-specific primers with 3' nucleotide matching each allele and a corresponding primer for W258X and R90H, respectively (Table 3).
- SYBR Premix Ex Taq® (Takara JAPAN).
- PCR tubes.
- StepOne (Applied Biosystems) Thermal cycler.

Reaction mixture was prepared as mentioned in the protocol section and PCR cycling performed with a starting annealing temperature at 70 °C in Phase 1 of cycling protocol. Phase 2 annealing temperature was maintained at 60 °C for 33 cycles of amplification.

Representative amplification plots of representative samples are shown in Fig. 2. We found that the respective allele-specific primers resulted in ΔCt values that were in accordance with our theoretical range for calling each allele. Difference in the Ct values between positive and negative reactions was at least five cycles, indicating that specific amplifications are greater than 32-fold in comparison to the nonspecific ones. To confirm the stringency of our method, we run the PCR products on 2 % agarose gel stained with ethidium bromide (0.5 µg/µl) and visualized them under UV light confirming the absence of any

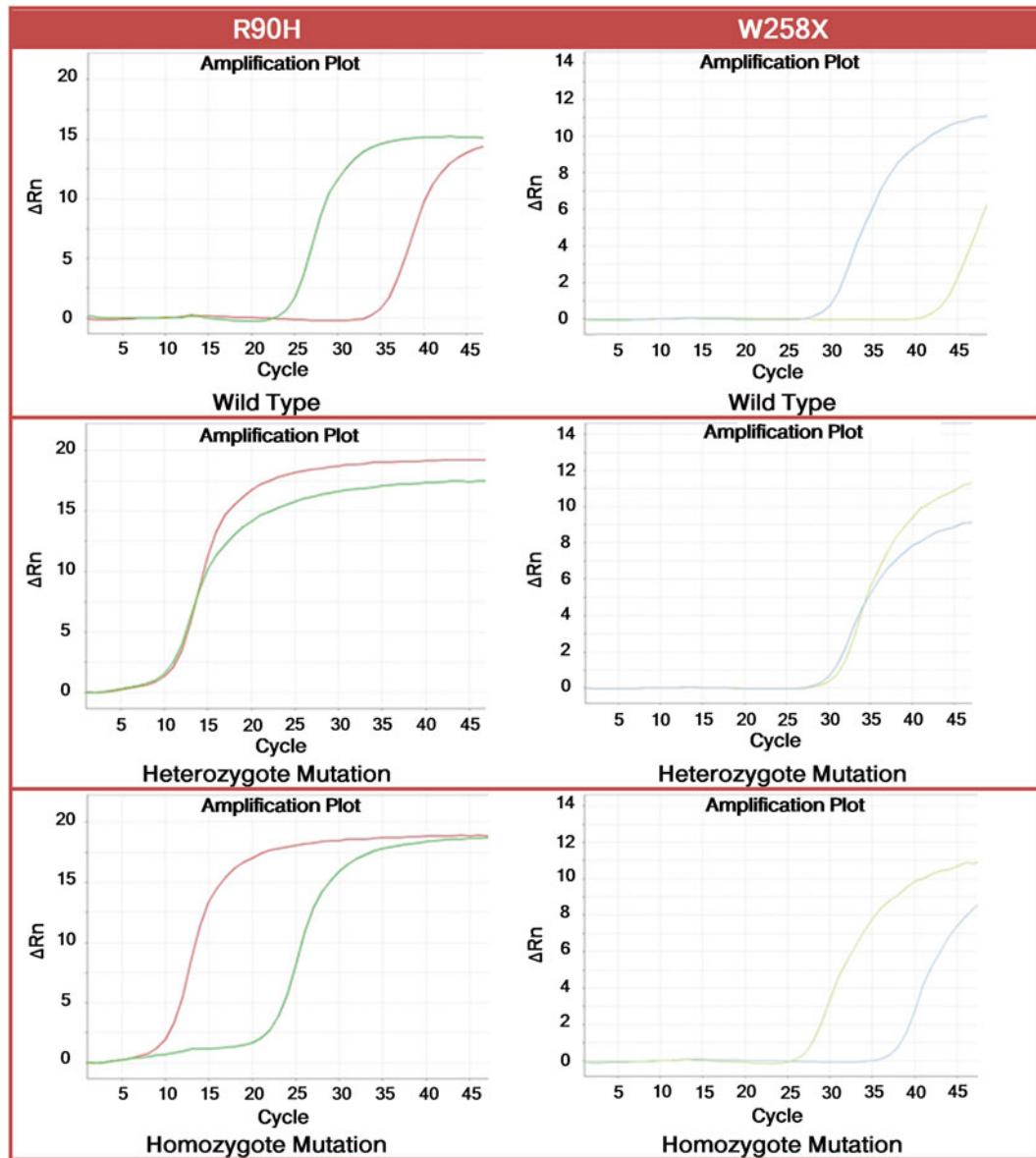


Fig. 2 Amplification plots showing representative results of W258X and R90H genotypes examined are shown. All genotypes were clearly determined and distinguished by difference in Ct values, ΔCt . ΔCt for wild-type samples was less than -5 and greater than 5 for homozygote samples

spurious PCR products (Fig. 3). DNA sequencing further validated the accuracy of the results. As mentioned earlier, end product analysis by gel electrophoresis and/or DNA sequencing is not necessary and was performed purely for validation purposes (Fig. 4).

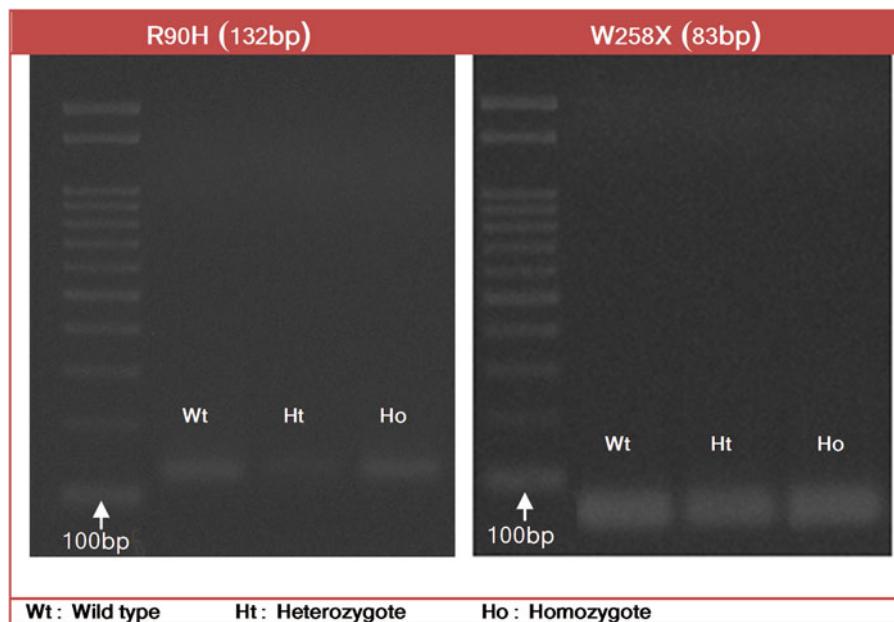


Fig. 3 Result verification. Amplified PCR products were run on agarose gels confirming the products were of single bands of expected size

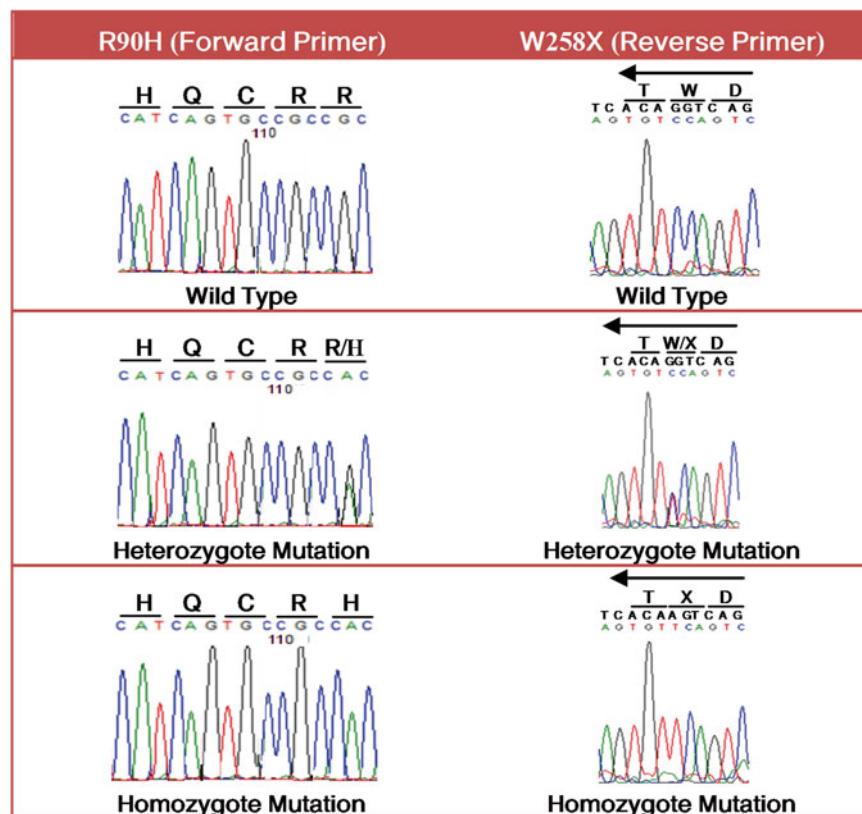


Fig. 4 Result verification. PCR product extracted from the gel and sequenced. DNA sequencing results confirmed authenticity of the results obtained from ΔCt analysis

4 Conclusion

Our method offered a simple, rapid, and relatively inexpensive means of successfully detecting W258X and R90H mutations in human urate transporter 1 (*SLC22A12*). As the method primarily entails primer design with 3' nucleotide modifications matching respective single nucleotide polymorphisms, it can be potentially applied to detect other mutations.

References

1. Newton CR, Graham A, Heptinstall LE et al (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Res* 17:2503–2516
2. Wu WM, Tsai HJ, Pang JH et al (2005) Touchdown thermocycling program enables a robust single nucleotide polymorphism typing method based on allele-specific real-time polymerase chain reaction. *Anal Biochem* 339:290–296
3. Takagi S, Omae R, Makanga JO, Kawahara T, Inazu T (2013) Simple and rapid detection method for the mutations in *SLC22A12* that cause hypouricemia by allele-specific real-time polymerase chain reaction. *Clinica Chimica Acta* 415:330–333
4. Enomoto A, Kimura H, Chairoungdua A et al (2002) Molecular identification of a renal urate-anion exchanger that regulates blood urate levels. *Nature* 417:447–452
5. Anzai N, Kanai Y, Endou H (2007) New insights into renal transport of urate. *Curr Opin Rheumatol* 19:151–157
6. Inazu T (2006) A case of renal hypouricemia caused by urate transporter 1 gene mutations. *Clin Nephrol* 65:370–373
7. Inazu T, Kawahara T, Ishikawa T (2007) Rapid detection of R90H mutations in the human urate transporter 1 gene. *Ann Clin Biochem* 44:189–191

Chapter 9

MultiPLX: Automatic Grouping and Evaluation of PCR Primers

Lauris Kaplinski and Maito Remm

Abstract

In this chapter we describe MultiPLX—a tool for automatic grouping of PCR primers for multiplexed PCR. Both generic working principle and step-by-step practical procedures with examples are presented.

MultiPLX performs grouping by calculating many important interaction levels between the different primer pairs and then distributes primer pairs to groups so that the strength of unwanted interactions is kept below user-defined compatibility level. In addition it can be used to select optimal primer pairs for multiplexing from list of candidates.

MultiPLX can be downloaded from http://bioinfo.ut.ee/?page_id=167. Graphical web-based interface to most functions of MultiPLX is available at <http://bioinfo.ut.ee/multiplex/>.

Key words PCR, Multiplex, Primer design

1 Introduction

As the scope of genomic studies is growing exponentially each year, the cost-effectiveness of each working step is of even bigger importance. Multiplexing PCR is one approach to save time and sample DNA if large number of genomic fragments has to be amplified by PCR. Multiplexing introduces new set of factors affecting the success of PCR, namely unwanted interactions between different primer pairs. The number of potential interactions is proportional to the square of the number of primer pairs in single tube (multiplexing level). For small number of PCR and low multiplexing levels groups are often created by hand using trial and error, but this is not a reasonable solution if thousands of PCR reactions have to be multiplexed with high levels.

To automate large-scale multiplexing, we have created a program called MultiPLX [1]. It analyzes many interactions between primer pairs and generates groups that effectively minimize the number of unwanted interactions in every tube.

MultiPLX can also be integrated with primer design to choose primers from a list of candidates, based on their suitability for multiplex PCR. We have also successfully used it for multiplexing hybrid primers and nested PCR reactions.

MultiPLX is a command-line tool, available both for Linux (UNIX) environment and Windows command line.

2 Working Principle

An important set of factors that affect the success of multiplexed PCR are unwanted bindings between primers and products from different PCR sets. Nearest-neighbor thermodynamics can be effectively used to predict the binding strength (Gibbs free energy ΔG) between oligonucleotides and thus estimate the success rate of multiplexed PCR. Program MultiPLX evaluates all possible pairwise combinations of PCR primer pairs and calculates up to five different thermodynamic “score” values (worst case ΔG) for each combination. These scores can be used to determine which pairs are compatible with each other and can be put into single group. The decision is based on user-selected cutoff values.

In addition to the thermodynamic binding energies, the differences between primer melting temperatures, the differences between product lengths, and one user-defined score can be taken into account. *See Note 1* for more detailed explanation of different factors that affect the success of multiplexed PCR.

The workflow of MultiPLX has the following steps:

1. Reading, calculating, and saving score data
2. Reading, calculating, and saving grouping data

Different tasks can be combined; for example, if scores are already calculated, there is no need to recalculate these, but they can be read from file instead.

MultiPLX can also be integrated into primer design workflow, so the final primers for each PCR are chosen from among the candidates, based on the compatibility with multiplexing.

3 Step-by-Step Overview

The generic syntax of MultiPLX is:

```
cmultiplx -primers FILE [PRIMERARGS]  
[SCOREARGS] [GROUPARGS]
```

The option “-primers” specifies PCR definition file. It specifies all primer pairs and products, one pair per line (Fig. 1). Values have to be tab-separated, different products from single primer pair space-separated. Specifying products is not obligatory if they are not known or will not be used in analysis.

```

pcr_A      TTAATCACCCGGCTCCCAGC      GGAGGGCTGACACAGGGAGG      TTAATCACCCGGCTCCC
AGCCGTGTTCTGCAGAAGGAGCTTTCTAATTCAAGTGTCCAGCCAGGAGGACATAAGTAGAACAGGTG
GAAGTGCCTGACTCTGCTATTCCCTCTCCATCCGTATTAGACCAGTTCCCTATCATTAATAAAAAAC
ACGAAACAAAGACAATGAGGGAAGGGTGGTCTCATTTATCACCCAGACCCAGCCTCCCTGTGTGAGGCCCTCC
pcr_B      GCCCAATTCCATGCTAAAGCGA      AGGGGAGCCCATTCTCGGATTGAG      GCCCAAT
TTCCATGCTAAAGCGAAGCATATAAGTAGCAGTAGAATCAGCTCATCTGCATTGTTCCCCAGCGAGGA
GCAATCAATTACCTGAGAGATGCCGTGGCTAAACGGTGACCCAACAAGGTGCAATTGCCAAAGTGTC
ACTTTATTGTCATGCCCTCAAATCCGAGAATGGGCTCCCT
pcr_C      GCCTTGAACAGAAAGGCAGGA      CAACTTGGCATGGATTGGACG      GCCTTGAACAGAA
AGGCAGGAGGATACTGTACAAGTTCAAGACAGAGAGAGGGATGGTAGGTGGTCAGTAGCAAGTGATGAATTCAT
AAAATGAGAATCTGCCTCTAGGAGGAGTTAAAGGAATACCAATTGAAATGAGACTCTATCTCCA
GGAGAGGAGTTCAAAGTAATACCAAGTGCCAAAGAGCCTCAGCAGAGATATGAACGTGACCGTGTGCCAGAA
GAAGGAAGGGCCCAGCTGCCAGATGTTACGTCCAATCCATGCCAGTT

```

Fig. 1 Sample from a PCR definition file. The columns are: PCR reaction name, left primer sequence, right primer sequence, and (optional) product sequence(s)

3.1 Step A: Calculating Compatibility Scores

In compatibility score tables, the worst-case thermodynamic binding energies between different PCR primer pairs are stored. As calculating of these is very time-consuming procedure, MultiPLX allows experimentator to save once calculated score table into files. This way the once calculated scores can be used many times, for example during adjusting parameters for optimal grouping solution.

The interactions between primers and products are calculated by using nearest-neighbor (NN) thermodynamic model. The binding energies of NN pairs are read from file, and the default file can be easily replaced if better NN parameters for specific experiment conditions are available.

Please see **Note 2** for the file format of the thermodynamic file.

There are two types of score files—one for primer-primer interactions (three scores for each PCR primer pair combination) and the other for primer-product (two scores for each PCR combination) interactions.

Please see **Note 3** for the explanation of different score types.

The scores will be calculated with the following option:

-calcscores SCOREIDS

The parameter SCOREIDS is a sequence of the numeric identifiers for scores (explained below).

An example of calculating score files:

```

test>./cmultiplx -primers primers-10.txt -calcscores 12345 -saveprimscores primer-scores.txt -saveprodscors product-scores.txt

```

“primers-10.txt” is the primer definition file. The option “**-calcscores 12345**” instructs MultiPLX to calculate all five thermodynamic compatibility scores (nonthermodynamic scores are always calculated on-the-fly, so they cannot be saved to intermediate score file). If all thermodynamic scores are not needed, some digits can be left out. For example “**-calcscores 123**” calculates only scores for primer-primer interactions.

primer-scores.txt and **product-scores.txt** are the output files where scores will be saved. Although they are text files, the information in these is normally not meant to be read by users.

The possible SCOREIDS can be listed with the following command:

```
cmultiplex -listscores
```

```
test>cmultiplex -listscores
Score codes:
1 - Primer-primer both 3' ends
2 - Primer-primer 3' end with any region
3 - Primer-primer any regions
4 - Primer-product 3' end with any region
5 - Primer-product any regions
```

Calculating primer-product scores takes much more processing time than calculating only primer-primer scores. We have also found that primer-product interactions have smaller influence to PCR quality than primer-primer interactions. For large datasets, especially if small multiplexing level is needed, these can be left out.

An example of different score types is given in Fig. 2.

3.2 Step B: Calculating Groups

Primer pairs are distributed into groups based on cutoff levels of score values. Any two primers are allowed to be in the same group only if all relevant scores are below cutoff values. Although this is simplification and does not distinguish the level of compatibility of the group as a whole, it makes computation much more straightforward.

In general, it is not possible to find the optimal solution without trying all possible combinations (unrealistic even on supercomputers).

Score type 1 - alignment of the 3' ends of two primers

```
5'-TTAATCACCCGGCTCCCAGC-3'
      |||   |
3'-GAGTTTAGGCTCTTACCCGAGGGGA-5'
```

Score type 2 - alignment of the 3' ends one primer with the internal region of other primer

```
5'-TTAATCACCCGGCTCCCAGC-3'
      ||
3'-AGCGAAATCGTACCTTAACCCG-5'
```

Score type 3 - alignment of the internal regions of two primers

```
5'-TTAATCACCCGGCTCCCAGC-3'
      |||||||
3'-GAGTTTAGGCTCTTACCCGAGGGGA-5'
```

Fig. 2 Different score types

MultiPLX uses a greedy algorithm for grouping, which gives reasonably good solution.

The most important parameters for grouping can be set with the following options:

```
-stringency LEVEL
-cutoff# VALUE
-calcgroups MAXGROUPS MAXITEMSINGROUP
```

Examples of calculating groups:

```
test> cmultiplx -primers primers-10.txt
      -loadprimscores primer-scores.txt -load-
      prodscores product-scores.txt -stringency
      normal -calcgroups 1000 10 -savegroups
      groups.txt

test>cmultiplx -primers primers-10.txt
      -loadprimscores primer-scores.txt -load-
      prodscores product-scores.txt -stringency
      low -cutoff1 -5 -cutoff8 4 -calcgroups 1000
      10 -savegroups groups.txt
```

Both examples read PCR primers and products from the **primers-10.txt** file and previously calculated scores from **primer-scores.txt** and **product-scores.txt**.

The first example performs grouping, using a generic balanced set of cutoff values (**-stringency normal**). The maximum number of groups is 1,000 and the maximum number of elements in group is 10 (**-calcgroups 1000 10**). Final groups will be saved to the **groups.txt** file. The format of grouping file is shown in Fig. 3.

The second example uses looser set of cutoff values (**-stringency low**) that normally would make bigger groups. Then it overrides two of these values. The minimum allowed dG for primer-primer binding from 3' ends is set to -5 kcal/mol (**-cutoff1 -5**) and the maximum allowed difference between the melting temperatures of primers to 4 °C (**-cutoff8 4**). Other parameters are identical to the first example.

Using “-” as the grouping file name will print the grouping information to screen instead of file.

3.2.1 Stringencies and Cutoffs

Stringency specifies generic set of balanced values for grouping cutoffs. The allowed levels are “low”, “normal”, and “high”. Low stringency results in the biggest groups but with bigger probability of PCR failure due to incompatibilities between primers.

The cutoff values of different stringencies are listed in Note 4.

For more detailed control, or if nonthermodynamic scores have to be taken into account, individual cutoffs can be overridden with “**-cutoff#**” option. The number sign “#” has to be replaced with an individual cutoff code (thus resulting in option names like **-cutoff1**, **cutoff2...**). Both stringency and cutoff options may be

```
# Libbdm build 1.0.1-06-05-2003
# Total 19 rows 6 column in table
MXMultiplexTable
Name      1      2      3      4      5      6
Group 1   1      59     47     82     63
Group 2   61     54     38     50     65
Group 3   10     37     67     77     33
Group 4   94     95     57     98     55     46
Group 5   97     8      22     18     30
Group 6   17     91     39     80     68
Group 7   31     41     79     5      69
Group 8   11     56     9      60     48
Group 9   75     78     70     2      83
Group 10  88     44     72     3      32     21
Group 11  20     58     14     28     52
Group 12  90     29     87     7      100
Group 13  34     96     71     27     53     76
Group 14  64     49     40     73     81
Group 15  42     62     35     93     23     43
Group 16  99     15     13     86     12     24
Group 17  6      19     45     4      85
Group 18  84     66     16     36     74
Group 19  25     92     26     89     51
```

Fig. 3 The format of grouping file. Groups are listed in rows and PCR pair names in columns

used together, in which case specific cutoff values replace the ones determined by stringency.

The list of possible cutoff codes can be displayed with the following option:

-listcutoffs

```
test>cmultiplx -listcutoffs
Cutoff codes:
1 - Primer-primer both 3' ends
2 - Primer-primer 3' end with any region
3 - Primer-primer any regions
4 - Primer-product 3' end with any region
5 - Primer-product any regions
6 - Product length max difference (range)
7 - Product length min difference (ladder)
8 - Maximum primer melting temperature difference
9 - Custom score (maximum allowed)
```

There is no direct way to request certain predefined group size or certain number of groups. If such solution is needed, one has to

set the maximum group size or the maximum number of groups to desired value, and experiment with different cutoffs, until acceptable solution is found. As the grouping is usually very fast (unless very big number of optimization iterations is used), this is only a minor inconvenience.

Please see **Note 5** for more information about grouping.

3.3 Integrating MultiPLX with Primer Design

Calculating groups for existing list of primers usually gives quite good multiplexing levels. Still, as only a single primer pair is available for any amplified region, MultiPLX can only rearrange primer pairs between different groups. Difficulties in grouping may appear if some primer pairs have very strong interactions (and thus low compatibility) with many other pairs. This difficulty can be alleviated by combining primer design and primer grouping.

Primer design usually involves picking a single pair out of many possible candidates for each target region. MultiPLX can read more than one primer pair candidate for each target region and select the one that gives the smallest number of unwanted interactions.

MultiPLX primer input file can list arbitrary number of candidate pairs for each target region. The target region is identified by name. The primer pair candidates from the same target region can share common primers. An example of PCR primer file with multiple candidate primer pairs is shown in Fig. 4.

There are two options specific to primer selection:

-maxcandidates NUMBER
-savefinalset FILENAME

An example of selecting primers for multiplexing:

```
test>cmultiplx -primers primer-candidates.txt
      -maxcandidates 10 -calcscores 123 -saveprimerscores primer-scores.txt

test> cmultiplx -primers primer-candidates.txt
      -maxcandidates 10 -loadprimscores primer-scores.txt -stringency normal -calcgroups 1000 1000 -savegroups groups.txt
      -savefinalset final-primers.txt
```

pcr_A	TTAATCACCCGGCTCCCAGC	GGAGGGCTGACACAGGGAGG
pcr_A	TTAATCACCCGGCTCCCAGC	CGTAGCTATGGCATCGATT
pcr_A	GCATGCCTATAAGCGATGGAC	CGTAGCTATGGCATCGATT
pcr_B	GCCCCAATTCCATGCTAAAGCGA	AGGGGAGCCCATTCTCGGATTGAG
pcr_B	GCCTTGAACAGAAAGGCAGGA	CAACTTGGCATGGATTGGACG

Fig. 4 Sample from a PCR definition file with several primer pair candidates for each target region. The columns are: PCR reaction name, left primer sequence, right primer sequence. The product sequences are missing in this example, but if present, they can be listed in the fourth column. Notice that some primer sequences can be shared between alternative primer pairs

In the first step of this example, primer candidates are read from the file **primer-candidates.txt**. Up to ten candidate primer pairs are selected from each target region (**-maxcandidates 10**). Primer-primer interaction scores are then calculated and written to **primer-scores.txt** file.

In the second step, the same number of candidate primer pairs is read again from the primer file and the scores for all these candidate primer pairs (calculated in previous step) are read from **primer-scores.txt**. Primer pairs are then grouped using normal stringency (**-stringency normal -calcgroups 1000 1000**). Using large value (**1000**) for both the number of groups and the number of group members ensures that the size of groups is actually limited by primer compatibility. If there is more than one primer pair candidate for PCR, MultiPLX automatically chooses the most compatible one (the one having the smallest number of unwanted interactions with all other primer pairs). Groups are then written to **groups.txt** and the chosen set of primer pairs to **final-primers.txt**.

It is important to use the same number of candidates (**-maxcandidates**) for both score table generation and grouping, because the values in score table are identified by the positions of primers and pairs in the primer file.

4 Specific Applications

4.1 Testing Groups

MultiPLX can be used for evaluating an existing multiplexing solution by pointing out these primer pairs that have high probability of unwanted pairings with other group members. Primer compatibility for existing groups will be evaluated by using the same score table and cutoff values of scores as for calculating groups. Existing groups have to be presented in the same file format as the groups generated by MultiPLX.

Existing groups can be evaluated with the following option:

-saveoffenders FILENAME

An example of evaluating existing multiplexing groups:

```
test>./cmultiplx -primers primers-100.txt
      -loadprimscores primscores-100.txt -strin-
      gency low -cutoff1 -5.5 -groups test-
      groups.txt -saveoffenders -
      1
      95      72      PrimPrimEnd2      -5.8
      95      32      PrimPrimEnd2      -5.5
      2
      58      82      PrimPrimEnd2      -5.6
      3
      8       7       PrimPrimEnd2      -5.9
```

17	75	PrimPrimEnd2	-5.8
4			
39	44	PrimPrimEnd2	-5.7
5			
6			
38	64	PrimPrimEnd2	-5.7
86	88	PrimPrimEnd2	-5.9
7			
96	61	PrimPrimEnd2	-5.8
61	18	PrimPrimEnd2	-5.5

In this example the existing groups are read from the **test-groups.txt** file and evaluated using low stringency and user-specified cutoff1. The output lists the groups and all of their members, which have some scores above cutoff values.

4.2 Universal and Complex Primers

Universal PCR primers that are linked to both ends of a studied DNA fragment are frequently used in large-scale genomic applications. The compatibility of universal primers is usually tested in design phase and does not need to be re-evaluated by the MultiPLX program. However, if hybrid primers with universal 5' end and target-specific 3' end are used, the MultiPLX might be useful to test their compatibility and/or to group them.

As complex primers are longer than simple ones, the score calculation with MultiPLX is slower than for simple primers. As the effect of universal (5') end will be the same for all primers, it is possible to limit calculating scores to only target-specific part of primer. To do that, new primer definition file, where all universal primer fragments are removed, has to be constructed using some external tool.

This approach does not take into account the possibility of alignment between the middle part of one primer (partly specific, partly universal) with the end or the middle of another primer or product.

4.3 Custom Scores

MultiPLX can use one additional score file to introduce custom factors into multiplexing. For example, the number of predicted PCR products amplified by any given pairwise combination of PCR primers from template genome can be calculated using the GenomeTester package [2] and used as input for MultiPLX. The custom version of GenomeTester for this is accessible from the webpage of the Department of Bioinformatics, University of Tartu (<http://bioinfo.ebc.ee/gt4multiplx/>).

MultiPLX allows using only a single custom score value. If a more than one custom score is needed, these have to be combined into single value beforehand. Also the value must have negative correlation with PCR success (i.e., big values are bad, small values are good).

5 Web Interface

MultiPLX can be accessed from the webpage of the Department of Bioinformatics, University of Tartu (<http://bioinfo.ut.ee/multiplx/>).

6 Notes

1. The parameters affecting the success of PCR multiplexing

The unwanted bindings can be broadly divided into different classes, based on whether they take place between two primers or one primer and one product and whether the 3' end of primer is bound or free.

In addition to bindings between primers and products, it is often useful to consider also the melting temperatures of primers and the differences in product lengths.

(a) Primer-primer interactions

For multiplexed PCR there are four different interactions between any pair of PCR primers. Thus all possible primer pairs have to be tested for unwanted interactions with all other primers in the same multiplexing group, and if their alignment is too strong, moved to another group.

All primer-primer interactions lower the concentration of free primers and thus the probability of binding the primer to the target site. Additionally, the interactions involving the 3' end of one of both primers create a new possible elongation site.

(b) Primer-product interactions

Primer-product interactions have similar effect as primer-primer interactions. But as product concentrations are much lower (at least in the beginning of PCR experiment), the effect is lower as well.

In the case of products we do not have to differentiate the bindings between the 3' end of the primer with any region of the product and the bindings between the 3' end of a primer with the 3' end of the product, as the overall effect is similar in both cases.

(c) Additional properties (melting temp, product length)

In addition to unwanted bindings, it is often useful to limit the maximum difference between the melting temperatures of primers in a single group. The melting temperature of primers is correlated with the speed of PCR and it is better to keep all individual PCR rates in multiplex group as similar as possible. The difference in product lengths has a similar effect, thus it may be useful to limit the maximum difference between product lengths as well.

If final PCR products will be detected by gel electrophoresis, it may be necessary to have different product lengths in a single group, so they will separate on gel and thus can be individually detected.

The melting temperatures of primers depend on PCR reaction mixture parameters (the concentration of monovalent and divalent cations and the concentration of DNA). These parameters can be specified by the following command-line options:

```
-csalt SALT
-cmg MAGNESIUM
-cdna DNA
```

SALT is the concentration of monovalent salts in mM
(default 50)

MAGNESIUM is the concentration of magnesium in mM
(default 1.5)

DNA is the concentration of primers in nM (default 50)

Although the salt and DNA concentrations affect absolute melting temperatures of primers, they are usually irrelevant for multiplexing, as only temperature differences between melting temperatures are used. As the effect is identical for all primers, differences remain the same. So if the exact experiment conditions are not known beforehand, defaults (or other reasonable values) are safe to use for multiplexing.

MultiPLX uses the following formula to calculate melting temperature:

$$T_m = \frac{dH}{dS} + 1.987 \times \ln(\text{cDNA}) + 16.6 \times \log(\text{Na+}[NORM])$$

Tm—melting temperature in Kelvins

dH—binding enthalpy in J/mol

dS—binding entropy in J/(mol K)

Na+[NORM]—normalized salt concentration

Na+[NORM] is calculated from the following formula [3]:

$$\text{Na+}[NORM] = (\text{Na+}[mM] + 120.0 \times \sqrt{(\text{Mg2+}[mM] - \text{dNTP}[mM])}) / 1000;$$

Na+[mM], Mg2+[mM]—the concentrations of monovalent and divalent cations in mM/l

dNTP[mM]—the total concentration of nucleotide triphosphates in mM/l

2. Thermodynamic data

Thermodynamic data table lists all dinucleotide combinations and corresponding thermodynamic nearest-neighbor parameters.

As MultiPLX has to calculate pairings with mismatches, both Watson-Crick pairs and pairs with single mismatch have to be listed in the table. In addition to these, missing nucleotides can be specified to take the contribution of dangling ends into account.

All missing values are treated as 0. The default parameters are from published sources [4–10].

Thermodynamic data is specified for MultiPLX with the following option:

```
-thermodynamics FILENAME
```

A sample from thermodynamic data file is shown in Fig. 5.

3. Score types used by MultiPLX

MultiPLX uses the following score types:

1. Maximum binding energy (ΔG) of two primers including 3' ends of both primers (PRIMPRIMEND2).
2. Maximum binding energy of 3' end of one primer with any region of another primer (PRIMPRIMEND1).
3. Maximum binding energy of any region of different primers (PRIMPRIMANY).
4. Maximum binding energy of 3' end of one primer with any region of PCR product (PRIMPRODEND1).
5. Maximum binding energy of any region of a primer with any region of PCR product (PRIMPRODANY).
6. Maximum product length difference between compared PCR primer sets.
7. Minimum product length difference between compared PCR primer sets.
8. Maximum difference in primer melting temperatures between compared PCR primer sets.

4. Cutoff values for different stringencies

Stringencies represent generic balanced set (according to our knowledge) of cutoffs for grouping. Depending on the number of PCR primer pairs to be grouped, high stringency may give groups of about 1–4, normal stringency 3–8, and low stringency 7–15 PCR primer pairs.

The default cutoff values of stringencies take neither product length differences nor melting temperature differences into account, so if these should be used, specifying explicit cutoff values is needed.

The cutoff values of stringencies are given in Fig. 6.

5. Grouping algorithm

Primer pairs are selected one by one, and tested against each member of each existing group. If a primer pair is compatible with an existing group, and it has fewer items than maximum number allowed, the item is placed in that group.

```
# libdna thermodynamic table
#
# Format: PQ/RS dH dS
# P,Q,R,S - nucleotides
# dH Enthalpy (cal/mol)
# dS Enthropy (kcal/mol*K)
#
Energy+Enthropy
#
# Watson-Crick-Pairs
# SantaLucia J. and Hicks D. 2004. The Thermodynamics of DNA
# Structural Motifs.
# Annu. Rev. Biophys. Biomol. Struct. 33:415-40
#
AA/TT -7600 -21.3
TT/AA
AC/TG -8400 -22.4
AG/TC -7800 -21.0
AT/TA -7200 -20.4
CA/GT -8500 -22.7
GT/CA -8400 -22.4
TG/AC
CC/GG -8000 -19.9
GG/CC -8000 -19.9
CG/GC -9800 -24.4
GC/CB -9800 -24.4
CT/GA -7800 -21.0
GA/CT -8200 -22.2
TC/AG
TA/AT -7200 -21.3
#
# A-C Mismatches
# Allawi H.T. and SantaLucia J. 1998. Nearest-neighbor
# thermodynamics of internal A-C mismatches in DNA: sequence
# dependence and pH effects.
# Biochemistry 37:9435-44
#
AA/CT 7600 20.2
TC/AA
AA/TC 2300 4.6
CT/AA
AC/CG -700 -3.8
GC/CA
AC/TA 5300 14.6
AT/CA
AG/CC 600 -0.6
CC/GA
CA/AT 3400 8
TA/AC
CA/GC 1900 3.7
CG/AC
CC/AG 5200 14.2
GA/CC
...
...
```

Fig. 5 Sample from a default thermodynamic file

```

Low stringency (-stringency low)

PrimPrimEnd2:          -6.0
PrimPrimEnd1:          -10.0
PrimPrimAny:           -10.0
PrimProdEnd1:          -14.0
PrimProdAny:           -14.0
Max prod len diff:    -
Min Prod len diff:   -
Max Melting temp diff: -
Max custom score:    -

Normal stringency (-stringency normal)

PrimPrimEnd2:          -4.0
PrimPrimEnd1:          -8.0
PrimPrimAny:           -8.0
PrimProdEnd1:          -12.0
PrimProdAny:           -12.0
Max prod len diff:    -
Min Prod len diff:   -
Max Melting temp diff: -
Max custom score:    -

High stringency (-stringency high)

PrimPrimEnd2:          -2.0
PrimPrimEnd1:          -6.0
PrimPrimAny:           -6.0
PrimProdEnd1:          -10.0
PrimProdAny:           -10.0
Max prod len diff:    -
Min Prod len diff:   -
Max Melting temp diff: -
Max custom score:    -

```

Fig. 6 Cutoff values of stringencies

Otherwise it is tested against the members of next group and so on. If it is incompatible with all existing groups and the maximum number of groups is not exceeded, a new group is created, otherwise the grouping fails with an error message “**ERROR: Cannot fit primers into groups**”.

The number of final groups is sensitive to the order of primers to be grouped. The order can be modified by the following option:

`-initialorder VALUE`

The allowed values are “**file**”, “**friends**”, and “**random**”.

File keeps primer pairs in the same order that they are in the input file.

Friends orders primer pairs by the number of compatible pairs, starting from the smallest (fewest compatible pairs) value. Thus the primers that have smaller probability to be compatible with others are distributed before the primers that have higher

probability of success. Usually this option gives the optimal solution (smallest number of groups).

Random distributes primers randomly. It can be useful, if nondeterministic grouping is desired, for example for testing the compatibility of different primer selection methods with multiplexing.

For random order grouping more than one iteration can be performed with the following option:

-groupiter NUMITERATIONS

The solution with the smallest number of groups among all iterations will be chosen as the final result.

Greedy grouping results in groups of very different sizes. Usually the first groups are the biggest, because all primers are placed into the first compatible group.

The size of the groups can be made more uniform with the following option:

-optimize NUMITERATIONS

The optimization is on by default and the default number of iterations is 10,000. To turn it off, the number of iterations has to be set to 0.

The optimization works by moving primer pairs from bigger groups into smaller groups and swapping the members between groups if no pair can be moved into smaller group. It does not guarantee a uniform group size, but under normal circumstances the sizes of groups do not differ by more than one.

Acknowledgements

This work was supported by the grant EU19730 from Enterprise Estonia. Development of primer design software in our group has been funded by Centre of Excellence in Genomics at Estonian Biocentre (EU European Regional Development Fund). The authors thank Katre Palm for a valuable help with English grammar.

References

1. Kaplinski L, Andreson R, Puurand T et al (2005) MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics* 21:1701–1702
2. Andreson R, Reppo E, Kaplinski L et al (2006) GENOMEMASKER package for designing unique genomic PCR primers. *BMC Bioinformatics* 7:172
3. von Ahsen N, Wittwer CT, Schutz E (2001) Oligonucleotide melting temperatures under PCR conditions: nearest-neighbor corrections for Mg²⁺, deoxynucleotide triphosphate, and

- dimethyl sulfoxide concentrations with comparison to alternative empirical formulas. *Clin Chem* 47:1956–1961
- 4. Allawi HT, SantaLucia J (1997) Thermodynamics and NMR of internal G-T mismatches in DNA. *Biochemistry* 36:10581–10594
 - 5. Allawi HT, SantaLucia J (1998) Nearest-neighbor thermodynamics of internal A-C mismatches in DNA: sequence dependence and pH effects. *Biochemistry* 37:9435–9444
 - 6. Allawi HT, SantaLucia J (1998) Nearest-neighbor thermodynamic parameters for internal G-A mismatches in DNA. *Biochemistry* 37:2170–2179
 - 7. Allawi HT, SantaLucia J (1998) Thermodynamics of internal C-T mismatches in DNA. *Nucleic Acids Res* 26:2694–2701
 - 8. Kaderali L (2001) Selecting target specific probes for DNA arrays. Universität zu Köln, Köln
 - 9. Peyret N, Senevirtane PA, Allawi HT et al (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A-A, C-C, G-G and T-T mismatches. *Biochemistry* 38:3468–3477
 - 10. SantaLucia J, Hicks D (2004) The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* 33:415–440

Chapter 10

In Silico PCR Primer Designing and Validation

Anil Kumar and Nikita Chordia

Abstract

Polymerase chain reaction (PCR) is an enzymatic reaction whose efficiency and sensitivity largely depend on the efficiency of the primers that are used for the amplification of a concerned gene/DNA fragment. Selective amplification of nucleic acid molecules initially present in minute quantities provides a powerful tool for analyzing nucleic acids. In silico method helps in designing primers. There are various programs available for PCR primer design. Here we described designing of primers using web-based tools like “Primer3” and “Web Primer”. For designing the primer, DNA template sequence is required that can be taken from any of the available sequence databases, e.g., RefSeq database. The in silico validation can be carried out using BLAST tool and Gene Runner software, which check their efficiency and specificity. Thereafter, the primers designed in silico can be validated in the wet lab. After that, these validated primers can be synthesized for use in the amplification of concerned gene/DNA fragment.

Key words Polymerase chain reaction (PCR), Primer3, Web Primer, Gene Runner, DNA polymerase, RAPD, DNA isolation, Purification, Validation

1 Introduction

The polymerase chain reaction (PCR) is an efficient molecular biology technique used to amplify a small amount of DNA up to million folds in a short span of time. It is an enzymatic reaction where replication of DNA takes place using DNA polymerase enzyme. Here, DNA replication is carried out at higher temperature by exploiting use of a thermo-stable DNA polymerase. It is based on the ability of DNA polymerase to amplify DNA for specific DNA template. The use of PCR can be exploited for various purposes viz. amplification of human-specific DNA sequences; differentiation of species, subspecies, and strains; DNA sequencing; detection of mutations; monitoring cancer therapy; detection of bacterial and viral infections; predetermination of sex; linkage analysis using sperm cells; ascertaining recombinant clones; and studying molecular evolution [1]. Primers are required as DNA polymerase can add nucleotides only to existing 3' OH group. That's why selectivity and efficacy of PCR mainly depends on the

efficiency of the primers required for the amplification of concerned DNA [1–5]. An oligonucleotide can serve as a primer; however, its efficiency depends on the following factors [3]:

- (a) Annealing and extension temperature, kinetics of association, and dissociation of primer–template duplexes.
- (b) Duplex stability of mismatched nucleotides and their location.
- (c) The efficiency with which the polymerase can recognize and extend a mismatched duplex.

For successful amplification, proper primer design is the most important step. As the name suggests, primer is the prime (first) nucleic acid for the attachment of DNA polymerase. For maximal specificity and efficiency of PCR, amplification of optimal primer sequence and appropriate primer concentration are essential. A poorly designed primer can result in little or no product due to nonspecific amplification and/or primer-dimer formation that may become competitive enough to suppress product formation.

Proper selection of oligonucleotide primer is critical for PCR, DNA sequencing, and oligo-hybridization. Proper PCR primer can be designed with the help of bioinformatics tools and software. Various bioinformatics programs are available for designing primers from the template sequence. All these programs help in designing primer, but wet lab validation is further required. All results from PCR primer design program must be validated in the laboratory before doing their synthesis.

2 Materials

2.1 *In Silico Primer Designing*

There are variety of software and tools available for the design of primers for PCR (Table 1). These are mostly freeware software, which are available on the internet. However, a few are commercial programs [6, 7]. A program may be a stand-alone program or a complex integrated networked version of the commercial software. These software packages may be for complete DNA and protein analyses, secondary structure predictions, primer design, molecular modeling, development of cloning strategies, plasmid drawing, or restriction endonuclease analyses. One cannot compare two programs using two different algorithms that will focus on different parameters. Even if the same parameters are equivalently set for both programs, discrepancies may be there due to differences in their calculation methods and the order in which the selection criteria are applied. Besides, different programs attack the task of primer selection differently. Here, we discussed a couple of tools for Primer designing that can be used to design primers for PCR.

Table 1
Different available tools for Primer designing

Tool	Description	URL
Primer3	Comprehensive PCR primer design tool; easy to accept defaults option	http://primer3.wi.mit.edu/
NetPrimer	For individual or primer pair	http://www.premierbiosoft.com/ netprimer/index.html
Gene Fisher2	For degenerate primer and can accept unaligned sequence	http://bibiserv.techfak.uni-bielefeld.de/ genefisher2/
Web Primer	Design primer for PCR as well as for sequencing	http://www.yeastgenome.org/cgi-bin/ web-primer
Primer BLAST	Target specific primer	http://www.ncbi.nlm.nih.gov/tools/ primer-blast/
PCR Designer	For restriction analysis of various sequence mutations	http://primer1.soton.ac.uk/primer.html
CODEHOP	Consensus-Degenerate hybrid oligonucleotide primers	http://bioinformatics.weizmann.ac.il/ blocks/codehop.html
Primo 3.4	A complete package for PCR primer, contains Primo Pro, Primer Degenerate, etc.	http://www.changbioscience.com/ primo/primo.html
eprimer3	Picks PCR primers and hybridization oligos	http://emboss.bioinformatics.nl/cgi-bin/ emboss/eprimer3
PrimerQuest	Design primers of 50 sequence at a time	http://www.idtdna.com/biotools/ primer_quest/
Methprimer	For designing bisulfite-conversion-based Methylated PCR Primers	http://www.urogene.org/methprimer/

2.1.1 Primer3

Primer3 was developed by Rozen and Skaletsky [8]. This program and its source code are freely available on internet (<http://primer3.wi.mit.edu/>). Primer3 is a computer program that suggests primers for PCR. This software has been provided by Whitehead Institute for Biomedical Research. In selecting primer, Primer3 tries to balance equally primer length, primer melting temperature, and product length. While giving input sequence, user selects specific characteristics of output primer called as constraints. The user selects these options through text boxes, check boxes, and pull down menus.

The top of output displays sequence id and informational notes. It is followed by best primer pairs and their characteristics including primer length, primer temperature, and so forth. The next information is quasi-graphical representation of the left (>>>>) and right (<<<<) primers in the source sequence. Finally the output contains a section headed statistics which indicates reasons why individual primers are unacceptable.

2.1.2 Web Primer

The Web Primer tool facilitates the design of primers for use in sequencing or PCR. It is freely available and it can be downloaded from the web site, <http://www.yeastgenome.org/cgi-bin/web-primer> [9]. Sequencing primers are evenly spaced along the DNA, whereas PCR primers are at the ends of DNA selected in a region of DNA which has to be amplified.

For using Web Primer:

1. The user must input sequence as the locus name or the actual DNA sequence.
2. Thereafter, experiment type for primer is selected that can be a sequence primer or a PCR primer.
3. As submit button is clicked, an intermediate parameter page is opened to customize the parameters for the primer such as location of primer (with respect to locus entered or DNA sequence), melting temperature (Tm), annealing temperature (Ta), etc. If window shows that the parameters are not satisfactory and primers cannot be designed, one must follow the instructions in the resulting error message and reset the parameters.

2.1.3 Gene Fisher2

Gene Fisher2 is mainly for degenerate primers, which are the primers based on the sequences of homologous genes [10]. It is freely available (<http://bibiserv.techfak.uni-bielefeld.de/genefisher2/>). At first, the consensus sequence is calculated and thereafter, the possible priming region for forward and reverse primers is determined. Input is given as either nucleotide or amino acid sequence and it can be a single or multiple sequence (less than 100).

Gene Fisher is an interactive web-based program that automates the task. The procedure leads to isolation of genes in a target organism using multiple alignments of related genes from different organisms. The term “gene fishing” refers to the technique where PCR is used to isolate a postulated but unknown target sequence from the pool of DNA.

2.1.4 Gene Runner

Gene Runner is a simple molecular biology analysis program. It is a freely available Windows program, which can be downloaded from the web site, www.generunner.net. DNA and protein sequences can be submitted manually or after downloading from other databases. It performs multiple sequence alignment and can search cleavage sites, restriction sites, and open reading frames (ORFs). Gene runner now has protein motif database searching abilities also. One may visit to NCBI web site and download Genbank flat files for searching. One can search dbEST or GenBank for secondary protein structure by protein motifs in all six translation frames using Motif runner.

2.2 In Silico Primer Validation

2.2.1 Basic Local Alignment Search Tool (BLAST)

BLAST is an algorithm for comparing primary biological sequence information [11]. It finds the region of local similarity between sequences. The program compares nucleotide or protein sequences of sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help in identifying members of gene families. Different types of BLASTs are available according to the query sequences like blastn, blastp, blastx, tblastx, tblastn, megablast, psi-blast, and phi-blast.

2.2.2 NetPrimer

NetPrimer combines the latest primer design algorithms with a web-based interface allowing the user to analyze primers over the internet (<http://www.premierbiosoft.com/netprimer/>). The designed primers are analyzed for secondary structures including hairpins, self-dimers, and cross-dimers in primer pairs. This ensures the availability of the primer for the reaction as well as minimizing the formation of primer dimer.

2.3 Wet Lab Validation

2.3.1 For DNA Isolation and Purification

- (a) Grinding medium: 100 mM Tris–HCl buffer, pH 8.0 containing 175 mM EDTA, 70 mM NaCl, 2 % CTAB, and 30 mM 2-mercaptoethanol.
- (b) Chloroform:Isoamylalcohol (24:1, v/v).
- (c) Isopropanol.
- (d) 70 % (v/v) Ethanol.
- (e) TE buffer: 10 mM Tris–HCl buffer containing 1 mM EDTA, pH 8.0.

2.3.2 For Polymerase Chain Reaction and RAPD Analysis

- (a) Reaction mixture, 25 μ l: 2.5 μ l of 10 \times polymerase buffer (supplied by the manufacturer), 100 μ M of each of the four dNTPs, 2.5 mM MgCl₂, 25 ng of each of the forward and backward primers, 20 ng of genomic DNA, 1.2 units of Taq DNA polymerase in a total of 25 μ l volume.

If large number of tubes have to be prepared, a master mix may be prepared using PCR Recipe [12, 13], and thereafter, the same can be dispensed in all the tubes. It will reduce pipeting errors.

3 Methods

3.1 In Silico Primer Designing Using Primer3/Web Primer

The use of software in biological applications has given a new dimension to the field of biology. Many different programs for the design of primers are now available as described above. Here, we described use of Primer3 and Web Primer as these are freely available user-friendly web services for designing primer.

3.1.1 Using Primer3

The following steps are used for in silico primer designing:

1. All nucleotide sequences of concerned gene are searched in RefSeq database (<http://www.ncbi.nlm.nih.gov/refseq>). The Reference Sequence (RefSeq) database aims to provide a comprehensive, integrated, nonredundant set of sequences, including genomic DNA, transcript (RNA), and protein products for major research organisms [14].
2. After that, multiple sequence alignment is done using Clustal Omega tool [15]. All the sequences that are retrieved from RefSeq are aligned to get the conserved region of the sequence.
3. Thereafter, the conserved region of the sequence is inserted in the input box of the GUI of Primer 3. Generally, the parameters are adjusted as follows: Tm min.= 52 °C, max.= 58 °C; Primer size min.= 18, max.= 24; GC% min.= 50, max.= 60.
4. Afterwards, **Pick Primer** button is clicked. The software designs forward and backward primers.

3.1.2 Using Web Primer

1. All nucleotide sequences of concerned gene are searched in RefSeq database (<http://www.ncbi.nlm.nih.gov/refseq>). The Reference Sequence (RefSeq) database aims to provide a comprehensive, integrated, nonredundant set of sequences, including genomic DNA, transcript (RNA), and protein products for major research organisms.
2. After that, multiple sequence alignment is done using Clustal Omega tool. All the sequences that are retrieved from RefSeq are aligned to get the conserved region of the sequence.
3. The conserved region of the sequence is fed in the input box of the GUI of Web Primer. The same parameters for Tm (melting temperature), GC%, and primer size are set with this software as mentioned for Primer3. This software also designs forward and backward primers (*see Note 1*).
4. The various properties namely stem size, loop, dG, Tm, dimer, internal loop, and bulge loop are calculated for all the forward and backward primers using Gene runner software.
5. All the primers satisfying the above-mentioned properties are selected as valid primers.

3.2 In Silico Primer Validation

In silico results should be validated before proceeding in the wet lab. For primer designing, primers obtained by using bioinformatics tools should be validated in silico before commencing the wet lab validation. This is done by using the following steps:

1. Specificity of the primers is validated by carrying out the local alignment against the NCBI's (nr) database through BLAST. The BLAST search is carried out against the nonredundant (nr) database and bacteria specified as the organism.

2. If the primers and the sequence are aligned (please see Note 2), then the primer is considered to be a validated specific primer.
3. Efficiency of a primer is validated by checking all the properties of the primer by using NetPrimer tool. A primer is efficient only when it does not form any secondary structures viz. dimer or hairpin like structure (please see Note 3). To facilitate the selection of an optimal primer, each primer is given a rating based on the stability of its secondary structures.
4. After checking specificity and efficiency of a primer, thereafter, it is validated in the wet lab.

3.3 Wet Lab Validation

3.3.1 DNA Isolation and Purification

The following is a general protocol for leaf tissue [16]. There will be variation in the protocol depending on whether DNA is being isolated from a plant or animal tissue or from bacterial cells.

1. Collect 5 g leaf tissue and grind in liquid nitrogen using chilled pestle and mortar.
2. Transfer the ground powder to a 50 ml sized polypropylene centrifuge tube containing 15 ml of grinding medium maintained at 60 °C and thereafter incubate for 1 h.
3. Emulsify the mixture with an equal volume of chloroform:Isoamylalcohol (24:1) for 5 min by inversion and centrifuge at 11,000 × g for 10 min in a centrifuge.
4. Pipet out aqueous phase into a fresh centrifuge tube and add 0.6 to one volume of isopropanol to it and mix by doing quick and gentle inversion to precipitate DNA.
5. Spool the precipitated DNA using disposable pipet tip and wash twice with 70 % ethanol.
6. Dry the pellet under vacuum using a centrifugal lyophilizer (speed vac) and dissolve in 1 ml TE buffer.

3.3.2 Polymerase Chain Reaction and RAPD Analysis

1. Transfer master mix (24 μl) having genomic DNA in PCR tubes. Add 1 μl of genomic DNA in each tube.

2. Mix the reaction mixture after addition of all the components gently and carry out PCR in a thermal cycler using the following conditions:

STEP 1: Denaturation at 94 °C for 4 min in first denaturation step and thereafter 1 min for every subsequent denaturation steps.

STEP 2: Do annealing at 52 °C for 2 min.

STEP 3: Set extension at 72 °C for 5 min.

3. Prepare the amplified samples by adding 1/6th volume of 6× loading dye (0.25 % bromo-phenol blue and 0.25 % xylene cyanol in 30 % glycerol).

4. Prepare a solution of 1.5 % agarose in 1× TAE buffer (40 mM Tris–HCl, 1 mM EDTA and 20 mM acetic acid, pH 8.3) by heating to boiling.
5. Cool the agarose solution to about 50 °C and thereafter, you may add ethidium bromide to a final concentration of 0.5 µg/ml. After mixing the ethidium bromide properly, pour the solution into the gel tray with comb and place on a horizontally levelled surface.
6. Allow the gel solution to cool up to room temperature and leave for about 15 min for proper gellification.
7. Remove the comb and put the gel in a submerged type electrophoresis chamber filled with 1× TAE buffer.
8. Load the samples in the wells carefully.
9. Connect the chamber to electrophoresis power supply. Generally, the electrophoresis is run using 100 mA current till bromo-phenol blue reaches at another end of the gel.
10. You may also run molecular marker (1 kb ladder) simultaneously in one adjacent well.

After electrophoresis, put the gel over the UV trans-illuminator for observing the fluorescent white bands. Take the required precautions for UV irradiation.

4 Notes

1. Parameters for primer designing include melting temperature (T_m), annealing temperature (T_a), GC contents, dimer formation, false priming, and 3' end sequence. These criteria must be set properly while using any software.
2. For the proper validation of primer, alignment of primer and sequence is considered to be correct only when identity and query coverage both have been considered.
3. Primer must not form dimer (self dimer and cross dimer) or hairpin-like structure.
4. Wet lab validation includes actual verification of proper alignment of the primers with the DNA templates. Therefore, it involves DNA isolation in pure form, annealing, and amplification.
5. Make sure that DNA isolated is free from contaminants.
6. Chloroform:Isoamylalcohol mixture is used in DNA isolation. Iso amyl alcohol allows the chloroform to mix better with the sample.
7. Be careful in pipeting out aqueous phase and do not collect any traces of other separated layer.

8. Ethidium bromide dye must be handled much carefully since it is carcinogenic.
9. Taq DNA polymerase is commonly used thermo-stable DNA polymerase. Nowadays, other thermo-stable DNA polymerases are also available in the market.

Acknowledgements

The authors acknowledge the facilities of the Department of Biotechnology, Ministry of Science and Technology, Government of India, New Delhi (DBT), under the Bioinformatics subcentre in the preparation of this manuscript.

References

1. Garg N, Kumar A (2006) Primer designing for *DREB2A*, a drought resistant gene in *Glycine max*. *J Cell Tissue Res* 6:807–813
2. Garg S, Sohani N, Pundhir S, Kumar A (2007) Primer designing for microbial endo-1-4- β -xylanase gene. *J Cell Tissue Res* 7:1147–1154
3. Garg N, Pundhir S, Prakash A, Kumar A (2008) PCR primer design: DREB genes. *J Comp Sci Syst Biol* 1:021–040
4. Dieffenbach CW, Lowe TM, Dveksler GS (1993) General concepts for PCR primer design. *Genome Res* 3:S30–S37
5. Garg N, Pundhir S, Prakash A, Kumar A (2008) Primer designing for *DREB1A*, a cold induced gene. *J Proteomics Bioinformatics* 1:037–046
6. Abd-Elsalam KA (2003) Bioinformatic tools and guideline for PCR primer design. *Afr J Biotechnol* 2:91–95
7. Singh VK, Kumar A (2000) PCR: software for setting up PCR reactions. *Biotechnol. Softw Internet Rep* 1:276–277
8. Rozen S, Skaletsky H (1999) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365–386
9. Inglis DO, Arnaud MB, Binkley J, Shah P, Skrzypek MS, Wymore F, Binkley G, Miyasato SR, Simison M, Sherlock G (2012) The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata*. *Nucleic Acids Res* 40(Database issue):D667–D674
10. Giegerich R, Meyer F, Schleiermacher C (1996) GeneFisher—software support for the detection of postulated genes. *Proc Intl Conf Intell Syst Mol Biol* 4:68–77
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
12. Doyle JJ, Doyle JL (1990) Isolation of DNA from small amounts of plant tissues. *BRL Focus* 12:13–15
13. Dieffenbach CW, Lowe TMJ, Dveksler GS (1995) General concepts for PCR primer design. In: Dieffenbach CW, Dveksler GS (eds) PCR primer, a laboratory manual. Cold Spring Harbor Laboratory, New York, pp 133–155
14. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2005) GenBank. *Nucleic Acids Res* 33:D34–D38
15. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Higgins DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948
16. Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res* 8:4321–4326

Chapter 11

Primer Design Using Primer Express® for SYBR Green-Based Quantitative PCR

Amarjeet Singh and Girdhar K. Pandey

Abstract

To quantitate the gene expression, real-time RT-PCR or quantitative PCR (qPCR) is one of the most sensitive, reliable, and commonly used methods in molecular biology. The reliability and success of a real-time PCR assay depend on the optimal experiment design. Primers are the most important constituents of real-time PCR experiments such as in SYBR Green-based detection assays. Designing of an appropriate and specific primer pair is extremely crucial for correct estimation of transcript abundance of any gene in a given sample. Here, we are presenting a quick, easy, and reliable method for designing target-specific primers using Primer Express® software for real-time PCR (qPCR) experiments.

Key words Real-time PCR, SYBR Green, Primer, Expression, Primer Express®

1 Introduction

After the emergence of PCR technique in early 1980s to amplify the DNA molecules, it has been extensively used and modified into several different forms to solve the problems in molecular biology. One of the important facets of PCR-based technique is utilized in assessing the gene expression by qualitative or quantitative measure of mRNA in the cell. Real-time quantitative RT-PCR (qPCR) has emerged as a powerful technique to estimate the relative quantitative differences in the transcript level of various samples. This technique is favored over other methods such as northern blotting, ribonuclease protection assays, and semi-quantitative RT-PCR for transcript analysis because lesser amount of RNA, highly reliable quantitative assessment, and lesser efforts to generate significant data in a relatively short period of time [1, 2]. Moreover, its ease of use and high sensitivity make it a desirable method for various applications in molecular biology and diagnostics [3–6]. Numerous laboratories, which perform functional genomics studies, utilize

qPCR method to validate tremendous amount of transcriptomic data generated through high-throughput techniques such as gene chip microarrays [7–14]. Also, during the time of publication, it is highly recommended to validate the expression data by qPCR analysis. Obtaining high-quality and accurate expression data depends on the design of the qPCR experiment. Like a conventional PCR reaction, qPCR reaction also consists of buffer, dNTPs, DNA polymerase, primers, and DNA template. The success of qPCR reaction depends on the selection and use of the best possible combinations of all these components. For the transcript or expression analysis the RNA template is transcribed into cDNA. For specific amplification, DNA and RNA templates should be relatively pure of contaminating proteins and carbohydrates, and their purity can be ascertained by measuring the OD_{260}/OD_{280} and OD_{260}/OD_{230} ratio of the sample using a spectrophotometer. Uncontaminated DNA and RNA template generally have OD_{260}/OD_{280} ratios of 1.8–2.0 and OD_{260}/OD_{230} ratio in the range 2.0–2.3. Genomic DNA contamination should also be removed from RNA samples by using a DNase enzyme and RNA samples should not be degraded. The RNA integrity should be analyzed on a gel or on a bioanalyzer chip. The ratio of 28S to 18S rRNA should be approximately 2.0 in an intact RNA sample. Primers are the most essential and important components of a PCR reaction as they determine the desired specific region on the template and bind there to initiate the polymerization of a DNA sequence. SYBR Green-based detection method is more suitable and feasible for expression analysis of multiple genes together. SYBR Green chemistry is appreciated as the simplest and cheapest chemistry for real-time PCR applications. This dye binds to the minor groove of double-stranded DNA and fluoresces thousands times brighter when bound than in unbound state. This implies that SYBR Green signal increases with the progress in PCR reaction with formation of more double-stranded DNA product. However, one major drawback with the SYBR Green chemistry is its nonspecific binding to any kind of double-stranded DNA and generation of non-specific fluorescent signal. With the detection chemistry as that of SYBR Green, designing of proper and specific primers becomes more important because nonspecific intercalation of dye within primer dimers may produce nonspecific fluorescence and lead to false positive results. Here, we present an easy and reliable method to design specific primers for real-time qPCR experiment using Primer Express® software. We also discuss about consideration of various factors such as range of GC content, melting temperatures, length of primer, amplicon size, primer dimer formation, stem-loop structure generation, for reliable and specific primer design using computational tools.

2 Materials

1. *Computer with Internet:* To begin the process of primer design, a computer system with Internet accessibility is must, as initial reference sequences will be searched, analyzed and downloaded from online available public databases.
2. *Primer Express® software:* Primer Express® is a program from Applied Biosystems and used to design the primers for real-time PCR for SYBR Green®-based assays. This software does not require any specific computer program to run and can be easily installed with a normal configuration with any version of Windows and Macintosh.
3. *Reference nucleotide sequence:* To design real-time PCR primers, a nucleotide sequence is prerequisite to target a specific region for amplification and select suitable primer pair. Generally, a full-length cDNA sequence with 5' and 3' UTR region is suitable as a reference sequence but if full-length cDNA is not available, coding region (ORF) of a gene can also be selected (*see Note 1*).
4. *Sequence analysis tools:* Once the primer(s) is selected based on the fulfillment of primary requirement, it is important to scan it for the specificity. Sequence alignment tool such as BLAST and similar tools can be used to align the primer with selected reference sequence and other similar sequence to ensure that the primer(s) bind only to the desired region on the reference sequence and not to any other nonspecific sequence. BLAST tool available with NCBI (National Centre for Biotechnology Information) is more often used for this analysis but specific databases such as TAIR (The Arabidopsis Information Resource) and RGAP (Rice Genome Annotation Project-TIGR) can also be used for species-specific homology search.

3 Methods

3.1 Retrieval of Reference Nucleotide Sequence

1. For the gene of interest, obtain the database accession ID by keyword search or homology search in the public databases.
2. Search in NCBI or species-specific database for the full-length cDNA or ORF sequence, using identified accession ID. Here, we are using a rice gene as an example for which the full-length cDNA was extracted from KOME (Knowledge-Based Oryza Molecular Encyclopedia) database (<http://cdna01.dna.affrc.go.jp/cDNA>).
3. After retrieval, copy and paste the sequence on a text file (notepad) FASTA format (*see Note 2*). However, the sequence file can also be directly imported as GeneWorks, GenBank sequence, EMBL, GCG, PHYLIP, PRIMER, and ASCII text in Primer Express® software.

3.2 Generation of Primers by Software

1. Open Primer Express® by double-clicking on the software icon in the computer.
2. At the extreme left of the screen, select “File” then “New” in the dropdown menu. Slide to the right and select “TaqMan® Probes and Primer Design” (Fig. 1).
3. A new screen will open and at the sequence tab a blank space will appear where reference sequence should be uploaded either by copy and paste (*see Note 1*) or by clicking on “Import DNA File” button and selecting the text sequence file in FASTA format (Fig. 2).
4. Go to the original window and click “Options” and then from the dropdown menu select “Find Primers/Probes now”.
5. A complete list of 200 primer pairs in version 2.0 and 50 pairs in version 3.0 will be generated. The entire list of primer combinations can either be viewed on the computer screen by clicking on the “Primers” button on sequence tab with all the important parameters such as start and end positions, length, Tm, and % GC content (Fig. 3) or it can be saved by clicking the on “Save List” button at bottom right of the sequence tab screen and it will be automatically imported in text format, to the folder which contains the input reference sequence.
6. This file can be opened with MS excel and list of primers can be analyzed for various important parameters at any time.

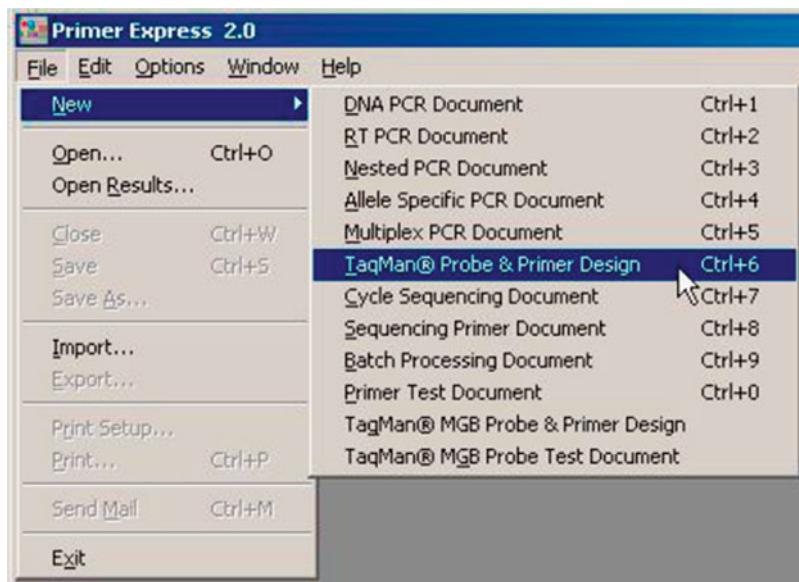


Fig. 1 A snapshot of Primer Express® 2.0 showing the starting screen where primer designing process starts. After selecting ‘File’ then ‘New’ through dropdown menu, slide toward ‘TaqMan® Probe and Primer Design’ and select it to start primer designing

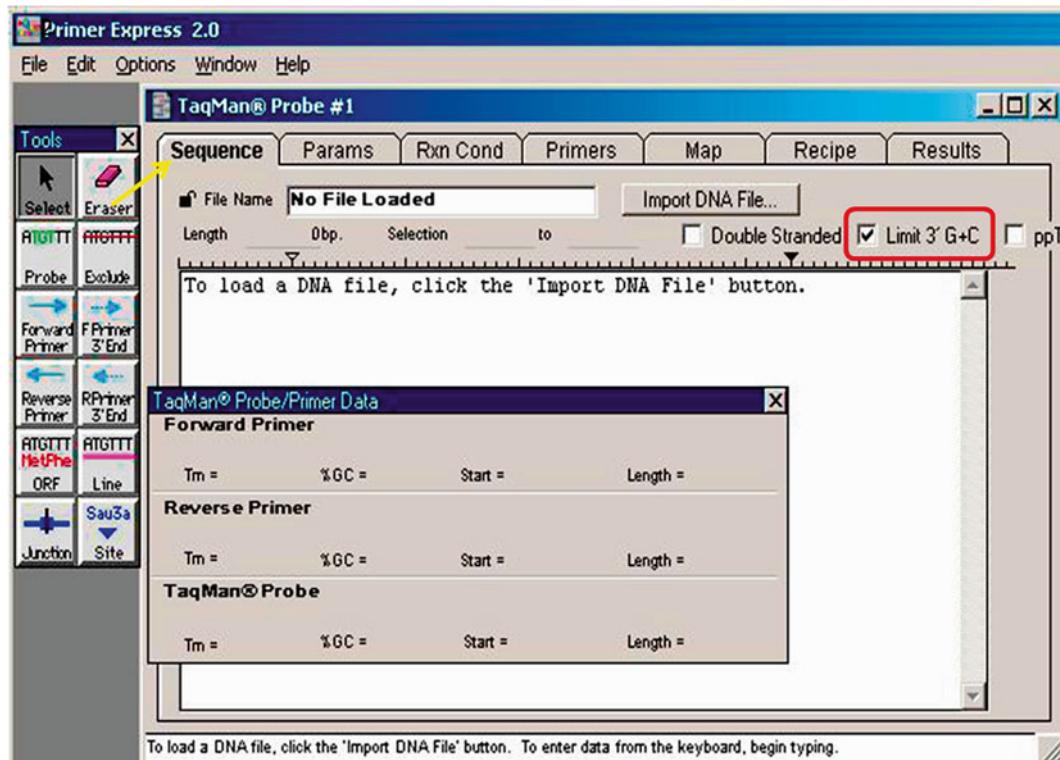


Fig. 2 A snapshot of Primer Express® 2.0 screen showing the space where reference sequence can be imported and the various features and tools used to obtain optimum primer(s)

	Fwd Start	Fwd Stop	Fwd Len.	Fwd Tm	Fwd %GC	Fwd Seq	Rev Start	Rev Stop	Rev Len.	Rev Tm	Rev %GC	Rev Seq	Probe Start	Probe Stop	Probe Len.	Probe Tm
1	9	30	22	58	50	CTCTCGCT	75	55	21	58	57	CCACTACC	32	46	15	69
2	9	30	22	58	50	CTCTCGCT	75	55	21	58	57	CCACTACC	33	49	17	70
3	9	31	23	58	48	CTCTCGCT	75	55	21	58	57	CCACTACC	33	49	17	70
4	8	27	20	58	55	TCTCTCGC	75	55	21	58	57	CCACTACC	29	45	17	68
5	8	27	20	58	55	TCTCTCGC	75	55	21	58	57	CCACTACC	29	45	18	70
6	8	27	20	58	55	TCTCTCGC	75	55	21	58	57	CCACTACC	30	46	17	69
7	8	27	20	58	55	TCTCTCGC	75	55	21	58	57	CCACTACC	31	46	16	69
8	8	27	20	58	55	TCTCTCGC	75	55	21	58	57	CCACTACC	32	46	15	69
9	8	27	20	58	55	TCTCTCGC	75	55	21	58	57	CCACTACC	33	49	17	70
10	8	28	21	58	52	TCTCTCGC	75	55	21	58	57	CCACTACC	30	46	17	69
11	8	28	21	58	52	TCTCTCGC	75	55	21	58	57	CCACTACC	31	46	16	69
12	8	28	21	58	52	TCTCTCGC	75	55	21	58	57	CCACTACC	32	46	15	69
13	8	28	21	58	52	TCTCTCGC	75	55	21	58	57	CCACTACC	33	49	17	70
14	8	29	22	60	60	TCTCTCGC	75	55	21	58	57	CCACTACC	31	46	16	69
15	8	29	22	60	50	TCTCTCGC	75	55	21	58	57	CCACTACC	32	46	15	69
16	9	30	22	58	50	TCTCTCGT	76	56	21	58	52	AACCACTACC	32	46	15	69
17	8	29	22	60	50	TCTCTCGC	75	55	21	58	57	CCACTACC	33	49	17	70
18	9	30	22	58	50	TCTCTCGT	76	56	21	58	52	AACCACTACC	33	49	17	70
19	9	30	22	58	50	TCTCTCGT	76	55	22	59	55	AACCACTACC	32	46	15	69
20	9	30	22	58	50	TCTCTCGT	76	55	22	59	55	AACCACTACC	33	49	17	70
21	9	31	23	58	48	TCTCTCGT	76	56	21	58	52	AACCACTACC	33	49	17	70
22	9	31	23	58	48	TCTCTCGT	76	55	22	59	55	AACCACTACC	33	49	17	70
23	8	27	20	58	55	TCTCTCGC	76	56	21	58	52	AACCACTACC	29	45	17	68
24	8	27	20	58	55	TCTCTCGC	76	56	21	58	52	AACCACTACC	29	46	18	70
25	8	27	20	58	55	TCTCTCGC	76	56	21	58	52	AACCACTACC	30	46	17	69
26	8	27	20	58	55	TCTCTCGC	76	56	21	58	52	AACCACTACC	31	46	16	69
27	8	27	20	58	55	TCTCTCGC	76	56	21	58	52	AACCACTACC	32	46	15	69
28	8	27	20	58	55	TCTCTCGC	76	56	21	58	52	AACCACTACC	33	49	17	70
29	7	27	21	59	52	ATCTCTCG	75	55	21	58	57	CCACTACC	29	45	17	68
30	8	27	20	58	55	TCTCTCGC	76	55	22	59	55	AACCACTACC	29	45	17	68

Fig. 3 A snapshot of the Primer Express® 3.0 showing the list of primer combinations obtained for one of the rice gene after clicking the Primers/Probes tab. For each pair of primers (both forward and reverse) all the important details such as start and end position on the target sequence, primer length, GC content, melting temperature (Tm), and sequence can be obtained in a single file

7. To view and analyze the different parameters of any selected primer(s) in the software itself, click on that primer sequence and a small window will open in Version 2.0 where different parameters such as GC content, Tm, and length are mentioned (Fig. 4a), while in version 3.0 go to ‘Tools’ tab and in dropdown menu select ‘Primer Probe Test Tool’ and then a small window will pop up, which gives these details about both forward and reverse primers and additionally about the secondary structures such as hairpin, self-dimers, and cross-dimers (Fig. 4b).

At step 4, if suitable primers are not found then a pop-up will appear with notice saying no acceptable primer pairs were found (in version 3.0). This may happen when the reference sequence contain a stretch of only one nucleotide repeats at one or both the ends, especially sequences with high GC content such as rice nucleotide sequence and hence both the primers does not fit into the parameters of the software, fixed to get a compatible primer pair. In this case, it is advisable to proceed for primer designing manually.

8. For the optimum efficiency of real-time PCR reaction, primer(s) should fulfill the following conditions:

- (a) Minimum primer length: 18 bases
- (b) Amplicon size: 50–150 bases
- (c) % GC content: 40–70 %
- (d) Tm: 58–60 °C
- (e) At the 3' end, last 5 nucleotides should not contain more than 2 G + C bases (*see Notes 3–6*).
- (f) Low or no self-complementarity to avoid primer dimers

To check all these parameters in Primer Express® use the ‘Primer Test’ tool. From the “File” menu select “New”, slide towards right and choose “Primer Test Document”. Now select the sequence tab and choose forward or reverse primer to test. This information will guide and suggest the users to adjust the default parameters.

To obtain the primer(s) with all these parameters in optimum range, instead of using default settings, adjustment can be done in many filters. For the adjustments click on the ‘params’ tab next to sequence tab (Fig. 5).

- (a) *Optimum annealing temperatures (Tm)*: Although the primers generated with default settings will have comparable Tm, which will work optimally with ABI real-time machine, but this can be modified according to user’s requirement to work with any real-time machine and obtain better results.
- (b) *GC content*: Sometimes if primers are not obtained in the desired range of GC content, the filter can be relaxed or constricted as per the requirement.

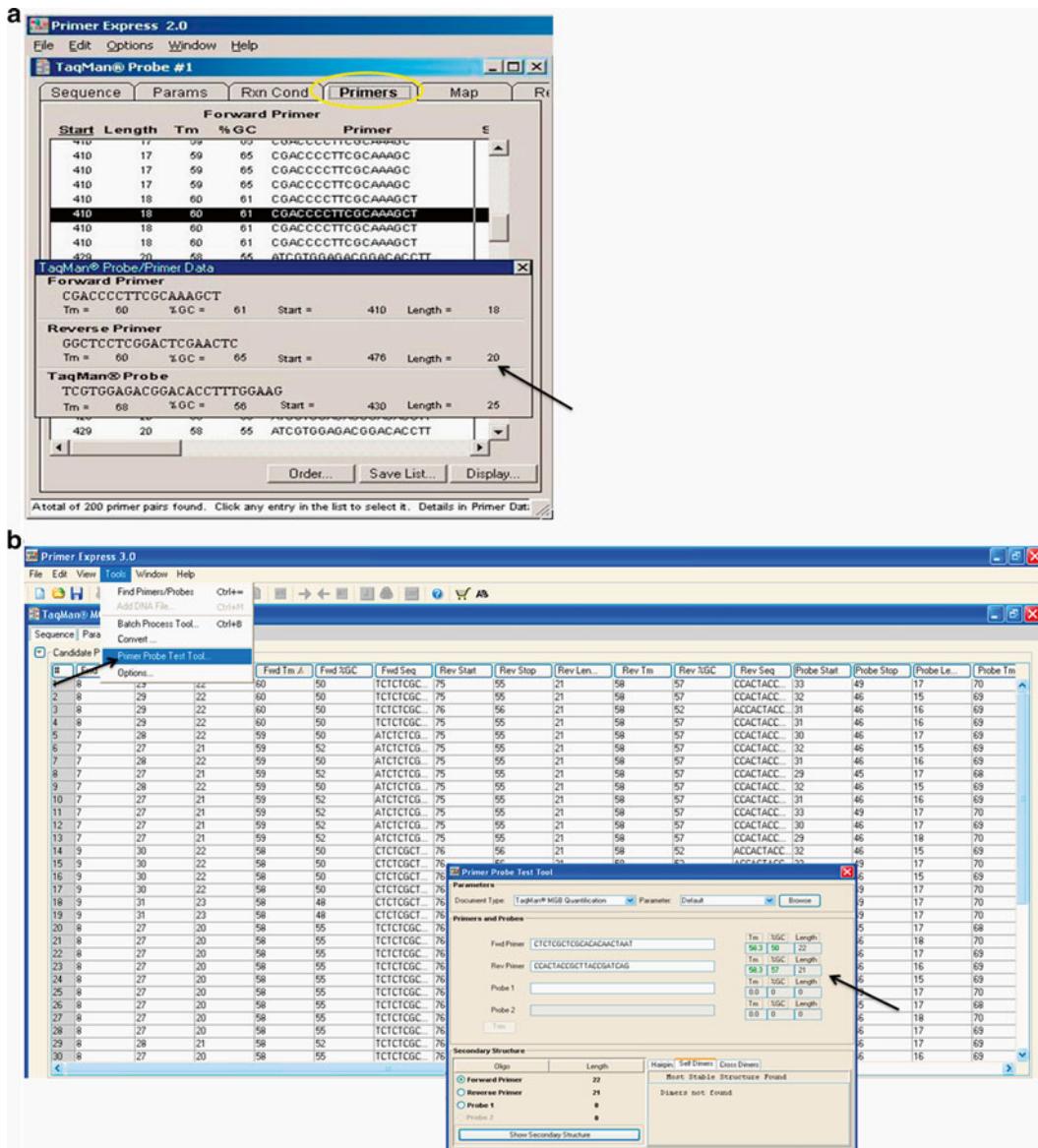


Fig. 4 Snapshots from the computer screens depicting the step to analyze different vital parameters of a primer. (a) In Primer Express® 2.0, details of different parameters can be viewed by clicking on the test primer from the list of primers. A small window appears, which shows Tm, % GC, start site, and length of the primer. (b) In Primer Express® 3.0, from the Tools tab a dropdown menu comes where Primer Probe Test Tool is selected, after that a small window appears with all the important parameters including secondary structures such as hairpin, self-dimers, and cross-dimers

(c) *Primer length:* Primers generated by the software are of standard size for most amplification experiments, but if different primer size is required (sometime bigger length primers are required to increase the specificity), it can be achieved through adjustments.

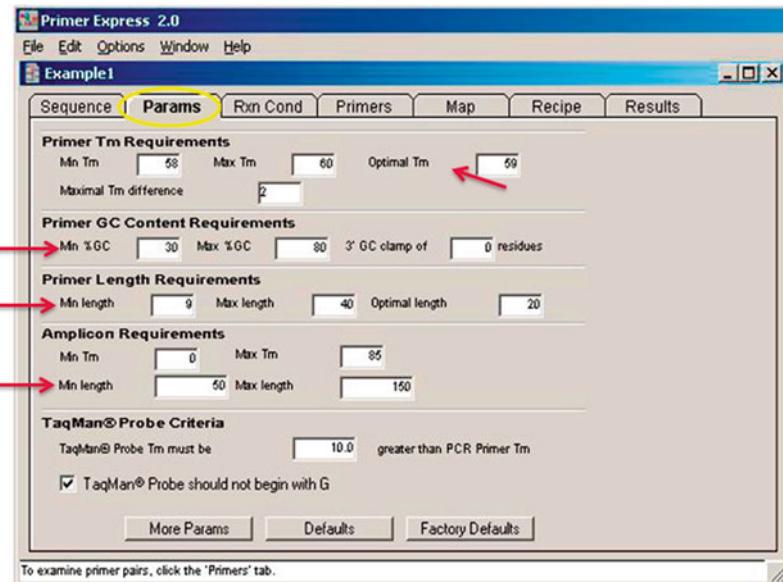


Fig. 5 A Primer Express® 2.0 screenshot showing different options for manual modification of various important primer parameters at the Params tab, to obtain an optimum combination. Red arrows are showing the positions where various primer parameters such as Tm, GC content, Primer length, and amplicon size can be modified according to requirement of user (color figure online)

(d) *Amplicon size*: Desired amplicon length can also be changed. Maximum amplicon size of ~150 bp results in close to 100 % PCR efficiency, and if this length is increased, it may lead to decrease in PCR efficiency. Minimum length of amplicon is rarely lowered and not recommended below 50 bp but it can also be increased to create a larger amplicon to visualize on the gel.

3.3 Analysis for the Primer Specificity

After verifying all the necessary parameters for a compatible primer pair, test the primer pairs for their sequence specificity.

1. From the list of primers, select the primer pair preferably towards the 3' end (especially when the cDNA is synthesized using oligo-dT primer) as 3' end UTR region is considered unique to a nucleotide sequence.
2. Take the primer sequence and perform homology search in the database preferably NCBI using BLAST tool. Here, select the organism genome and choose BLASTN tool for nucleotide alignment. Copy and paste the primer sequence in the blank space, and after choosing highly similar sequences (megablast) click the BLAST button.
3. Analyze the hits obtained from BLAST and make sure that the test primer binds to the reference nucleotide (cDNA) sequence with 100 % coverage and does not bind to cDNA sequence of any other gene of same species with 100 % coverage. If primer binds to any nonspecific cDNA with less than 100 % coverage,

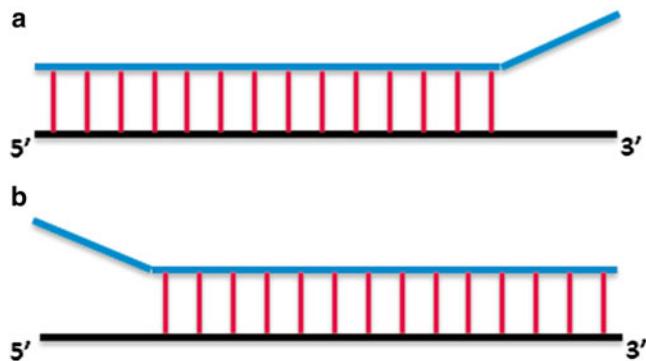


Fig. 6 Depiction of nonspecific binding of the primers to any unwanted sequence. Primer may bind to a nonspecific sequence either completely with 100 % coverage or lesser. When it is less than 100 % coverage, two main possibilities could exist, and they are shown here. (a) Primer may bind to the sequence from 5' end but remain unzipped at 3' end of nonspecific sequence. (b) Primer may not have the complementary bases (2–3 bases) at the 5' end but might find the complementary bases at 3' end, then it will zip at 3'end and amplification of nonspecific sequence will be propagated

it can be used provided it should not zip toward 3'end of nonspecific target sequence (mismatch of 2–3 bases) so that DNA polymerase halt there and does not propagate to produce nonspecific amplicon (Fig. 6a). On the other hand if the primer remains unzipped at 5' end but can bind complementary at the 3'end of nontarget sequence, amplification of off target can be propagated (Fig. 6b).

4. Primer specificity can also be checked by dissociation curve analysis after real-time PCR run, on the real-time PCR instrument itself. Ideally, a dissociation curve should contain a single peak, which indicates the specific amplicon generation. If primers bind nonspecifically then multiple peaks could be observed on the dissociation curve. As a case study, we have performed the dissociation curve analysis for one of the rice gene with different samples of cDNA. Two primer combinations were selected for this analysis, one pair with complete specificity (100 % coverage) to the target sequence and other pair with 70 % coverage and it also binds to the off targets. qPCR amplification with nonspecific primer combination showed multiple peaks in different cDNA samples after dissociation curve analysis (Fig. 7a). While amplification from specific primers showed a single peak in all the samples (Fig. 7b). Analysis of expression data showed that this gene has higher expression values with nonspecific primer combination than specific primers. This higher expression is mainly due to the detection of additional fluorescence of SYBR Green from nonspecifically amplified products and hence led to false positive results.

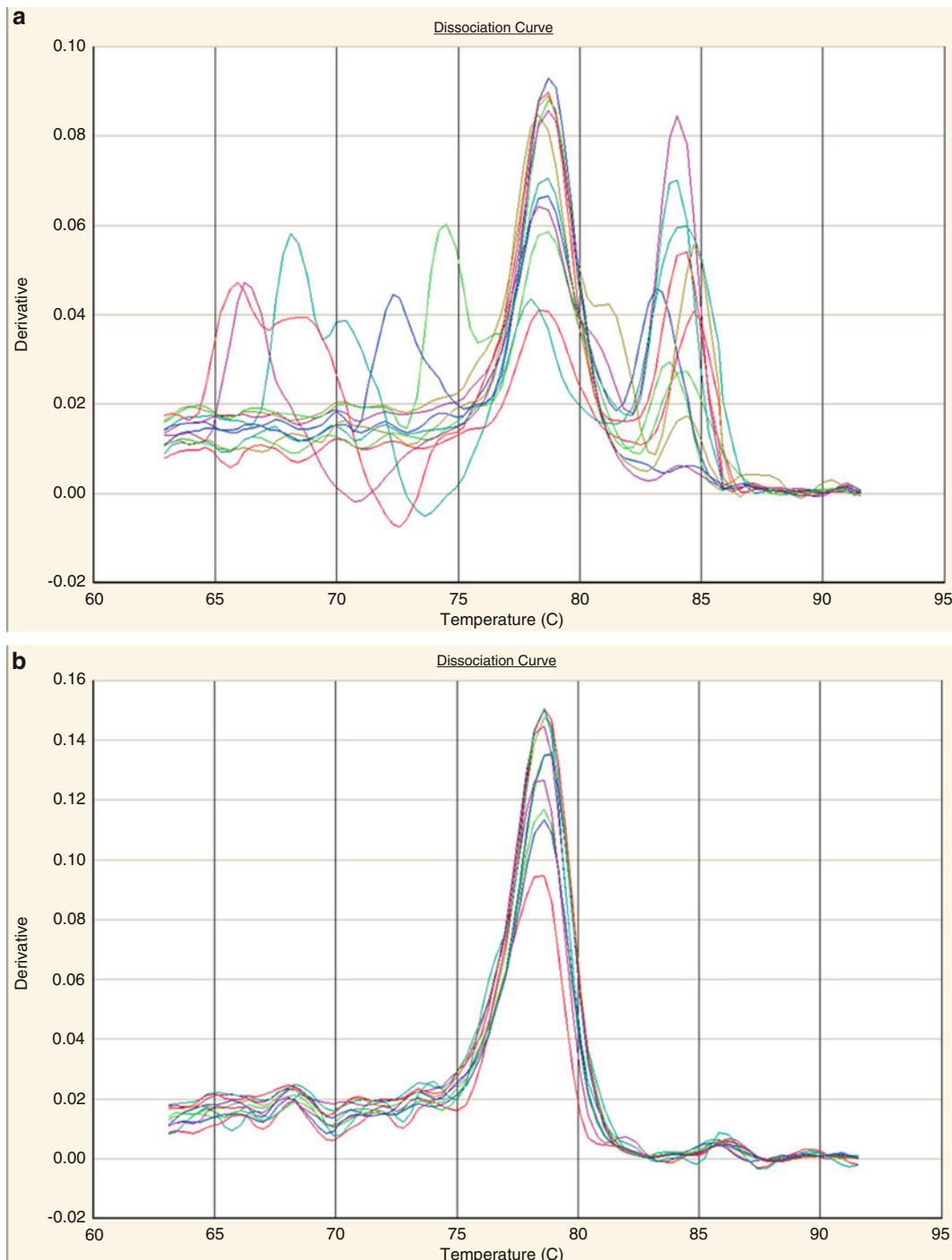


Fig. 7 Dissociation curve analysis performed after qPCR reaction to show primer specificity. During qPCR analysis as a case study in rice, (a) one of the genes showed multiple peaks in different cDNA samples depicting nonspecific binding to template whereas, when the primers were changed to be more specific (b) then a single specific peak was observed in most of the samples at a particular annealing temperature

4 Notes

1. When using Primer Express®, it is better to start primer designing with about 500 bp of reference nucleotide sequence to obtain more specific and efficient primers.
2. When importing a reference sequence file to Primer Express® software, make sure that the sequence is in a tab-delimited text format and does not include any extraneous information such as detail description of the sequence. This may lead to nonrecognition of the sequence by the software.
3. Repeats, e.g., AGAGAGAG and runs of a nucleotide, e.g., GTTTTTTCG should be avoided in the sequence as they lead to the mispriming on the template.
4. When searching for the primers (at **step 3** in generation of primers by software, methods), make sure that “Limit 3’G + C” box is selected, otherwise primer generated without this limitation will contain more than 2 G + C in the last 5 bases at 3’end.
5. For relative expression analysis using SYBR Green for several genes, it is advisable to keep the length of the amplicon very close because larger size product will produce more fluorescence, hence the expression values obtained will not reflect the actual relative expression of multiple genes.
6. It is beneficial to design a primer crossing intron/exon boundary, as it will lead to amplification of specific cDNA sample and not the genomic DNA amplification. In such cases, DNase treatment of RNA samples can be avoided.

Acknowledgement

Research work in GKP’s lab is partially supported by grants from University of Delhi, Department of Biotechnology (DBT), Department of Science and Technology (DST), and Council of Scientific and Industrial Research (CSIR), India. AS acknowledges CSIR for research fellowship.

References

1. Sambrook J, Russell DW (2001) Molecular cloning: a laboratory manual, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
2. Bustin SA (2000) Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol* 25:169–193
3. Thornton B, Basu C (2010) Real-time PCR (qPCR) primer design using free online software. *Biochem Mol Biol Educ* 39: 145–154
4. Ginzinger DG (2002) Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. *Exp Hematol* 30:503–512

5. Belgrader P, Bennett W, Hadley D et al (1999) PCR detection of bacteria in seven minutes. *Science* 284:449–450
6. Johnson MP, Haupt LM, Griffiths LR (2004) Locked nucleic acid (LNA) single nucleotide polymorphism (SNP) genotype analysis and validation using real-time PCR. *Nucleic Acids Res* 32:e55
7. Singh A, Giri J, Kapoor S et al (2010) Protein phosphatase complement in rice: genome-wide identification and transcriptional analysis under abiotic stress conditions and reproductive development. *BMC Genomics* 11:435
8. Singh A, Baranwal V, Shankar A et al (2012) Rice phospholipase A superfamily: organization, phylogenetic and expression analysis during abiotic stresses and development. *PLoS One* 7:e30947
9. Singh A, Pandey A, Baranwal V et al (2012) Comprehensive expression analysis of rice phospholipase D gene family during abiotic stresses and development. *Plant Signal Behav* 7:847–855
10. Singh A, Kanwar P, Pandey A et al (2013) Comprehensive genomic analysis and expression profiling of phospholipase C gene family during abiotic stresses and development in rice. *PLoS One* 8:e62494
11. Singh A, Kanwar P, Yadav AK et al (2014) Genome-wide expressional and functional analysis of calcium transport elements during abiotic stress and development in rice. *FEBS J* 281:894–915
12. Sharma R, Agarwal P, Ray S et al (2012) Expression dynamics of metabolic and regulatory components across stages of panicle and seed development in indica rice. *Funct Integr Genomics* 12:229–248
13. Trevisan S, Manoli A, Begheldo M et al (2011) Transcriptome analysis reveals coordinated spatiotemporal regulation of hemoglobin and nitrate reductase in response to nitrate in maize roots. *New Phytol* 192:338–352
14. Xu J, Sun J, Du L et al (2012) Comparative transcriptome analysis of cadmium responses in *Solanum nigrum* and *Solanum torvum*. *New Phytol* 196:110–124

Chapter 12

Designing Primers for SNaPshot Technique

Greiciane Gaburro Paneto and Francisco de Paula Careta

Abstract

The SNaPshot technique, also known as minisequencing, is a primer extension-based method developed for the analysis of Single Nucleotide Polymorphisms (SNPs). Using this technique, it is possible to analyze more than 50 SNPs distributed throughout the genome in a single multiplex reaction, making it an advantage when compared with traditional sequencing reaction. In this chapter, you will find a step-by-step guide to design a multiplex primer assay for SNaPshot reaction.

Key words SNP, Genotyping, Multiplex, SNaPshot, Primer

1 Introduction

Single Nucleotide Polymorphisms (SNPs) are the most abundant class of human polymorphisms occurring at an average frequency of approximately 1 per 1,000 bp in the genome. SNPs can be defined as single-base inheritable variations in a given and defined genetic location that occur in at least 1 % of the population [1, 2].

To date, more than 61 million SNPs were reported in public databases [3]. This abundance makes SNPs useful markers for genetic association studies that strive, by means of the statistical association of neighboring alleles or linkage disequilibrium, to localize the genes involved in disease susceptibility or adverse drug reactions [4]. Also, SNPs are useful in evolutionary studies and forensic genetics analysis [5–7]. Therefore, tools to routinely analyze a growing number of SNPs are of great importance nowadays.

SNaPshot technique, also known as minisequencing, is a primer extension-based method developed for the analysis of SNPs. For this, first, SNPs are initially amplified in a conventional multiplex PCR, using specific primers flanking its location. Later,

SNaPshot reaction is performed using allele-specific primers annealing one base before the SNP you want to analyze and extended with dideoxynucleotides (ddNTPs) labeled with fluorescence. SNPs are, then, detected in capillary electrophoresis. This technique allows the detection of multiple SNPs simultaneously in a single multiplex assay. The reaction with the largest number of SNPs developed allows the analysis of 52 SNPs in multiplex [8]. Several SNaPshot multiplex assays are described in the literature [9–11].

Though, using SNaPshot technology is mandatory to design and test multiplex primers, both to conventional PCR and to SNaPshot reaction. Primers can also be designed by hand or using specific software. For this, there are general rules that must to be followed. Primer lengths determine the specificity and significantly affect their annealing to the template. Too short primers produce low PCR specificity, resulting in nonspecific amplification and too long primers decrease the template-binding efficiency at normal annealing temperature due to the higher probability of forming secondary structures such as hairpins. The primer length is usually between 18 and 30 bases, with ideal GC content of 40–60 % and the optimal melting temperature between 52 °C and 60 °C, for most of the PCR reactions. Amplicon length should be between 50 b and 3 kb. However, exceptions of those rules can be also found in some published articles.

In this chapter, you will find a step-by-step guide to build your multiplex primer assay for SNaPshot reaction using PerlPrimer and AutoDimer software.

2 Materials

Many different primer design software are commercially available. A good example, explained in this chapter, is PerlPrimer Software. PerlPrimer is a free, open-source GUI application written in Perl that designs primers for standard PCR, bisulphite PCR, real-time PCR, and sequencing. It has the ability to calculate possible primer dimers; retrieve genomic or cDNA sequences from Ensembl; BLAST primers using the NCBI server or a local server, among others. The program is designed to be cross-platform compatible and has been developed and tested on both Microsoft Windows and GNU/Linux-based operating systems. Users have also reported success using the program under Mac OS X [12].

Another software, explained in this chapter, is called AutoDimer. This software was developed to rapidly screen previously selected PCR primers for primer dimer and hairpin interactions in short DNA oligomers (<30 nucleotides). AutoDimer was originally

created to assist in the development of multiplex PCR assays for probing STR and SNP markers for forensic purposes [13].

Both software complement each other to generate the best condition for the assay.

3 Methods

When working with SNaPshot reactions, it is important to know that it is necessary to design primers for both, PCR reaction and SNaPshot reaction, independently. PCR primers can be designed either by hand or using software (as Perlprimer), but SNaPshot primers need to be designed only by hand (using software just to confirm annealing temperatures for the multiplex).

3.1 Perl Primer

Software: Building PCR Primers

1. Before start, create, or download from a database, a fasta file (or .txt file) flanking DNA sequence that you need to analyze (flanking the SNP);
2. Open Perl Primer software;
3. Select the option **STANDARD PCR**;
4. In **Sequence** box, click on **Open DNA Sequence**  , select your fasta or .txt file to be opened;
5. Use default values for **Primer Tm** and **Primer Length** (see Note 1);
6. In **Amplified range** indicate nucleotides range to the program build your forward and reverse primers;
7. Then, click on **Find Primers** (Fig. 1);
8. The program will show you, as result, many primer pair options, you can save this result if you want. If you double-click a primer pair, it will show you all parameters about those primers;
9. Choose the best option taking a look in **Dimers** box: structure with values below -2.00 kcal/mol and, if possible, without **extensible primer dimer** (Fig. 2);
10. After selecting the best primer pair for your assay, go back to the main menu, generate report, and save.
11. Before continuing, you need to confirm the specificity of the primer chosen using BLAST. For this, after double-clicking a primer pair, the option **BLAST Primers** will appear in the left corner of the window. Click on it and check if the primer pair selected is specific to the desired organism.
12. Repeat this protocol for each SNP you want to analyze.

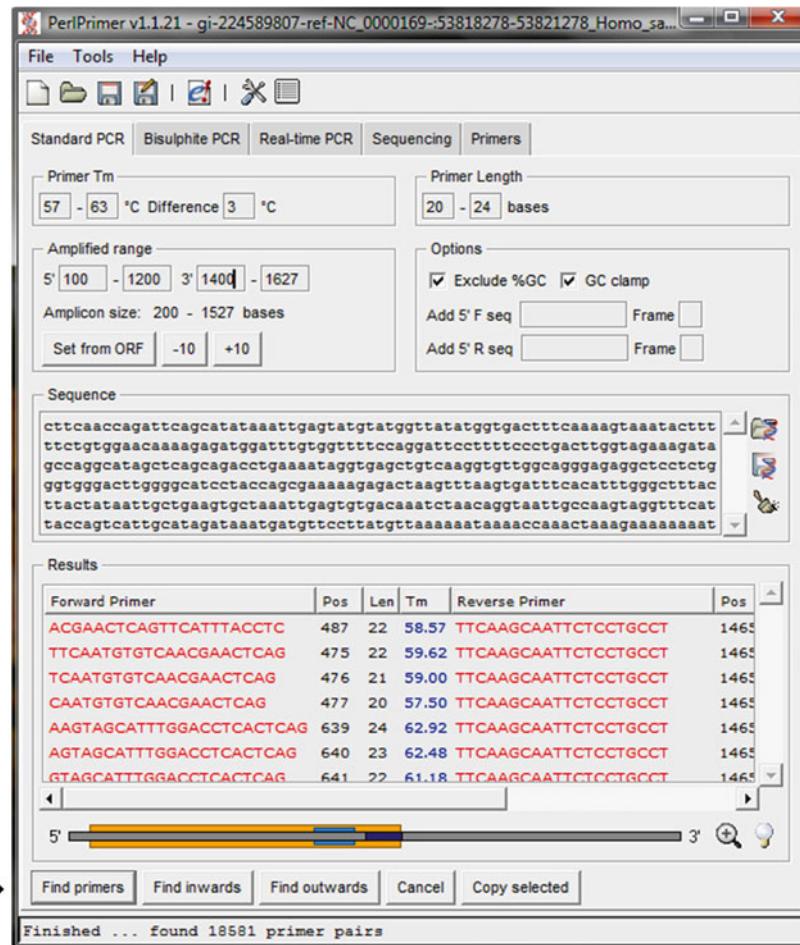


Fig. 1 PerlPrimer software layout. After filling out all the fields, click on the **Find primers** button (indicated with black arrow), and the program will generate several primer pair options

3.2 Building Snapshot Primers: By Hand

1. Created or downloaded from a database, the **double-stranded** DNA sequence (flanking the SNP you need to analyze) to design SNaPshot primers;
2. Localize the SNP in the sequence;
3. Start to build SNaPshot primer one base before or one base after SNP (you can create forward or reverse primers) (Fig. 3);
4. Primer length will be according to the annealing temperature desired (*see Note 1*); however, a noncomplementary nucleotide tail should be included in the 5' end to allow electrophoretic separation of the fragments (*see Note 2*);
5. You must check the primer annealing temperature using a software (Perl Primer, for example);
6. Select the best SNaPshot primer for each SNP you want to analyze.

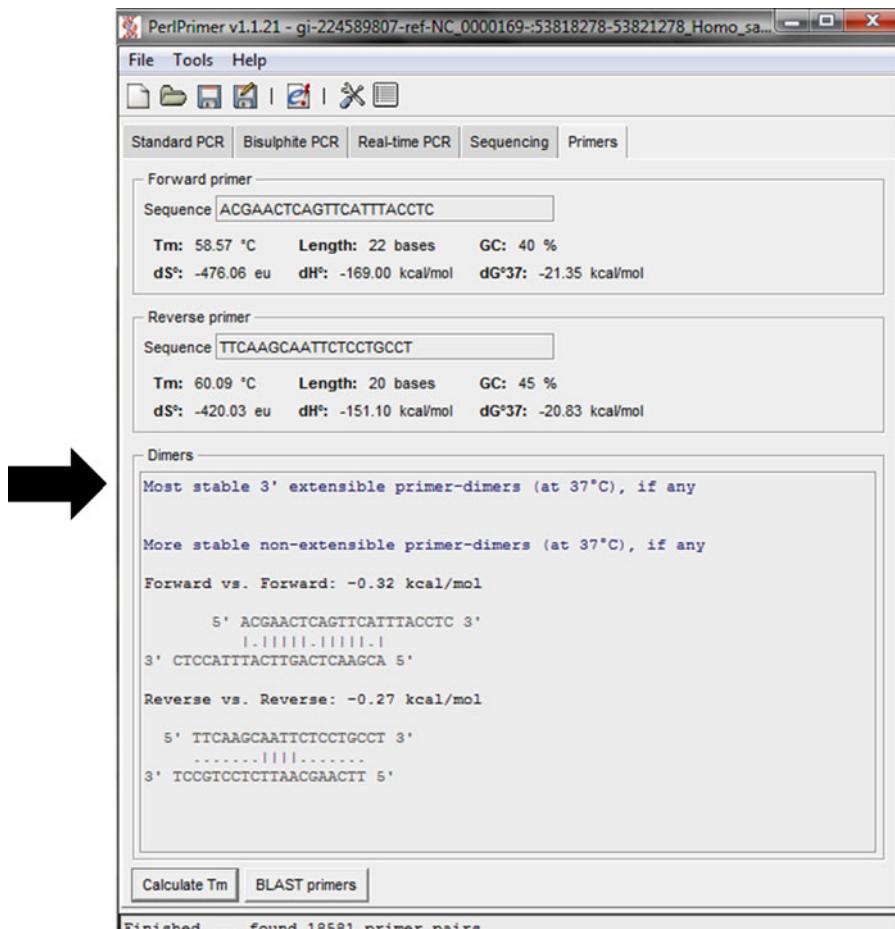


Fig. 2 Dimers box shown in PerlPrimer software after analysis of a selected primer pair

6670	6680	6690	6700	6710	6720
TAATCTCCA	TATTGTA	ACT TACTACTCCG	GAAAAAAAAGA	ACC ATTTGGA	TACATAGGTA
ATTAGAGGGT	ATAACATTGA	ATGATGAGGC	CTTTTTTCT	TGGTAAACCT	ATGTATCCAT
6730	6740	6750	FWD 6760	6770	SNP 6780
TGGTCTGAGC	TATGATATCA	ATTGGCTTCC	TAGGGTTTAT	CGTGTGAGCA	CACCA <u>TATAT</u>
ACCAGACTCG	ATACTATAGT	TAACCGAAGG	ATCCCAAATA	GCACACTCGT	GTGGTATATA
6790	6800	6810	6820	6830	6840
TTACAGTAGG	AATAGACGTA	GACACACGAG	CATATTCAC	CTCCGCTACC	ATAATCATCG
AATGTCATCC	TTATCTGCAT	<u>CTGTGTGCTC</u>	<u>GTATAAAGTG</u>	GAGGCGATGG	TATTAGTAGC
REV					
6850	6860	6870	6880	6890	6900
CTATCCCCAC	CGGCGTCAAA	GTATTTAGCT	GA CTCGCCAC	ACTCCACGGA	AGCAATATGA
GATAGGGGTG	GCCCGAGTTT	CATAAAATCGA	CTGAGCGGTG	TGAGGTGCCT	TCGTTATACT

Fig. 3 Example of a double-stranded DNA sequence used to design SNaPshot primers. SNP marked in red. PCR primers are underlined (fwd → forward and rev → reverse). SNaPshot primers are featured. Numbers indicate nucleotide position in DNA sequence (color figure online)

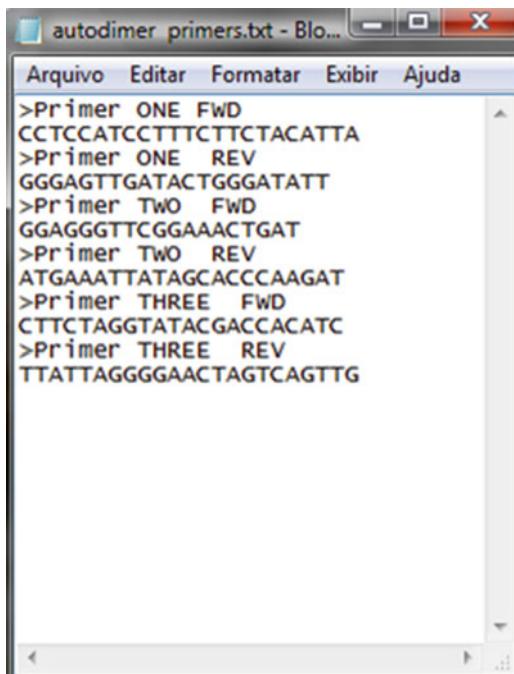


Fig. 4 Example of a .txt file in Notepad with all primers chosen for multiplex. This file must to be created to run on Autodimer software

3.3 Autodimer Software: Analyzing Interactions Among Primers in the Multiplex

1. Create a .txt file with all primers chosen for multiplex (example in Fig. 4) (*see Note 3*);
2. Open Autodimer software;
3. Click **File – Open** – select your .txt file with multiplex primers (for PCR or SNaPshot);
4. Use SCORE 7 or superior (*see Note 4*);
5. Now, click on **Primer Dimer Screen** and then **Hairpin Screen**;
6. Check for primer dimer above 50 °C;
7. Best primers have no primer dimer and no hairpin (Fig. 5), if it is not possible, design new primers;
8. If it is also not possible to design new primers, be sure that primer dimer does not occur in 3'.

4 Notes

1. All primers designed to be used in multiplex must have similar annealing temperatures to work correctly.
2. A noncomplementary tail (example, poly(T)) should be included in 5' of each SNaPshot primer since SNaPshot reactions do not produce fragments of different length (required for electrophoresis separation). It is desirable that each tail have, at least, 5 bp of difference in length for each SNP.

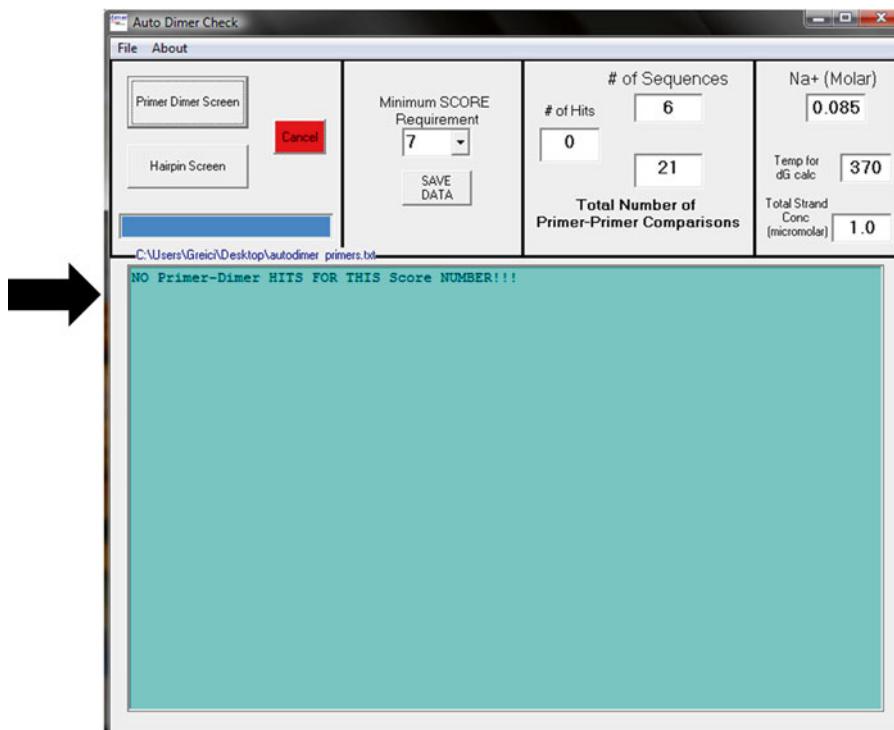


Fig. 5 Autodimer software layout. Black arrow indicates the box with primer analysis result (no primer dimer and no hairpin found in the example)

3. Autodimer need to be used to check PCR and SNaPshot primers, independently.
4. The score reflects the general stability or tendency of the potential interaction to exist in solution. A score threshold of 7 or 8 works well when designing multiplex PCR primers.

References

1. Sachidanandam R, Weissman D, Schmidt SC et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933
2. Taillon-Miller P, Gu Z, Li Q et al (1998) Overlapping genomic sequences: a treasure trove of single nucleotide polymorphisms. *Genome Res* 8:748–754
3. Smigelski E, Sirotnik K, Ward M et al (1999) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 28:352–355
4. Ramos E, Doumatey A, Elkahloun AG et al (2014) Pharmacogenomics, ancestry and clinical decision making for global populations. *Pharmacogenomics J* 14:217–222
5. Stonking M, Hedgecock D, Higuchi RG et al (1991) Populations variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *Am J Hum Genet* 48:370–382
6. Chemale G, Paneto GG, Menezes MA et al (2013) Development and validation of a D-loop mtDNA SNP assay for the screening of specimens in forensic casework. *Forensic Sci Int Genet* 7:353–358
7. Kidd KK, Speed WC, Pakstis AJ et al (2014) Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int Genet* 10:23–32
8. Borsting C, Sanchez JJ, Hansen HE et al (2008) Performance of the SNPforID 52 SNP-plex assay in paternity testing. *Forensic Sci Int Genet* 2:292–300
9. Magnin S, Viel E, Baraquin A et al (2011) A multiplex SNaPshot assay as a rapid method

- for detecting KRAS and BRAF mutations in advanced colorectal cancers. *J Mol Diagn* 13:485–492
10. Coutinho A, Valverde G, Fehren-Schmitz L et al (2014) AmericaPlex26: a SNaPshot multiplex system for genotyping the main human mitochondrial founder lineages of the Americas. *PLoS One* 26:e93292
11. Yang L, Sun H, Chen D et al (2014) Application of multiplex SNaPshot assay in measurement of PLAC4 RNA-SNP allelic ratio for noninvasive prenatal detection of trisomy 21. *Prenat Diagn* 34:139–144
12. Marshall OJ (2004) PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics* 20:2471–2472
13. Vallone PM, Butler JM (2004) AutoDimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques* 37:226–231

Chapter 13

Rapid and Simple Method of qPCR Primer Design

Brenda Thornton and Chhandak Basu

Abstract

Quantitative real-time polymerase chain reaction (qPCR) is a powerful tool for analysis and quantification of gene expression. It is advantageous compared to traditional gel-based method of PCR, as gene expression can be visualized “real-time” using a computer. In qPCR, a reporter dye system is used which intercalates with DNA’s region of interest and detects DNA amplification. Some of the popular reporter systems used in qPCR are the following: Molecular Beacon®, SYBR Green®, and Taqman®. However, success of qPCR depends on the optimal primers used. Some of the considerations for primer design are the following: GC content, primer self-dimer, or secondary structure formation. Freely available software could be used for ideal qPCR primer design. Here we have shown how to use some freely available web-based software programs (such as Primerquest®, Unafold®, and Beacon designer®) to design qPCR primers.

Key words Quantitative real-time polymerase chain reaction, qPCR, Real-time PCR, Primer design, Free online software, SYBR Green primers

1 Introduction

Quantitative real-time polymerase chain reaction (qPCR) is an analytical tool used to measure the expression level of a target DNA sequence, or specific nucleotide sequence of a gene of interest (GOI). If performed correctly, it can be a highly reliable, robust method for estimating the relative expression ratio of gene expression [1–4]. It is fundamentally different from qualitative PCR, in that it allows the experimenter not only to determine whether the GOI is expressed under certain conditions but also to calculate the expression of GOI level relative to a controlled sample, in most cases a housekeeping gene.

The qPCR method requires the use of fluorescent markers, in addition to the normal PCR reagents (*Taq polymerase*, primers, dNTPs, and buffering reagents). These are chemical markers which can be in the form of oligonucleotide probes (i.e. TaqMan) that bind to a specific nucleotide sequence or dyes that intercalate any double-stranded DNA (i.e. SYBR® Green) appearing during the

PCR cycles. With probe-based assays such as TaqMan, emission of a fluorescent signal occurs only when the probe interacts with the target-specific nucleotide sequence. SYBR® Green, on the other hand, emits a fluorescent signal when the dye comes into contact with any double-stranded nucleotide sequence. Since the fluorescent dye will intercalate into any double-stranded nucleotide sequence, there is always concern of detecting false signals that are produced by the presence of primer dimers, hairpin loops, or amplification of nonspecific products. However, research has shown that when proper qPCR protocol is followed, and primers are carefully designed, SYBR® Green-based detection assays accomplish the same high-performance results as that of TaqMan, but with less setup and running cost [5–8]. The lower cost of using SYBR® Green-based detection assays makes them more appealing to many laboratories.

Designing effective primers for SYBR® Green-based assays does not have to involve purchasing expensive software. There are powerful, free, web-based software programs that can create and test primers for qPCR [9]. For many years, Primer 3 online software was the tool of choice [9, 10]. It has since been replaced by a more advanced version, called Primer3Plus. This web-based program contains multiple option settings making it an optimal choice for design of primers, but could be challenging for the novice user. Another, user-friendly, online primer design software for qPCR primers is PrimerQuest®. This software allows users to upload nucleotide sequences directly from NCBI using only the accession number, and can be used to create qualitative, probe-based, and intercalating dye-based primers.

Once primers are created, both forward and reverse primers and their predicted amplicon should be checked for secondary structure formation. In SYBR® Green-based qPCR the presence of primer dimers, hairpin loop formation, cross dimers, and nonspecific target sequences can skew results. Detecting potential secondary structures can also be accomplished using free software such as NCBI's BLAST, UNAFold, or Beacon Designer, Free Edition. This chapter explains how to effectively design SYBR® Green-based qPCR primers using free, online PrimerQuest® software, and how to check both the amplicon and primers for secondary structure formation.

2 Materials

Personal or laptop computer with Internet access, NCBI Genbank accession number JQ240295 or any other nucleotide sequence of interest.

3 Methods

Proper design of qPCR primers is important when using SYBR® Green-based assays. Since the fluorescent dye intercalates into double-stranded DNA formation of primer dimers, trace DNA, hairpin loops, or nonspecific amplicons may increase the fluorescent signal, resulting in inaccurate quantification of the GOI. Knowledge of the common parameters used for qPCR primers is therefore essential. The method below explains the procedure for using free, online PrimerQuest® to design SYBR® Green-based qPCR primers. Since the parameters for making primers for use with SYBR® Green dyes are generalized for use with any qPCR, this method for primer design can be used with any online primer design software. In this protocol, the nucleotide sequence for *Pinus contorta* (+)-alpha pinene synthase, NCBI Genbank accession number JQ240295.1 is used; however, any nucleotide sequence could be used in its place. The methods are given in a step-wise manner, and when appropriate, denote optimal parameters.

3.1 Accessing Website

Using a personal computer, or laptop, with Internet access, go to the PrimerQuest® website at <http://www.idtdna.com/Primerquest/Home/Index>.

3.2 Entering Sequence

Nucleotide sequence information can be entered by one of three methods: entering sequence manually, downloading sequence using Genbank or Accession ID, or uploading the nucleotide sequence in Microsoft Excel format. The choice of options depends upon the user's resources. If primers are needed for a nucleotide sequence that is not in Genbank, the user should copy and paste the sequence using the manual or Excel file option. If an NCBI's Genbank sequence exists, the user only needs to know the accession number for that sequence. We used the option for "Download sequence(s) using Genbank or Accession ID".

Select the dropdown menu, for "Download sequence(s) using Genbank or Accession ID" and enter JQ240295.1 into the input box next to "NCBI ID#", then click on "Get Sequence". The nucleotide sequence is automatically retrieved from NCBI's Genbank and placed in the sequence box.

3.3 Primer Design

Under the subtitle "Choose Your Design" users have a choice of four selections: (1) PCR, 2 primers; (2) qPCR, 2 primers+probe; (3) qPCR, 2 primers intercalating dyes; and (4) show custom design parameters. Option 1 should be used for designing qualitative PCR primers. Option 2 is used to create primers for qPCR probe-based assays, such as TaqMan PCR. Option 3 is used to design intercalating dye-based qPCR primers such as SYBR® Green, and Option 4 is used to customize primer parameters.

To design SYBR® Green qPCR primers it is important to consider the primer melting temperature, GC content, number of GC clamps, and the amplicon length. Optimal melting temperatures for primers used in qPCR should be between 50 °C and 60 °C. The GC content of the primers should be between 50 % and 60 %. Optimal GC clamp number is 1 with the maximum number of GC clamps not exceeding 2. Optimal amplicon length should be between 75 bp and 150 bp.

Select “Customize Parameters,” and change the following under “Primer Criteria”: Next to Primer Tm (°C): change the minimum melting temperature from 59 to 50 and optimum from 62 to 60. Next to Primer GC (%): change the minimum from 35 to 50, the optimum from 50 to 55, and the maximum from 65 to 60. Next to Primer size (nt): change the optimum from 22 to 20 and the maximum from 30 to 24. Under the “Misc Settings” subtitle, change the value for 3' GC Clamp (nt) from 0 to 1.

The optimal amplicon length in a qPCR is between 75 bp and 150 bp. Under the subtitle “Amplicon Criteria”, set the minimum amplicon size to 75, the optimal size to 100 and the maximum size to 150. Then, press the “Get Assays” button at the bottom of the page. Results are calculated automatically and appear in a new window. Each primer set shows the amplicon length, and the primer forward and start and stop position, length, melting point, and GC% content.

The first three primer results of the assay are as follows (Fig. 1):
Set 1 ACC# JQ240295.1—Forward Primer=GACCACCTCC

PrimerQuest

Assay Design	Results	Help	About								
ACC# JQ240295.1 Assay Set 1 Details											
Back to Results											
Parameter Set:	General PCR (Primers only)										
Sequence Name:	ACC# JQ240295.1										
Amplicon Length:	98										
Forward	GACCACCTCCGTCTTAATG (Sense)	Start	216	Stop	236	Length	20	Tm	60	GC%	50
Reverse	AAGTGGAGAGAGGGCTATG (AntiSense)		294		314		20		60		50
Base	Sequence										
1	GTTGTCTTGAGCCGTCTATGAGGCCAACGATCATATAACATATAAGCCACAGTAAATGGCTCTGGTTCTGCTGTCCCCTGTAATTCCAACGTGTGC										
101	TGTGCRGAACGTTGTCGGTTTAGTCATGAGCTGAAAGCTATCCATAGTACAGTCCCCAATCTTGGAAATGTCRAGGGGAGGGAAATCCATAGCRCCTTC										
201	TATGAGCATGAGTC GACCACCTCCGTCTTAATG AGGATGGGTACCAAGACGCATAGCTGGTCATCATTCACCCACCTTGGACGATGATTC CATAGCC										
301	TCTCTCTCCTTCACTT CCTATGAGGCCACCTCTTACCGTGAAGGCCTCTATAGAGTTATAGGGAAAGTAAAAAATATCTCGATTAAATGTCAGTGGGGATG										
401	GAGTATCACCAAGTCCCCCTCACTGACCTCATACCGCCCTCTGGATGGTCGATACGCTGTAACGGTGGGATCGATAGGCATTCAACACGAGATAAA										
501	TTCTGCTCTGGATCATGTTAACAGTTATGGACCGAAAAAGGCATTGGACGTGGCAGAGGAAGTGGTGTGACTGATCTCAACTCAACTGCTGGGTCT										
601	CGAACTCTCGACTACACGSGATAACACAGTGTCTTCACATGTTCTGATCACTTCACAAAACGAGAAGGGCGACTTGTGCTGCATTCAACACAGAAG										
701	GAGAGATAAGAGATGTCCTGAAATTATTCGGGCATCTCATGCCTTCTGGCGAAAAAAATTATGGAGGCGGCTGAAATCTCTACAATGTATTT										
801	AAAAGATGCCCTACAAAAGATTCGGCCCTCCGGCTTTACAAAGAGATAGAGTATCTTCTGAAATTGGTTGTCACAGRACTGGCAAGGATGSAARCA										
901	AGAATGTACATGACGTATTGGAGAGGAGACATACATTGAGACGCCATATTGATCAGGGAGAAACTGTTAGAGCTTGCAAAACTGGAATTCAACATTT										
1001	TTCACTCTCTAGTAAAGAGAGAATTCAACATCTCACAGAGATGGTGGAAAGATTATGGATTCCCTGAAATAACATCTCACGGCATGTCACGTGGAAATA										

Fig. 1 A screenshot of Primerquest® program showing output of qPCR primers

GTTTCTAATG, Reverse Primer = AAGTGGAGAGAGAGGGCTA TG. Set 2 ACC# JQ240295.1—Forward Primer = AGGCT GGGATACAGTCAG, Reverse Primer = GTCGAACGTGGGAAG ATAAC. Set 3 ACC# JQ240295.1—Forward Primer = ATAG CCTCTCTCTCCACTTC, Reverse Primer = ACTCCATCCTC CACTGAC.

3.4 Checking Amplicon for Secondary Structures

To view the primer assay details, click on “View Assay Details” under Set 1 ACC# JQ240295.1. This opens a window showing the forward and reverse primers and the amplicon location within the GOI. Find, highlight, and copy the amplicon into Windows Notepad. Remove any spaces or numbers from the nucleotide sequence. The copied amplicon is as follows: GACCACCTCCG TTTCTAATGAGGATGGGGTACCAAGACGCATAGCTG GTCATCATTCCAACCTTGGGACGATGATTCC ATAGCCTCTCTCCACTT.

Secondary structures formed by the amplicon can result in fluorescence, which can alter the qPCR quantification of the targeted GOI. One of the most important steps in primer design is to check for the presence of secondary structures. This can be accomplished by using free, online software called UNAFold.

Access the UNAFold open access website: <http://www.idtdna.com/UNAFold>. Paste the copied amplicon into the sequence input box. In the “Sequence Name” input box, type “Set 1 Primers”. Change the Mg concentration to 3 mM and click “submit”. The resulting secondary structures are presented in table format (Fig. 2). Check to assure that all secondary structures have a TM (°C) less than the qPCR annealing temperature (normally 55–60 °C). The results of this analysis show that there are 15 structures with a melting temperature of 49.5°C and below, all of these are below the annealing temperature. However it is advisable to use a ΔG value more than -9 kcal/mol.

3.5 Checking Primers

Each primer set must be checked for cross and self-dimers and hairpin formation. These steps are crucial for primers intended to be used in intercalating dye-based assays such as SYBR® Green. Beacon Designer, Free Edition, is an online software that allows users to analyze primers designed specifically for qPCR assays. It can be used to rule out cross dimers, self-dimers, and hairpin formations which can skew results during qPCR analysis.

Access Beacon Designer, Free Edition, at <http://www.pre-mierbiosoft.com/qpcr/>. Select the option for “SYBR® Green”. From the UNAFold website, copy and paste the forward (sense) primer and reverse (anti-sense) primer into the input boxes in Beacon Designer. Click the “analyze” button. The results will pop up in a new window.

The results from primer set 1 show that there are no self-dimers or hairpin formations. There are cross dimers that have a ΔG of

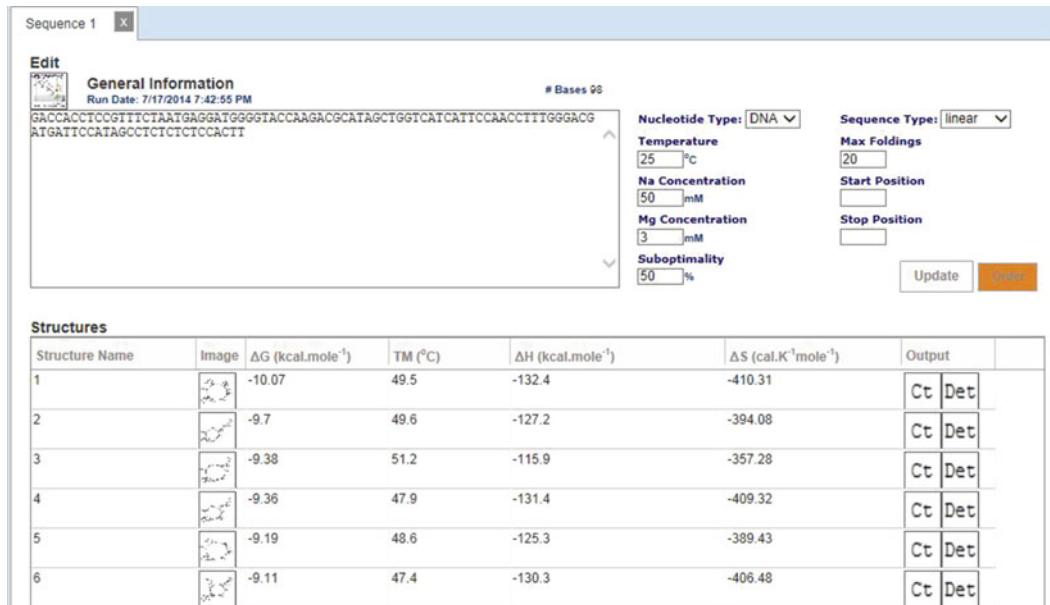


Fig. 2 A screenshot of UNAFold® program showing some possible secondary structure of primers

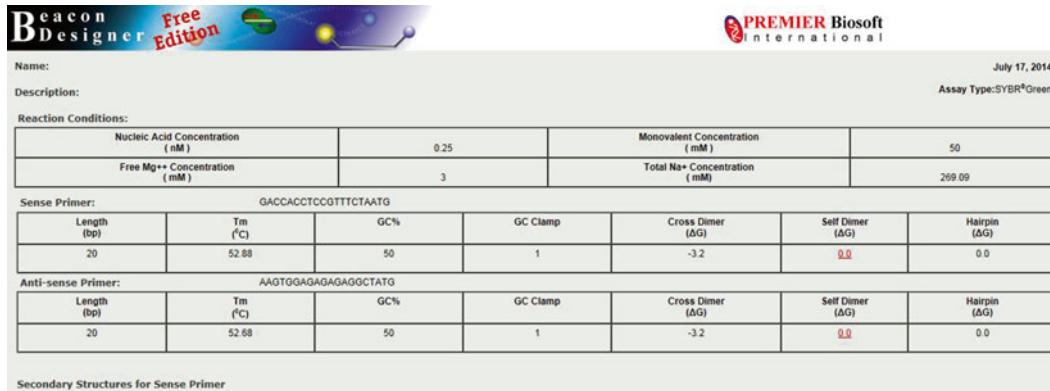


Fig. 3 UNAFold screenshot showing possibilities of cross, and self-dimer and hairpin formation

–3.2. Scroll down to view the structures that are formed. Secondary formations with a ΔG of –3.5 and below (–3.6, –4.0, etc.) should be avoided (Fig. 3).

4 Notes

Should the UNAFold or Beacon Designer, Free Edition software turn up secondary structures with ΔG values below –3.5, or melting temperatures of secondary structures that are above the

annealing temperature, first try another set of primers from the list. If none of the primers are satisfactory, then try altering the “Primer Criteria” for the following parameters: Primer GC content (between 45 % and 65 %, leave optimum blank); Primer melting temperature (between 50 °C and 68 °C, leave optimum blank). Under template criteria, increase the amplicon size from 150 bp to 250 bp, leave the optimum blank.

References

1. Karlen Y et al (2007) Statistical significance of quantitative PCR. BMC Bioinformatics 8:131
2. Vermeulen J et al (2001) Measurable impact of RNA quality on gene expression results from quantitative PCR. Nucleic Acids Res 39(9):e63
3. D'haene B, Hellemans J (2010) The importance of quality control during qPCR data analysis. Int Drug Disc:18–24
4. Bustin S, Nolan T (2004) Pitfalls of quantitative real-time reverse transcription polymerase chain reaction. J Biomol Tech 14:155–166
5. Paudel D et al (2001) Comparison of real-time SYBR Green dengue assay with real-time TaqMan RT-PCR dengue assay and the conventional nested PCR for diagnosis of primary and secondary dengue infection. N Am J Med Sci 3(10):478–485
6. Maeda H et al (2003) Quantitative real-time PCR using TaqMan and SYBR Green *Actinobacillus actinomycetemcomitans*, *Porphyromonas gingivalis*, *Prevotella intermedia*, *tetQ* gene and total bacteria. FEMS Immunol Med Microbiol 39:81–86
7. Tajadini M et al (2014) Comparison of SYBR Green and TaqMan methods in quantitative real-time polymerase chain reaction analysis of four adenosine receptor subtypes. Adv Biomed Res 3:85
8. Andersen C et al (2006) Equal performance of TaqMan, MGB, Molecular Beacon and SYBR Green-based detection assays in detection and quantification of roundup ready soybean. J Agric Food Chem 54:9658–9663
9. Thornton B, Basu C (2011) Real-time PCR (qPCR) primer design using free online software. Biochem Mol Biol Educ 39(2):145–154
10. Untergasser A et al (2012) Primer3—new capabilities and interfaces. Nucleic Acids Res 40(15):e115

Chapter 14

PRIMEGENSw3: A Web-Based Tool for High-Throughput Primer and Probe Design

Garima Kushwaha, Gyan Prakash Srivastava, and Dong Xu

Abstract

Highly specific and efficient primer and probe design has been a major hurdle in many high-throughput techniques. Successful implementation of any PCR or probe hybridization technique depends on the quality of primers and probes used in terms of their specificity and cross-hybridization. Here we describe PRIMEGENSw3, a set of web-based utilities for high-throughput primer and probe design. These utilities allow users to select genomic regions and to design primer/probe for selected regions in an interactive, user-friendly, and automatic fashion. The system runs the PRIMEGENS algorithm in the back-end on the high-performance server with the stored genomic database or user-provided custom database for cross-hybridization check. Cross-hybridization is checked not only using BLAST but also by checking mismatch positions and energy calculation of potential hybridization hits. The results can be visualized online and also can be downloaded. The average success rate of primer design using PRIMEGENSw3 is ~90 %. The web server also supports primer design for methylated sequences, which is used in epigenetic studies. Stand-alone version of the software is also available for download at the website.

Key words Primer, Probe, PCR, Primer-design, High-throughput design, DNA methylation, PRIMEGENS

1 PRIMEGENSw3: Web Server for PRIMEGENSw2

PRIMEGENSw3 (<http://primegens.org>) is a web-based tool for high-throughput primer and probe design. It is the web-based version of the existing stand-alone tool PRIMEGENSw2 [1–7] with several new features for genome-scale primer/probe design. It allows users to upload their one or many sequences on a high-performance server—not only for primer and probe design—but also for checking cross-hybridization with another set of uploaded sequence pools or a locally stored genomic sequence database on the server. Also, the specificity of the designed primers is checked using primer sequence along with several intuitive primer-specific 3'-end hybridization filters and free energy calculation. PRIMEGENSw3 automates the process of Primer3 [8], BLAST

[9] and considers additional hybridization conditions in order to design a highly reliable set of specific primers and probes.

Another unique feature of PRIMEGENSw3 is that it allows users to select various types of input formats and primer design algorithms to design primers and probes customized for their experimental requirements. There is currently no other software tool that allows such flexibilities to incorporate various types of design constraints and a customized algorithmic workflow suitable for different applications for PCR, microarray, and targeted sequencing.

PRIMEGENSw3 provides a set of web-based utilities for high-throughput primer/probe design. These utilities allow users to select genomic regions and to design primer/probe for selected regions in an automatic fashion. It runs embedded PRIMEGENSv2 algorithms at the backend on a high-performance server and uses the locally stored genomic database or any user-provided dataset for a cross-hybridization check. Cross-hybridization is assessed not only using BLAST but also by checking mismatch positions and energy calculation of potential hybridization hits. The results can be visualized online in a tabular format and can also be downloaded. In addition, the web server supports primer design for bisulfite-treated sequences for DNA methylation studies. A stand-alone version of the software is also available for download from the website.

2 Design Process and Protocol

PRIMEGENSw3 uses Primer3 to design primers, and it uses BLAST to check their specificity against the sequence database (custom or locally stored whole genome). Apart from using these well-known third-party tools to design primers/probes, PRIMEGENSv2 utilizes its own algorithms to optimize and customize the whole design process. One of PRIMEGENSv2's primer design algorithms is Sequence-Specific Primer Design (SSPD), which takes each input query sequence and designs multiple primer pairs scattered all over the sequence when using Primer3 and then selects the unique 15-mer oligonucleotides from the 3'-end of each primer (15 bases by default but optional). In its next step, SSPD runs Mega-BLAST [10] to perform gapless alignments of the 15-mer oligonucleotides against all the sequences in selected database files and removes all primer pairs with nonspecific PCR amplification. The second algorithm that PRIMEGENS provides is Fragment-Specific Primer Design (FSPD). FSPD can be used to design multiple primers for very long sequences, e.g., to cover long sequences like CpG island or another long genomic region of interest. In FSPD algorithm, PRIMEGENS first breaks down the whole query sequence into small overlapping/non-overlapping fragments, and then designs primers for each of these fragments similar to the SSPD algorithm. The third PRIMEGENS algorithm

is Probe-Specific Primer Design (PSPD), where it also designs sequence-specific segment (probe) for each input query sequence and then designs primer pairs of these probes to target a local region within the query sequence. The PSPD algorithm is different from other algorithms as it carries out a BLAST search for each query sequence against all sequences in the database file—first to find query-sequence-specific probes and then to design a primer for each probe. All these primer design algorithms are explained in more detail in [4–7].

During the cross-hybridization check for designed primers, PRIMEGENSv2 does not run Mega-BLAST on the whole primer sequence but on unique oligo at the 3'-end of all primers designed for faster processing. From all the recorded hits, it checks the occurrence of any hit of a reverse primer close to each of the hits of the forward primer. Presence of two opposite hybridization hits within amplifiable distance is considered as potential cross-hybridization hit for that primer pair. Only these hits are subjected for filtering in its subsequent steps for further optimization. Other than the uniqueness of primer sequence, PRIMEGENSv2 next checks the order and orientation of primer binding for a successful amplification of a sequence. Even if a primer pair binds in correct order and orientation, it checks if the resulting amplicon size is within a specific range for a successful PCR. PRIMEGENSw3, thus validates all potential cross-hybridization hits using all of these constraints.

After applying all the above-mentioned filters, sequence similarity at the 3'-end of each of the potential primer pair's cross-hybridization is checked. Sequences of selected potential cross-hybridization hits are extracted for each primer. These sequences are then compared to the potentially binding primer sequences in order to check for any mismatches between them that do not form a Watson–Crick base pair. Hybridization hits that have exact complementary matches in the last 2 bases in left primer and first 2 bases in right primer with no continuous mismatches in the next 3 bases are considered as potential alternative primers (Fig. 1). Sequence matching only for such primers may form a stable duplex and thus may cross-hybridize. The energy of each of the potential primer pair hybridizations thus obtained is computed using the Nearest-neighbor (NN) model [11–16].

Previous studies [17–22] showed that an NN model is sufficient to accurately predict the thermodynamics of a DNA duplex composed of Watson–Crick base pairs and internal single mismatches. The stability of a cross-hybridization duplex is calculated from its primary sequence based on the total free energy change (ΔG_o) of the DNA helix from its individual strand [23, 24]:

$$\Delta G_o(\text{total}) = \sum_i n_i \Delta G_o(i) + \Delta G_o(\text{init w/ term GC}) + \Delta G_o(\text{init w/ term AT}) + \Delta G_o(\text{sym}),$$

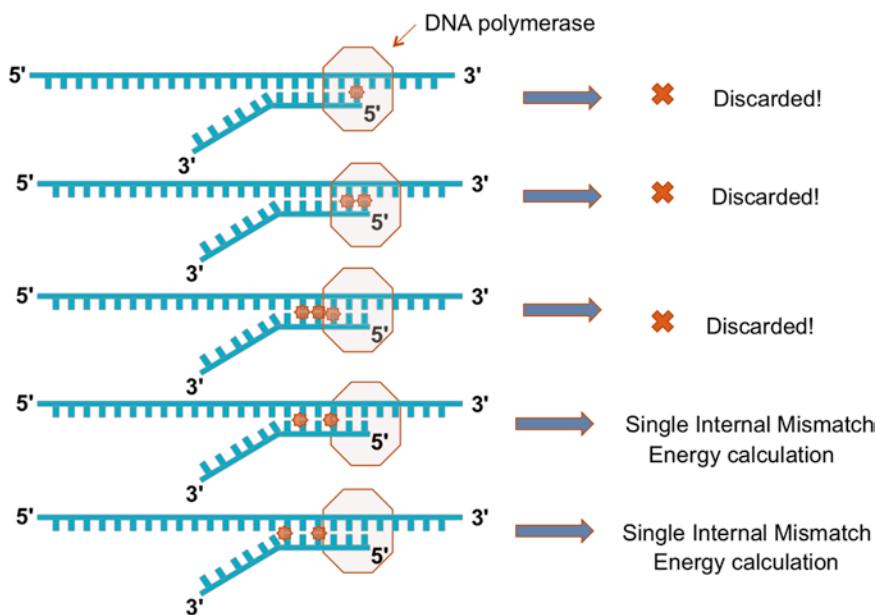


Fig. 1 Mismatch position validation to filter potential cross-hybridization candidates. For successful primer binding, six nucleotide positions at the 3'-end of DNA duplex are considered to be most important and analyzed for mismatch and energy calculation. The first three hybridizations are discarded due to the presence of mismatching at the first two nucleotide positions at their 3'-end. The fourth hybridization is discarded due to the presence of three consecutive mismatches in the internal nucleotide positions. The fourth and fifth conditions with more stable nonconsecutive internal mismatch positions at the 3'-end of hybridization are evaluated for their free energies

where $\Delta G_o(i)$ is the standard free energy change for Watson–Crick NNs; n_i is the number of occurrences of each nearest neighbor i , and $\Delta G_o(\text{sym})$ self-complementary sequences. $\Delta G_o(\text{init } w/\text{term GC})$ and $\Delta G_o(\text{init } w/\text{term AT})$ account for the differences between duplexes with terminal AT and terminal GC pairs, respectively. Single internal mismatch NN-parameters [20–28] are used where mismatches are found. The “Minimum 3'-end Stability” value provided by the user (also used by Primer3 for primer design) is used as the filter threshold. Therefore, only those hybridizations that pass these specificity filters are reported as cross-hybridizations for each primer pair designed.

3 Basic Primer Design Steps Using PRIMEGENS

Designing primers/probes using PRIMEGENSw3 has a simple and interactive process (Fig. 2). The whole process can be broadly divided into two basic steps: (1) uploading data files (PCR template files for primer design and optional database for cross-hybridization check); and (2) providing primer-design specifications for job submission and execution.

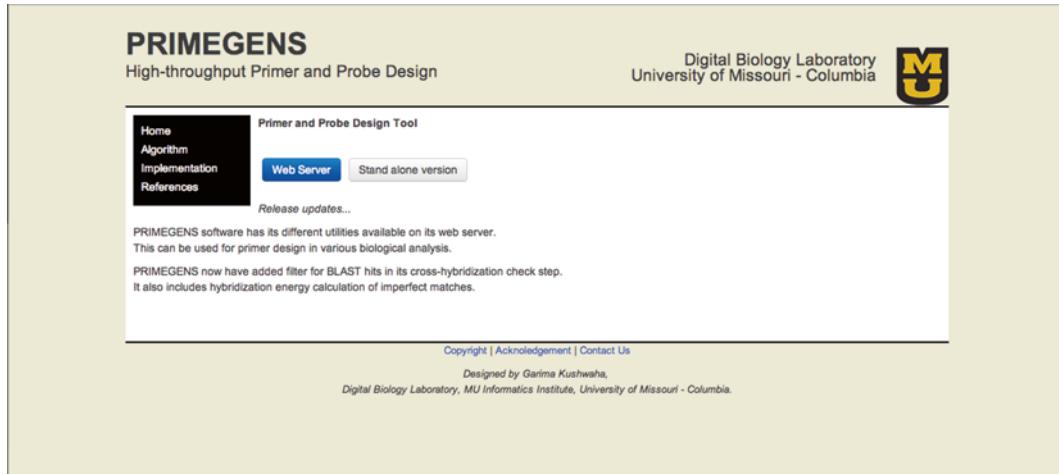


Fig. 2 PRIMEGENSw3 snapshot for first page showing options for web server and downloadable stand-alone version

1. Uploading data files

The first input is the query file having the sequence for which primers/probes need to be designed, and the second is the database file having all the sequences that are present in the PCR reaction. Sequences in the database files are used for PRIMEGENSw3 to check for any potential cross-hybridization and thereby select primer/probes that are specific to the query sequence. The user can upload their own customized database file or use available genomes supported by the server. We do not recommend that the user upload a large database file (e.g., a whole genome). Any new genome can be made available on the web server based on user request. PRIMEGENSw3 provides different sample data for both query and database sequences to enable users to test primer/probe design.

2. Primer/probe-design specifications

The next stage of PRIMEGENSw3 is to provide all input parameters for primer design. Input parameters on this page of the server are divided into same five sections as described in Subheading 3.1.3. All parameters have been set to the default values for best primer design. The default values are visible on the browser, and they can be changed by the user. After running PRIMEGENSw3, the server will show the link to all result files generated by the tool. PRIMEGENSw3 also generates files similar to those generated by the stand-alone version. These files can be viewed within the browser or downloaded on the user's local computer. In case the PRIMEGENS-w3 execution takes long time to complete due to the large size of input files, the server provides the user with a link to be used later to check for the result files. Along with

this, a user can also provide an email address (optional) to enable PRIMEGENS-w3 to send them an email notification for job completion and the link for the result files.

4 Web-Server Design Steps

Running primer design and selection by PRIMEGENSw3 is a step-wise process. A user has the ability to set all possible parameters required by primer3, BLAST, and PRIMEGENSv2 to run their primer design, interactively:

1. Select a task type
2. Input user data
 - (a) Query sequences file for primer/probe design
 - (b) Input sequence pool for cross-hybridization check
3. Choose an algorithm
 - (a) Sequence-specific primer design (SSPD)
 - (b) Fragment-specific primer design (FSPD)
 - (c) Probe-specific primer design (PSPD)
4. Set parameters
 - (a) Primer3 parameters
 - (b) BLAST parameters
 - (c) Algorithm-specific parameters
 - (d) Cross-hybridization-specific parameters
5. Run PRIMEGENS
6. Visualize results

4.1 Select Task

The first page on PRIMEGENSw3 web-server allows user to select the various features/services PRIMEGENS provides. There are four options for users to select, Regular primer design, Covering CpG Island, Covering Transcription Start Sites (TSS), and Covering maximum cut-sites region during amplification or probe design. Figure 3 shows the snapshot of its first page.

PRIMEGENSw3 provides some web-based utilities for users to design primers for selected genomic regions of interest like the nearest CpG island of a gene, a region around TSS of a gene and the local region with the highest density of restriction enzyme cut-sites in a given sequence or genomic locations along with regular primer design. These unique features of PRIMEGENSw3 can be used for custom applications like methylation patterns of various oncogenes/tumor suppressor genes.

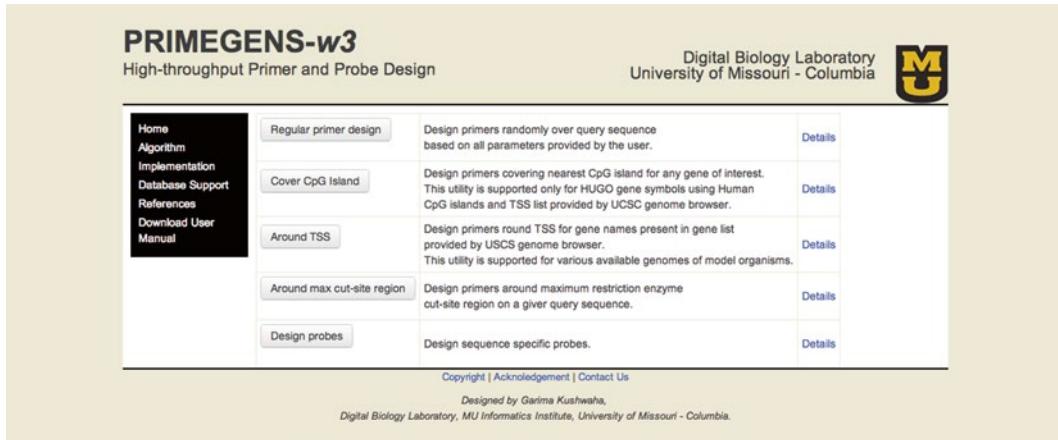


Fig. 3 PRIMEGENSw3 snapshot for presenting the utilities

(a) Utility-1: Primer design covering CpG island

This utility is used to design primers for genes that have CpG islands in close proximity to their respective TSSs. PRIMEGENS locates CpG islands to the left, right, or across the TSS of each gene and tries to cover either the whole or partial CpG islands along with its TSS. PRIMEGENSw3 extracts the final query sequence for the primer design with the user-specified length. This feature is supported only for the human genome at present with its TSS and CpG locations downloaded from the UCSC genome browser. Other supported genomes will be added and tested in future updates or, if requested, for a specific species.

(b) Utility-2: Primer design around TSS of a gene

“Primer design covering TSS” is a feature of PRIMEGENSw2, which helps the designed primers to cover the region around TSS of any gene. PRIMEGENSw3 is capable of extracting respective TSSs for each given gene name from the UCSC Genome Database (currently for the March 2006 assembly) and obtaining the final query sequence to be used for primer design. Query sequence is chosen according to the length of upstream and downstream sequence provided by the user on the query upload page. The user is only required to provide gene symbols for which the primer design is required.

(c) Utility-3: Primer design around the maximum cut-site region

This feature is used to search for regions of user-specified length having maximum enzyme digestion sites (cut-sites) density within each given query sequence and design primers around it. Using this feature ensures the presence of given cut-sites in the PCR product, which is very useful in Methyl-Specific PCR. It uses a sliding window protocol to search for such region and then design primers. PRIMEGENSw2 locates

the maximum cut-site density region in the whole query sequence and designs primer pairs within this region.

(d) Primer and probe design for bisulfite-converted sequences

PRIMEGENSv2 can design PCR primers for bisulfite-specific PCR (BSP) to study DNA methylation after the bisulfite treatment of the DNA sequence. Primers used for BSP contain no CpG sites; thus, both methylated and unmethylated DNA sequences can be amplified. It requires the input of bisulfite-converted query sequences and four variants of the database to check for cross-hybridization. PRIMEGENS provides four chromosomal files for each human chromosome sequence on its server as the model of the bisulfite-treated sequences: (1) bisulfite-methylated forward sequence, (2) bisulfite-methylated reverse sequence, (3) bisulfite-unmethylated forward sequence, and (4) bisulfite-unmethylated reverse sequence. It thus uses these sequences to check cross-hybridization while designing primers for bisulphate-converted sequences.

(e) Probe design algorithm for target enrichment

The probe design algorithm is used to search for sequence-specific probes with no BLAST hits against some set of sequences (database). PRIMEGENS allows users to design probes to certain regions of a genome rather than a whole genome. It takes user input to obtain the genomic region of interest in supported data format and uses a sliding window protocol to find fragments of user-given length. It then uses BLAST to ensure that it outputs only those probes that are specific to the target region and that it will not hybridize anywhere else.

4.2 Input Sequences

After defining the design task, PRIMEGENSw3 asks for the input files on its next page. The user needs to prepare two types of input data. First, a list of target sequences in one of the formats supported by the tool, and then, a database of sequences to be used for the cross-hybridization check. Figure 4 shows a webpage snapshot for Input sequence page at PRIMEGENSw3.

Different formats of input query sequences that can be used are as follows:

(a) Input: target sequences

PRIMEGENSv2 allows users to input query/target data with or without actual nucleotide sequences as long as these sequences can be retrieved from the input sequence database with the same names/identifiers. If users have sequences, they can provide input data in the FASTA format having header line with a “>” sign followed by a sequence starting at the following line. In case of query data without actual sequences, a user

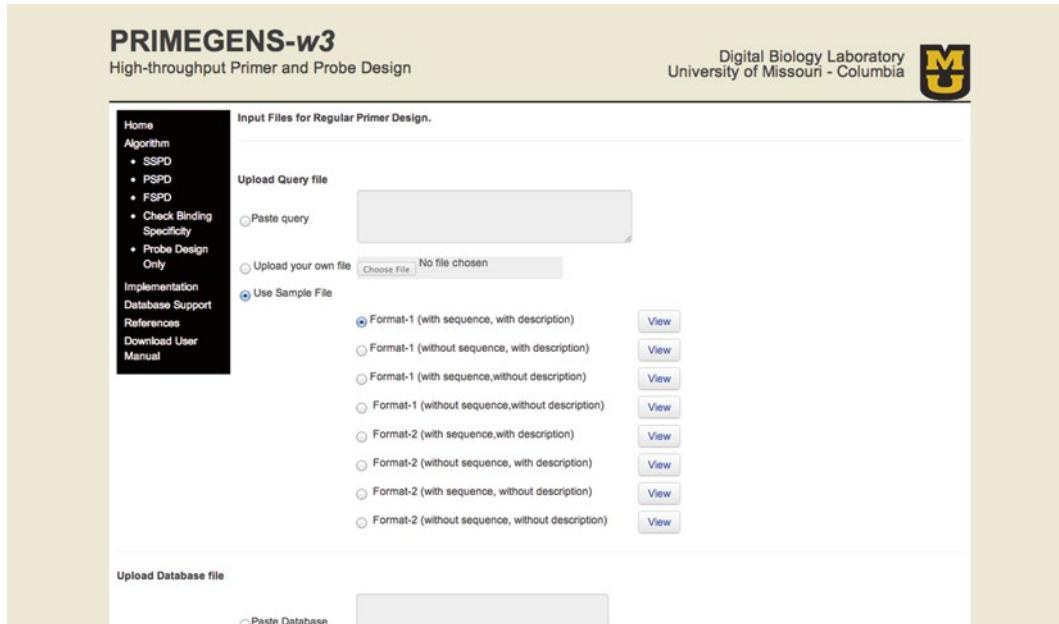


Fig. 4 PRIMEGENSw3 snapshot for the query and database input page

can employ one of the four input formats, which are (1) query name/identifier, (2) query name/identifier along with description, (3) chromosome position, and (4) chromosome position along with description. The chromosome position of genes from any genome, for example >chr21:33031597-33041570 can be used. In this case, the user has to provide the whole genome as the database. The genome must be in the chromosomal files, and there must be one file for each sequence containing the full sequence of the corresponding genome. Samples for all the different input format types supported by PRIMEGENS are provided on the input sequence webpage. Different formats are as follows:

Format-1 (with sequence): This is a normal multi-Fasta formatted file with multiple header lines followed by query sequence. In this format, a header line consists of query sequence name followed by its description separated by space.

Format-1 (without sequence): In this format, a user can provide just the header line with description. PRIMEGENS will extract the sequence from the “Database” input file.

Format-2 (with sequence): This is a multi-Fasta formatted file with just the sequence identifier. It consists of query sequence after each header line.

Format-2 (without sequence): In this format, only a list of sequence identifiers preceded by ‘>’ is given with no sequence. PRIMEGENS extracts sequence based on identifiers from the “Database” input file.

Format-3 (with sequence): This is a multi-Fasta formatted file with a headerline in the “>chr*:start-end<TAB>description” format followed by query sequence in the next line.

Format-3 (without sequence): In this format, just the chromosomal positions are provided from the species genome in the “>chr*:start-end<TAB>description” format. PRIMEGENS extracts the sequence from the “Database genome” selected by the user.

Format-4 (with sequence): This format is similar to Format-3 (with sequence) but without description like “>chr*:start-end” followed by query sequence in next line.

Format-4 (without sequence): This format is similar to Format-3 (with sequence) but without description like “>chr*:start-end”.

Format-5 (without sequence): In this format, a user can provide a list of gene symbols (HUGO/ENSEMBL ids), and PRIMEGENS will extract the corresponding sequences to use as input for primer design by PRIMEGENS.

View buttons with a link in front of each sample input format is provided to look at the file with the corresponding format. For all the input files without sequence, PRIMEGENS provides the Fasta formatted extracted sequence file with suffix “_fasta.txt” used as query input for primer design.

(b) Input: database (sequence pool/genome)

In the second input file, the user has to provide PRIMEGENS a database consisting of all the sequences against which cross-hybridization has to be checked for each primer/probe designed. Database input can either be a FASTA format file or one of the available genomes on the server. For query input in Format-4 (without sequence), PRIMENGENS needs a Database genome selected out of the listed available model organism genomes. Currently, PRIMEGENS provides the following Database genomes divided into parts sequence database and whole genome database.

Sequence database:

- Arabidopsis TAIR9 cDNA
- Arabidopsis TAIR9 CDS
- Maize CDS
- Medicago CDS

- Rice all cDNA
- Rice all CDS
- Sorghum all CDS
- Soybean Glyma1 cDNA
- Soybean Glyma1 CDS

Whole genome database:

- Human genome
- Bisulfite human genome
- *Anopheles gambiae*
- *Caenorhabditis elegans*
- *Drosophila melanogaster*
- Guinea Pig
- *Mus musculus*
- *Saccharomyces cerevisiae*
- Zebrafish

For cDNA and CDS, a user can also provide the data in Format-1/2/5 (with/without sequence). Additional genomes can be made available upon request.

4.3 Algorithm Selection

The next page is to choose a primer design algorithm supported by PRIMEGENS. It designs primers using three different algorithms: (1) Sequence-Specific Primer Design (SSPD), (2) Fragment-Specific Primer Design (FSPD), and (3) Probe-Specific Primer Design (PSPD). All three algorithms are briefly explained in Subheading 2.

4.4 Set Parameters

The next page onwards, the web server starts asking for all the parameters required to run PRIMEGENSw2. The different types of parameters required are:

(a) Primer3 parameters

After selecting design algorithm, PRIMEGENS' web-server brings up the page for user to set all required parameters for primer3. Different Primer3 parameters are divided into four sections, “General parameters,” “Advanced parameters,” “Penalty weights for primers,” and “Penalty weights for primer pairs,” as shown on the website and in Fig. 5.

There is a help icon button provided next to each of the Primer3 parameters for user to see the meaning and significance of each parameter as provided in Primer3. All the parameters listed are set with default values, which can be used as is or can be changed as desired. Press “Next” button to go to next page.

The screenshot shows the PRIMEGENS-w3 web application. The header includes the title "PRIMEGENS-w3" and "High-throughput Primer and Probe Design". It also features the "Digital Biology Laboratory" logo and the "University of Missouri - Columbia" text. A navigation menu on the left lists "Home", "Algorithm" (with sub-options: SSPD, PSPD, FSPD, Check Binding Specificity, Probe Design Only), "Implementation", "Database Support", "References", and "Download User Manual". The main content area is titled "Input Parameters" with a note "(All parameters have been set to a default value.)". It contains several input fields for Primer3 parameters, grouped under "Advanced setting Parameters". The fields include "Primer min quality" (default:0), "Primer quality range min" (default:0), "Primer inside penalty" (default:-1.0), "Primer explain flag" (default:1), "Primer min end quality" (default:0), "Primer quality range max" (default:100), "Primer outside penalty" (default:0.0), and "Primer file flag" (default:0). Below the input fields are "Next" and "Back" buttons. At the bottom, there are links for "Copyright | Acknowledgement | Contact Us" and a note: "Designed by Garima Kushwaha, Digital Biology Laboratory, MU Informatics Institute, University of Missouri - Columbia."

Fig. 5 PRIMEGENSw3 snapshot for setting Primer3 parameters

(b) BLAST parameters

The next page is to set parameters for MegaBLAST, such as word size and gap penalty, to look for cross-hybridization of primers in database sequences. BLAST parameters are also set to default values as shown in the input area. A user can change the default values if they want, or use default values by not putting anything in any of the textboxes and pressing the “Next” button to continue to following steps.

(c) Algorithm-specific parameters

The next page asks for the parameters in the algorithm-specific primer design parameters. First, a few parameters are for general PRIMEGENSv2 program and then parameters specific to the primer design algorithm are listed in its second page like hybridization parameters and oligo design parameters (e.g., fragment length for FSPD and probe length for probe design).

4.5 Run PRIMEGENS

After setting all the desired primer design parameters, the user can press the “Run PRIMEGENS” button to start executing the PRIMEGENS tool on our server as shown in Fig. 6. User can select (check box) option ask PRIMEGENS to send a link to primer design results on our server via email after successful completion of his or her job. Alternatively, if the computational job is small they can click “refresh” button in order to update their web-page and visualize their results.

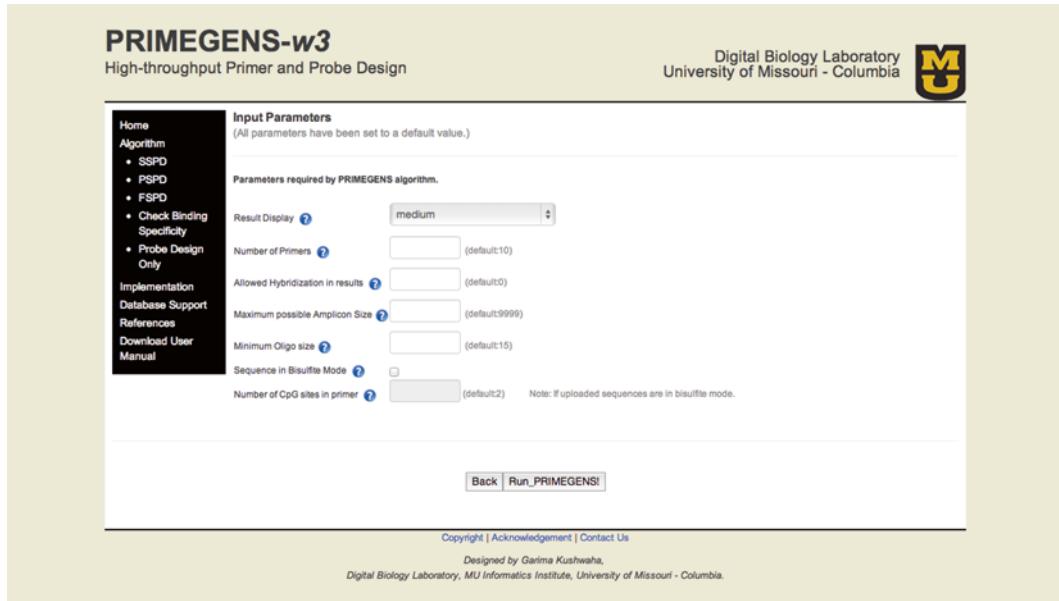


Fig. 6 PRIMEGENSw3 snapshot for launching PRIMEGENSv2 execution page

4.6 Visualize Results

After successful execution of primer/probe design, best results are shown in the form of a table with all information about each designed primer pair including query names with their left and right primer sequences and their respective characteristics, such as start position, length, melting temperature, amplicon size, and number of cross-hybridizations for the primer pair as shown in Fig. 7. Double-clicking on any row of this table enables the visualization of the primers in its corresponding query sequence as shown in Fig. 8. Also, this result webpage provides links to all output files generated by PRIMEGENSw3, that enables user to see the results in their browser or download by right click on each link.

Different output files generated by PRIMEGENSv2 are described as follows:

(a) Output: best primer pairs

Output file with best primer pairs is the most important output file generated by PRIMEGENSv2. It is the Excel sheet containing the best primer pair for each input query sequence provided by the user. It also contains various types of primer pair-related information and product size from each primer pair.

(b) Output: alternate primer pairs

In addition to the best primer pairs, PRIMEGENSv2 produces another file that contains alternate primer pairs for each input query sequence. In case a user wants to select alternate primer pairs instead of the best primer, or if the best primer

PRIMEGENS-w3
High-throughput Primer and Probe Design

Digital Biology Laboratory
University of Missouri - Columbia 

Design Results

Best primers

*Double click on any primer pair result in the following table to visualize their position on its corresponding sequence.

Query name	Left Primer-Sequence	Right primer-Sequence	Product size	Diagram		
>S11529022 Zea ...	ATGGTCAGGTCTCCCTG	AAGTGAAGTCATTTCGGGGTG	Start position : 1149 Length : 20 Tm : 57.5 GC-content : 50	Start position : 1318 Length : 20 Tm : 60.0 GC-content : 50	170	
>S22334604 Zea ...	GCCATTTCCTGGTGGGAA	CAACTTTCTGGTGGGTT	Start position : 332 Length : 20 Tm : 59.7 GC-content : 50	Start position : 486 Length : 20 Tm : 60.0 GC-content : 50	155	

Following are the result files available after running PRIMEGENS:

QueryFormat1_ZeaMays.fa Query file
 QueryFormat1_ZeaMays.fa_query_primer_list.txt alternate primers for each sequence
 QueryFormat1_ZeaMays.fa_primer.xls best primer sequences

Copyright | Acknowledgement | Contact Us
 Designed by Garima Kushwaha,
 Digital Biology Laboratory, MU Informatics Institute, University of Missouri - Columbia.

Fig. 7 PRIMEGENSw3 snapshot for design result visualization

PRIMEGENS-w3
High-throughput Primer and Probe Design

Digital Biology Laboratory
University of Missouri - Columbia 

Design Results

Best primers

*Double click on any primer pair result in the following table to visualize their position on its corresponding sequence.

Query name	Sequence
>S11529022 Zea ...	GAACGAGAACCGCTCTCTGCGATGGCATGGCTGCTCATGCTCTCGCTACACCGTGCGACACTCAGGGAGG
>S22334604 Zea ...	GGGGCGGCCTGTCGAGTTCTCGGGGTCAAGGAGCTCGGGCTCGCTGCCATCGCCGAATGCGACCTCCGAG

Following are the result files available after running PRIMEGENS:

QueryFormat1_ZeaMays.fa Query file
 QueryFormat1_ZeaMays.fa_query_primer_list.txt alternate primers for each sequence
 QueryFormat1_ZeaMays.fa_primer.xls best primer sequences

Copyright | Acknowledgement | Contact Us
 Designed by Garima Kushwaha,
 Digital Biology Laboratory, MU Informatics Institute, University of Missouri - Columbia.

Fig. 8 PRIMEGENSw3 snapshot for query sequence visualization page

pair fails to amplify or match other desired criteria, this file provides multiple choices for additional primer pairs for each query sequence. These primer pairs are ranked based on the number of potential cross-hybridizations.

(c) Output: failed sequences

PRIMEGENS-v2 also generates one file containing all FASTA formatted sequences whose the primer design failed. There could be various reasons for the failure, including design failure from the Primer3 program. The user can analyze these sequences further and rerun PRIMEGENS only for failed sequences using modified parameters.

(d) Output: probe for target sequencing

In addition to three standard output files, the PSPD algorithm generates this additional output file, i.e., gene-specific fragments (recorded with name of the query file followed by “uniseg.txt”). This file contains gene-specific fragments (probe) for each input query sequence that PSPD finds. These are the gene-specific fragments that PSPD ultimately uses to design primers for their corresponding query sequences.

In case a primer/probe design is not complete at the time the user tries to visualize the results, the web-server provides a link to the directory on the server where the user’s job is submitted. This link can be used anytime to access the prime/probe design results. The design results folder consists of all the different output files stated above along with the following additional files:

(a) Primergens_report.txt

This file consists of the detailed output from PRIMEGENSv2 execution. A user can track the detailed execution process through this file.

(b) Run_status.txt

This file consists of the status of PRIMEGENSv2 execution and tells if its primer/probe design run was successfully completed or is still running on the server.

(c) Primegens.log

This is log file for PRIMEGENSv2 execution. If PRIMEGENSv2 fails to run for some reason, an error message is logged in this file, which can be checked to discover the reason for its failure.

(d) Format_database.log

This file compiles the log while running formatdb on database input used by MegaBlast tool.

(e) Config.txt

This file consists of all the parameter values used by PRIMEGENSv2 to run primer/probe design. This file is also required by PRIMEGENSv2 to run on the command line.

If “cover CpG” design task is chosen, then three additional files are generated in the result folder. They are:

- (a) cpg_util_output.txt

This is the file Web-server generates to be used as the formatted query input for PRIMEGENSv2 execution.

- (b) cpg_util_output2.txt

This file consists of information about the location of CpG island around the TSS for each gene symbol provided in the query file input.

- (c) cpg_symbol_error.txt

This file consists of all those gene symbols from query gene list for which no CpG island information could be extracted from the UCSC genome browser tables.

If “Around TSS” design task is chosen, three additional files are generated in the result folder. They are:

- (a) tss_util_output.txt

This is the file Web-server generates to be used as the formatted query input for PRIMEGENSv2 execution during “Around TSS” utiliy. Query positions are generated for each gene symbol around their respective TSS based on the design parameters.

- (b) tss_util_output2.txt

This file consists of the information about locations of TSS and TES (Transcription End Site) for each gene symbol provided in the query file input.

- (c) tss_symbol_error.txt

This file consists of all those gene symbols appearing on the query gene list for which no TSS information could be extracted from the UCSC genome browser tables.

5 Discussion

PRIMEGENS $w3$ can be used for large-scale primer and probe design for various applications, such as PCR, DNA synthesis, qRT-PCR (gene expression), and targeted next-generation sequencing (454, Illumina, Agilent Sure-select technology, etc.) for normal and bisulfite-treated sequences. Web-server version of PRIMEGENSv2 removes the necessity for users to install and configure the software locally and also enables them to use stored genomic sequences on the server in performing cross-hybridization checks. PRIMEGENS $w3$ allows users to upload their own genomic sequences, design primers, filter them based on cross-hybridization, and then view the resulting primers. All results are retained on the server temporarily and is accessible through a unique URL

provided by the website. Website also contains links to example datasets for each of supported utilities to test the tool and visualize results. Also, a detailed instruction about running PRIMEGENSw3 is provided in a user manual, and description of each algorithm is also provided in help pages on its website.

PRIMEGENSw3 can be used for targeted or genome-wide primer design in a very interactive and user-friendly manner. Its various add-on utilities can be used to design specific primers of a target region, e.g., CpG island and promoter region. It has also been successfully used for splice variants' specific primer designs [18].

6 Future Development

We will further improve the computational efficiency and accuracy of the software. For example, we currently use BLAST for cross-validation check. The seed word used by BLAST for perfect match to extend in both directions can be at any position within the primer sequence, which instead should be at the 3'-end region of the primer where it binds the template sequence. It can also be less than the minimum allowed word-size of eight nucleotides for primer binding. For this, there is a need for an alignment tool like BLAST to do the sequence identity search as fast as BLAST but use the thermodynamic score for each base pairing to overcome the above-mentioned problem. We will explore a new indexing approach to address this issue.

Thermodynamic calculations for cross-hybridizing duplexes in PRIMEGENSw3 use the NN model of Watson–Crick base-paired duplexes and single internal mismatches. It can be extended to salt dependence, terminal-dangling ends, and also terminal mismatches for energy calculations in order to minimize mispriming and searching for potential hybridization.

Applicability of this tool can also be extended to design small artificial interfering RNAs for RNA interference (RNAi) [29], which is another similar area involving sequence hybridization studies [30, 31]. Recently, artificial miRNA (amiRNA) and tasiRNA (atasiRNA) [32] technologies have been developed as silencing technologies. AmiRNA and atasiRNA utilize miRNA (19–24 nt) and siRNA (21 nt), respectively, that are complementary to their target sequences and, therefore, are highly sequence-specific, minimizing off-target effect and capable of silencing gene family members sharing a high degree of homology. It is critically important to be able to design the 21 nt tasiRNA accurately for effective silencing of targets. Existing small RNA design tools design only one sequence at a time and do not consider cross-hybridization to off-target silencing using actual thermodynamics energy. Cross-hybridization check used in for primer selection in PRIMEGENSw3 can be extended for designing amiRNA and atasiRNA.

7 Availability and Requirements

PRIMEGENSw3 is a fully functional, open-source software tool. It is available as a web service backed by perl CGI scripts and hosted on an in-house server. It is available at <http://primegens.org/>. This web server has been tested on various PCR applications with different organisms. The stand-alone version of PRIMEGENS has been publicly available since 2002, and has extensive usage; its source code is at <http://code.google.com/p/primegens>. Average success rate of primer design using PRIMEGENS algorithm reaches ~90 % [5, 6].

Acknowledgements

This work has been supported by National Institute of Health (Grant 1R01-DA025779) and the Congressionally Directed Medical Research Programs, U.S. Army Medical Research and Materiel Command (Grant BC062135). This research is also supported in part by the Paul K. and Diane Shumaker Endowment in Bioinformatics.

References

1. Xu D, Li G, Wu L, Zhou J, Xu Y (2002) PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics* 18:1432–1437
2. Srivastava GP, Xu D (2007) Genome-scale probe and primer design with PRIMEGENS. *Methods Mol Biol* 402:159–176
3. Srivastava GP, Guo J, Shi H, Xu D (2008) PRIMEGENS-v2: genome-wide primer design for analyzing DNA methylation patterns of CpG islands. *Bioinformatics* 24:1837–1842
4. Srivastava GP, Kushwaha G, Shi H, Xu D (2010) High-throughput primer and probe design for genome-wide DNA methylation study using PRIMEGENS. In: A practical guide to bioinformatics analysis. iConcept Press Ltd, G. Fung ed. Hong Kong, pp 151–171
5. Lee EJ, Pei L, Srivastava GP, Joshi T, Kushwaha G, Choi JH, Wang X, Mockaitis K, Colbourne J, Zhang L, Schroth GP, Xu D, Zhang K, Shi H (2011) Targeted bisulfate sequencing by solution hybrid selection and massively parallel sequencing. *Nucleic Acids Res* 39(19):e127
6. Srivastava GP, Hanumappa M, Kushwaha G, Nguyen HT, Xu D (2011) PRIMEGENS-v2: homolog-specific PCR primer design for profiling splice variants. *Nucleic Acids Res*:39(10):e69
7. Kushwaha G, Srivastava GP, Xu D (2011) PRIMEGENSw3: a web-based tool for high-throughput primer and probe design. In: Bioinformatics and Biomedicine (BIBM), 2011 IEEE international conference, 2011, 345, 351. doi: [10.1109/BIBM.2011.6143](https://doi.org/10.1109/BIBM.2011.6143)
8. Rozen S, Skaltsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365–386
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
10. Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7:203–214
11. Crothers DM, Zimm BH (1964) Theory of the melting transition of synthetic polynucleotides: evaluation of the stacking free energy. *J Mol Biol* 9:1–9
12. DeVoe H, Tinoco I Jr (1962) The stability of helical polynucleotides: base contributions. *J Mol Biol* 4:500–517
13. Gray DM, Tinoco I Jr (1970) A new approach to the study of sequence-dependent properties of polynucleotides. *Biopolymers* 9:223–244
14. Borer PN, Dengler B, Tinoco I Jr, Uhlenbeck OC (1974) Stability of ribonucleic acid double-stranded helices. *J Mol Biol* 86:843–853
15. Tinoco I Jr, Borer PN, Dengler B, Levine MD, Uhlenbeck OC, Crothers DM, Gralla J (1973) Improved estimation of secondary structure in ribonucleic acids. *Nat New Biol* 246:40–41

16. Uhlenbeck OC, Borer PN, Dengler B, Tinoco I Jr (1973) Stability of RNA hairpin loops. *J Mol Biol* 73:483–496
17. Breslauer KJ et al (1986) Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A* 83(11):3746–3750
18. Rychlik W, Spencer WJ, Rhoads RE (1990) Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res* 18(21):6409–6412
19. Owczarzy R, Vallone PM, Gallo FJ, Paner TM, Lane MJ, Benight AS (1997) Predicting sequence-dependent melting stability of short duplex DNA oligomers. *Biopolymers* 44(3):217–239
20. Ke S-H, Wartell RM (1993) Influence of nearest neighbor sequence on the stability of base pair mismatches in long DNA: determination by temperature-gradient gel electrophoresis. *Nucleic Acids Res* 21(22):5137–5143
21. SantaLucia J Jr, Allawi HT, Seneviratne PA (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* 35:3555–3562
22. Santalucia JA (1998) Unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA* 95:1460–1465
23. Peyret N, Seneviratne PA, Allawi HT, SantaLucia J (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal AA, CC, GG, and TT mismatches. *Biochemistry* 38:3468–3477
24. Hyndman D, Cooper A, Pruzinsky S, Coad D, Mitsuhashi M (1996) Software to determine optimal oligonucleotide sequences based on hybridization simulation data. *Biotechniques* 20:1090–1094, 1096–1097
25. Allawi HT, SantaLucia J Jr (1997) Thermodynamics and NMR of Internal G-T Mismatches in DNA. *Biochemistry* 36: 10581–10594
26. Allawi HT, SantaLucia J Jr (1998) Nearest-neighbor thermodynamics of internal A-C mismatches in DNA: sequence dependence and pH effects. *Biochemistry* 37: 9435–9444
27. Allawi HT, SantaLucia J Jr (1998) Thermodynamics of internal C-T mismatches in DNA. *Nucleic Acids Res* 26:2694–2701
28. Allawi HT, SantaLucia J Jr (1998) Nearest neighbor thermodynamic parameters for internal G-A mismatches in DNA. *Biochemistry* 37:2170–2179
29. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391:806–811
30. Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, Mao M, Li B, Cavet G, Linsley PS (2003) Expression profiling reveals off-target gene regulation by RNAi. *Nat Biotechnol* 21:635–637
31. Baulcombe DC (2007) Molecular biology. Amplified silencing. *Science* 315:199–200
32. Allen E, Howell MD (2010) MiRNAs in the biogenesis of trans-acting siRNAs in higher plants. *Semin Cell Dev Biol* 8:798–804

Chapter 15

Selecting Specific PCR Primers with MFEprimer

Wubin Qu and Chenggang Zhang

Abstract

Selecting specific primers is crucial for polymerase chain reaction (PCR). Nonspecific primers will bind to unintended genes and result in nonspecific amplicons. MFEprimer is a program for checking the specificity of PCR primers against the background DNA. In this chapter, we introduce: (1) the factors that affect the specificity of primers; (2) the principle of MFEprimer and its settings; (3) how to use the MFEprimer to examine the specificity of primers.

Key words PCR, Primer specificity, MFEprimer

1 Introduction

Designing specific primers is a key step in PCR. The forward and reverse primers act like two anchors, which determine the region of interest to be amplified. However, in real experiments, due to the presence of background DNA, which usually including genomic DNA (gDNA) and complementary DNA (cDNA), even the well-designed primers can bind to mistargeted DNA regions and result in nonspecific amplicons. The background DNA here acts as competitor of target DNA. Nowadays, with the availability of the genomic DNA sequences of most model organisms, it is now possible to check the specificity of primers against the background DNA.

MFEprimer(<http://biocompute.bmi.ac.cn/CZlab/MFEprimer-2.0/>) is a program to check the specificity of PCR primers against the background DNA. The first version of MFEprimer was released in 2009 [1] and the current version is 2.0 [2]. Although there are several programs available [3–6] for checking the specificity of PCR primers before our work, none of them treats primer hybrid process on the base of the biochemical reactions. These programs mainly use the sequence alignment tools such as National Center for Biotechnology Information's Basic Local Alignment Search Tool (NCBI BLAST) [7] program to find the primer binding sites and neglect the important factors which influence the specificity of

primers, such as the binding stability of the 3'-end of a primer duplex (formed by the primer and its target DNA). Other limitations of these programs include fewer organisms support and slower running speed. Therefore, we developed the MFEprimer program with several important features: (1) Mimicking the thermodynamics hybrid process of the primers and the target amplicons. (2) Using the k -mer algorithm to speed up the binding sites searching process, for example, the running time is usually 1–10 s even for the whole genomic DNA searching with only one CPU used in popular personal computer. Here, a k -mer is defined as a short DNA sequence with a length of k nucleotides. (3) Comprehensive supported organisms and support for custom database upon users' request.

In the following sections, we (1) introduce the factors that affect the specificity of primers, (2) explain the principle of MFEprimer and its settings, and (3) use an example to illustrate the usage of MFEprimer.

2 Factors That Affect the Specificity of PCR Primers

Here, we focus on the nonexperimental factors that affect the specificity of PCR primers. The nonspecific binding sites of the PCR primers are the major sources responsible for nonspecific amplifications. There are three factors which can influence the nonspecific binding sites of a primer: (1) Background DNA is the source with a large number of potential binding sites. Therefore, it is critical to know the composition of background DNA. (2) Stable binding sites (including nonspecific ones) can form stable primer duplex, which in turn have chance to attract the DNA polymerase to bind for next elongation process. (3) After DNA polymerase binds to primer duplex, only the stable 3'-end of a primer duplex can attract DNA polymerase to start the elongation process. In other words, for a given primer and background DNA, the binding stability of the whole primer determines whether a binding site would be bound by DNA polymerase, while the stability of its 3'-end (default is the last nine residues [*see Note 1*]) determines whether a non-specific amplicon would be amplified.

2.1 The Background DNA

The background DNA acts as a competitor of target DNA and is the major source of the unexpected primer binding sites. Generally, the user aims to selectively amplify the specific region from the target DNA, which mixed with the background DNA. Therefore, it is obviously that we should design primers specific to target DNA and avoid nonspecific binding to background DNA. As a result, knowing background DNA is the precondition of designing specific PCR primers.

2.2 The Stability of the Primer Duplex

The annealing and extension are two key steps, which determine the anchor sites of a PCR primer. The annealing process determines which sites a primer could bind to, while the extension step decides whether the binding site could be extended.

In the annealing process of PCR, although there are plenty of DNA sites (from both target DNA and background DNA) available for binding, the primer can only bind to the sites which can form stable duplex. The binding stability of the binding site is measured by Gibbs free energy (ΔG) [8, 9]. The lower of the ΔG , the more stable is the duplex. The most stable binding site of the PCR primer is the site that has perfect match (its complement reverse sequence, shown in Fig. 1a). Some binding sites though have one or two mismatches (shown in Fig. 1b, c), still can form stable duplex with the primer [8–10]. That is because some mismatches still contribute much to the duplex stability. For instance, a G–G mismatch can contribute as much as -2.2 kcal/mol to duplex stability at 37°C [8] (see Note 2).

2.3 The 3'-End Stability of the Primer Duplex

After the annealing process, the DNA polymerase will bind to the 3'-end of the primer duplex, and then synthesize a new DNA strand complementary to the DNA template strand by adding dNTPs that are complementary to the template in 5'-3' direction. Therefore, a stable 3'-end of a primer is essential for elongation, but not for the 5' end [10, 11]. Therefore, the 5'- and 3'-ends of a primer have different scoring weights [10]. An unstable 3'-end (e.g., mismatches in 3'-end, shown in Fig. 1c) will stop the elongation process, while the 5'-end allows an addition of a tagging sequence or mismatches and did not affect the 3'-end elongation (shown in Fig. 1b). However, a high stable 3'-end will also contribute substantially to nonspecific extension in PCR reactions [1, 11] (see Note 1).

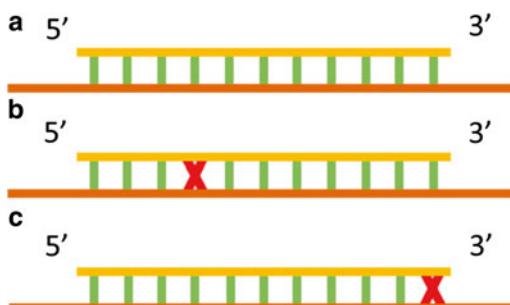


Fig. 1 Primer binding diagrams. Primer binds perfectly (a), with one mismatch near the 5' terminal (b) and with one mismatch at the 3' terminal (c). Yellow bar: the PCR primer. Long red bar: the DNA template sequence. Red cross: mismatch region

3 MFEprimer

3.1 Principle

MFEprimer is a fast and thermodynamics-based PCR primer specificity-checking program [1, 2]. The genomic DNA (gDNA) and complementary DNA (cDNA) were used as background DNAs for analysis. The PCR primer binding sites were searched against the whole background DNA with a k -mer-based algorithm and were sorted by the stability of the primer duplex. The searching results include both target amplicon and nonspecific potential amplicons. The main MFEprimer workflow consists of the below four steps:

1. Find the stable binding sites of the 3'-end of both forward primer and reverse primer among the target and background DNA sequences. In default, the stability of the last nine residues at the 3'-end of a primer was used to represent the primer 3'-end stability (*see Note 3*). The Nearest-Neighbor (NN) model [12] was used to calculate the stability (*see Note 4*). The unstable binding sites of the 3'-end will be filtered out directly and will not go through the following steps because only the primer duplex with stable 3'-end has the ability to start the extension reaction (*see Note 1*).
2. Calculating the binding stability of the entire primer sequence. The more stable the primer-template duplex, the higher the probability of the binding reaction.
3. Running a virtual PCR analysis and filtering out the unreasonable amplicons, such as the amplicon with 100,000,000 bp in size.
4. Output the predicted PCR amplicons. All the amplicons were sorted by the joint binding stability of both forward and reverse primers. The joint binding stability is the geometric mean of the binding stabilities of forward primer and reverse primer. The amplicons include not only the target amplicon but also the potential nonspecific amplicons.

3.2 Settings

Figure 3 shows the main features of the MFEprimer-2.0 home page. There are four sections in the home page:

1. *Background database selection*: This section is mandatory. Currently, there are 27 species databases available and some of them are added upon the request from the users, such as the dog's genomics DNA database [13].
2. *Input the primer sequences (5' → 3')*: This section is also mandatory and the primers should be in 5'-3' direction. Multiple primers are allowed and will be considered as either forward primer or reverse primer.
3. *Results filter settings*: This section is optional. Usually, MFEprimer-2.0 will output many predicted amplicons, including the ones that probably won't be amplified in real PCR. A very long list of amplicons will be useless, and users will still don't know how to evaluate the specificity of the primers. Results

filter settings can filter out these low probability amplicons. For example, in most cases, there are some predicted amplicons which have very low T_m values (*see Note 3*) and these amplicons would not be amplified in real PCR due to the T_m difference between these amplicons and the target amplicon. In this situation, the low T_m amplicons should be filtered out and let users focus on the amplicons that will be probably amplified (*see Note 5*).

4. *Experimental settings:* This section is optional. These settings are used by the nearest-neighbor (NN) model to calculate the melting temperature and Gibbs free energy. Different values of these parameters will result in different T_m values. We recommended the users to specify real PCR conditions to get the most accurate estimation of the T_m value.

4 How to Use MFEprimer to Check the Specificity of Primers

MFEprimer can check most types of primers, including conventional PCR primers, real-time PCR primers, degenerate primers, and multiplex PCR primers. Besides the primers, users should also provide the accurate background DNA, which is the DNA from the user's own sample. For example, if the user wishes to amplify a region of a human gene, the human gDNA and cDNA should be selected as the background DNA. However, if the gDNA was completely digested when preparing the sample during the experiment, then only the cDNA sequence database should be selected as the background DNA (*see Note 6*).

4.1 Preparing Primers

We have shown an example PCR primer pair, 5'-GACCAGTAGA CCTCGGGCGAA-3' and 5'-ACCTTGGCAAGTGCTCCTCT-3', to amplify and distinguish the two transcript variants (GenBank No. NM_010923 and NM_180960) of the mouse Nnat (GeneID: 18111) gene. The gene structure of these two transcript variants and the locations of the two primers are shown in Fig. 2 (*see Note 7*). And the figure is from the VizPrimer [14] web site (<http://biocompute.bmi.ac.cn/CZlab/VizPrimer>).

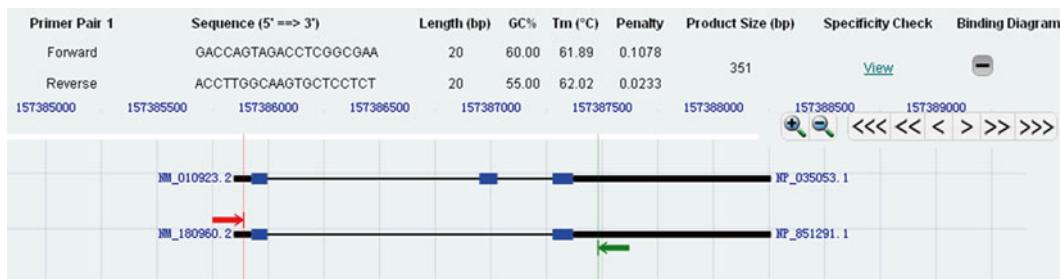


Fig. 2 The gene structure of two transcript variants of the mouse Nnat gene and the locations of the two primers

Background database selection: (Multiple database selection)

Mouse - RNA & Genomic ▾

Enter the primer sequences (5' → 3'): (Batch primer sequences)

Sequence 1: GACCAAGTAGACCTCGGGCGAA Clear
Sequence 2: ACCTTGGCAAGTGCTCCTCT Clear

Add + Del - Run Reset

Results filter settings [optional]:

Tm (°C): From: 10 To: 80.0 PPC (%): From: 10
Size (bp): From: 50 To: 2000

Experimental settings [optional]:

Concentration of monovalent cations (usually KCl, mM): 50.0 Concentration of dNTPs (mM): 0.25
Concentration of divalent cations (usually MgCl₂, mM): 1.5 Annealing oligo concentration (nM): 50.0

Fig. 3 Settings in the home page of MFEprimer-2.0

MFEprimer-2.0 Report

Home | Query | Histogram | Descriptions | Details | Parameters | Citation

Query

Query ID	Sequence (5' → 3')	Length	GC%	Tm
Seq1	GACCAAGTAGACCTCGGGCGAA	20	60.0	61.9
Seq2	ACCTTGGCAAGTGCTCCTCT	20	55.0	62.0

[View Hairpins & Dimers](#)

Fig. 4 The “Query” section of the MFEprimer-2.0 result page

4.2 MFEprimer Settings

As Fig. 3 shows, we first selected the “Mouse—RNA & Genomic” as the background DNA to avoid the sample from being potentially contaminated with the genomic DNA. Secondly, the primer sequences were written in 5' → 3' direction. In the “Results filter settings” section, we set more strict cutoff values for T_m “From” = 10 and primer pair coverage (PPC) “From” = 10. For other parameters we use the default value. Finally, we can click the “Run” button to launch the MFEprimer program.

4.3 The MFEprimer Results

The first line of the result page begins with the title “MFEprimer-2.0 Report,” followed by several shortcut links for each section of the result (shown in Fig. 4). Usually, the result page is long and these links can guide users to the right section quickly. The MFEprimer-2.0 result page contains six sections below.

1. *Query*: Fig. 4 shows the list of the input primer sequences with general information such as size, GC content, and T_m value annotation.
2. *Histogram of size, T_m , and ΔG of x potential amplicons*: This section provides histograms of the size, ΔG , and T_m when the number of predicted amplicons >10. The users can see what the worst-case scenario is for their primers and decide how to proceed (shown in Fig. 5).
3. *Descriptions of x potential amplicons*: Fig. 6 shows the brief description of all the potential amplicons predicted by MFEprimer-2.0. The amplicons in red are the ones that could be probably amplified in PCR.
4. *Amplicon details*: Information on the hybridization details for each predicted amplicon and the amplicon sequence. Figure 7 shows the first amplicon.
5. *Parameters*: The top part of Fig. 8, the parameters used during the evaluation process.
6. *Citation*: The bottom part of Fig. 8, these papers report the MFEprimer program.

Histogram of size, T_m and ΔG of 344 potential amplicons

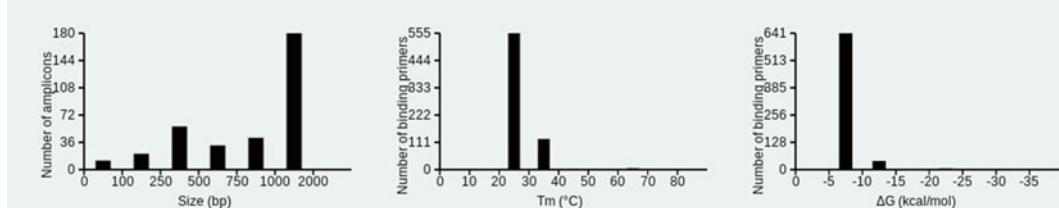


Fig. 5 The “Histogram of size, T_m , and ΔG of x potential amplicons” section of MFEprimer-2.0 result page

Descriptions of 344 potential amplicons

ID	Accession	Fp x Rp							What's this?		How to explain the result?	
			Size (bp)	PPC (%)	Fp T_m ($^{\circ}$ C)	Rp T_m ($^{\circ}$ C)	Fp ΔG (kcal/mol)	Rp ΔG (kcal/mol)	Fp 3' ΔG (kcal/mol)	Rp 3' ΔG (kcal/mol)		
1	NC_000068.6	Seq1 x Seq2	1632	100.0	61.9	62.0	-22.9	-22.6	-4.0	-2.9		
2	NM_010923.2	Seq1 x Seq2	432	100.0	61.9	62.0	-22.9	-22.6	-4.0	-2.9		
3	NM_180960.2	Seq1 x Seq2	351	100.0	61.9	62.0	-22.9	-22.6	-4.0	-2.9		
4	NC_000068.6	Seq2 x Seq2	92	-36.0	38.3	38.3	-11.6	-11.6	-2.9	-2.9		
5	NC_000067.5	Seq2 x Seq2	1083	-30.3	35.2	35.2	-10.8	-10.8	-2.9	-2.9		
6	NC_000067.5	Seq2 x Seq2	321	-26.5	30.7	33.0	-9.7	-10.1	-2.9	-2.9		
7	NC_000079.5	Seq2 x Seq2	302	-26.5	30.7	33.2	-9.7	-10.2	-2.9	-2.9		
8	NC_000074.5	Seq2 x Seq2	1874	-25.6	35.2	30.7	-10.8	-9.7	-2.9	-2.9		
9	NC_000070.5	Seq2 x Seq2	1110	-25.6	30.7	35.2	-9.7	-10.8	-2.9	-2.9		
10	NC_000069.5	Seq2 x Seq2	332	-25.6	35.2	30.7	-10.8	-9.7	-2.9	-2.9		
11	NC_000069.5	Seq2 x Seq2	252	-25.6	35.2	30.7	-10.8	-9.7	-2.9	-2.9		

Fig. 6 The part of “Descriptions of x potential amplicons” section of the MFEprimer-2.0 result page. This figure is reproduced from (MFEprimer-2.0: A fast thermodynamics-based program for checking PCR primer specificity 2012) with permission from Oxford University Press [2]

2. Analyzing the red amplicons. In most cases, the target amplicon is located in this red group and the real PCR conditions should be optimized for amplifying the target DNA. Other amplicons in red are the ones which have thermodynamics features close to the target amplicon and could be probably amplified in the PCR. Therefore, if there are other amplicons in red group, the users may run a multiple align [15, 16] or cluster [7] analysis with the amplicons to disclose the relationship between the target amplicon and these amplicons:
 - (a) If they are homologous sequences, such as transcript variants from one gene, then these amplicons should be regarded as specific amplicons.
 - (b) If they are not homologous sequences, these amplicons might usually be nonspecific amplicons, and the primer pair should be discarded. However, if we have to use the primer pair or the primer pair is quite difficult to redesign, a further analysis is required. In this situation, if the detection method for amplicons can distinguish the target and nonspecific amplicons, the primer pair is also acceptable. For example, if the user uses electrophoresis to detect the amplicons and none of them has similar size with the target amplicon, then the primer pair is acceptable. However, if we use hybridization (e.g., probe) to detect the amplicons, the primer pair would be accepted only when the probe hybrids to the target amplicon specifically.
3. Be careful to the rest of the amplicons. If there are too many predicted amplicons (*see Note 8*), then the primer pair should be redesigned as recommended [17]. On the other side, although these amplicons have lower binding energy (lower T_m and higher ΔG value), they also have the possibility to be amplified [18], especially when they have small sizes or the annealing temperature is low. As a result, our suggestion to avoid amplifying these amplicons is using a higher annealing temperature.

4.5 Analyzing the Example MFEprimer Results

Figure 6 shows the brief description of 344 potential amplicons predicted by MFEprimer-2.0 for the primer pair of the mouse Nnat gene. Following the steps described in the last section:

1. We can find the target amplicons: NM_010923 (ID: 2) and NM_180960 (ID: 3).
2. Besides the two target amplicons, there is another amplicon (ID: 1, GenBank No. NC_000068) with size of 1,632 bp in the red group. Multiple sequence alignment result shows that this amplicon contains the other two amplicons as shown in Fig. 9. The search result of NM_010923 on GenBank (http://www.ncbi.nlm.nih.gov/gene/?term=NM_010923.2) indicated that the sequence NC_000068 (ID: 1) is the genomic DNA sequence region of the mouse Nnat gene. Therefore, this amplicon should

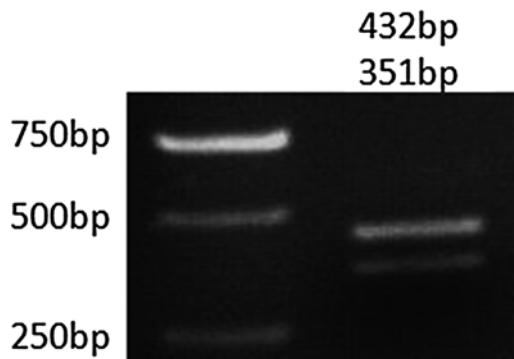


Fig. 10 Electrophoresis analysis of the two transcript variants of the mouse Nnat gene

not be considered as nonspecific amplicon. To be noticed that, if the sample was contaminated by genomic DNA, then the amplicon would be amplified.

3. Besides the amplicons in red font, there are also 343 amplicons predicted by MFEprimer-2.0. Compared to the target amplicon, these amplicons have significantly lower T_m values and higher ΔG values. Therefore, when we use an annealing temperature optimized for the target amplicon, these amplicons should not be amplified. In other words, we can use a higher annealing temperature to suppress the amplification of these nonspecific amplicons.
4. Directed by the MFEprimer-2.0 result, we experimentally validated the primer pair and obtained a high-quality result as shown in Fig. 10.

5 Notes

1. According to Rychlik's work [18], a continuous primer template duplex stable more than -11 kcal/mol will initialize the extension reaction. Therefore, if the stability of 3'-end of a primer duplex is equal or smaller than -11 kcal/mol, the duplex will probably be amplified. MFEprimer uses -9 kcal/mol as a strict cutoff value for safety consideration.
2. We would like to clarify here that the sequence alignment results (e.g., perfect match, mismatch) only indicate the nucleotide composition of the binding sites but not reflect the binding stability of the duplex. For example, the G-C and A-T matches have equal scores in the search results of National Center for Biotechnology Information's Basic Local Alignment Search Tool (NCBI BLAST) [7], but it is well known that the G-C match is more stable than the A-T match. Therefore, it is inappropriate to use NCBI BLAST or other similar tools [19] to check the specificity of the PCR primers.

3. The annealing temperature of a successful PCR is usually greater than 55 °C to suppress nonspecific PCR amplification. Therefore, primers with T_m values less than 55 °C would not be amplified in PCR reaction [2]. MFEprimer-2.0 uses a strict cutoff value of 48 °C to check the 3'-end stability of the primer duplex. To determine the maximum size of the 3'-end subsequence to represent its stability, we calculated all of the k -mer values, with k ranging from 5 to 9. The result shows that $k=9$ (the T_m of most stable 9-mer sequence is 48.4 °C) is suitable for the maximum size of the 3'-end subsequence. Moreover, the command-line version of MFEprimer-2.0 also supports custom k values, for example to set $k=8$ for a more strict 3'-end stability checking.
4. Although there are several methods available for predicting melting temperature, according to Panjkovich's [20] and Ahsen's [21] works, the most accurate T_m predicting method is the Nearest-Neighbor method [22].
5. We defined a parameter primer pair coverage (PPC) to score the ability of the primer pair (a forward primer coupled with a reverse primer) binding to the DNA template using the following formula (Fig. 11):

Where Fm and Rm are sequence overlaps of the forward primer and reverse primer with the DNA template, and Fl and Rl are the full lengths of the forward primer and reverse primer, respectively. $CVfr$ is the coefficient of variability of matched length of forward primer (Fm) and reverse primer (Rm). The maximum value of PPC is 100 %, indicating a pair of primers with the same length and both of them binding to the template completely.

6. It is recommended to select the “Genomic DNA” as background DNA to prevent the sample from being contaminated with genomic DNA.
7. Knowing the gene structure is essential for PCR primer design. For the gene has several transcript variants, it is important to know: (1) Which exon did the primers locate? (2) Can this primer pair distinguish these transcript variants? (3) Is the primer located at the junction site between the two adjacent exons?
8. According to Andreson's [17] work, the more the predicted amplicons, the larger the experimental failure rate. In detail, if the number of predicted amplicons is in 2–20, the failure rate is about 20 %; but if the number is in 10–100, the failure rate is about 40 %; and if the number is >10,000, the failure rate is >70 %.

$$\text{PPC} = \frac{Fm}{Fl} \times \frac{Rm}{Rl} \times (1 - CVfr)$$

Fig. 11 The formula of PPC

Acknowledgement

This work was supported by the National Basic Research Project (973 program) (2012CB518200), the General Program (31371345, 30900862, 30973107, 81070741, 81172770) of the Natural Science Foundation of China, the State Key Laboratory of Proteomics of China (SKLP-O201104, SKLP-K201004, SKLP-O201002), and the Special Key Programs for Science and Technology of China (2012ZX09102301-016).

References

- Qu W, Shen Z, Zhao D et al (2009) MFEprimer: multiple factor evaluation of the specificity of PCR primers. *Bioinformatics* 25(2):276–278
- Qu W, Zhou Y, Zhang Y et al (2012) MFEprimer-2.0: a fast thermodynamics-based program for checking PCR primer specificity. *Nucleic Acids Res* 40(W1):W205–W208
- Lexa M, Horak J, Brzobohaty B (2001) Virtual PCR. *Bioinformatics* 17(2):192–193
- Lexa M, Valle G (2003) PRIMEX: rapid identification of oligonucleotide matches in whole genomes. *Bioinformatics* 19(18):2486–2488
- Boutros PC, Okey AB (2004) PUNS: transcriptomic- and genomic-in silico PCR for enhanced primer design. *Bioinformatics* 20(15):2399–2400
- Andreson R, Kaplinski L, Remm M (2007) Fast masking of repeated primer binding sites in eukaryotic genomes. *Methods Mol Biol* 402:201–218
- Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402
- SantaLucia J Jr (2007) Physical principles and visual-OMP software for optimal PCR design. *Methods Mol Biol* 402:3–34
- SantaLucia J, Hicks D (2004) The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* 33:415–440
- Onodera K, Melcher U (2004) Selection for 3' end triplets for polymerase chain reaction primers. *Mol Cell Probes* 18(6):369–372
- Miura F, Uematsu C, Sakaki Y et al (2005) A novel strategy to design highly specific PCR primers based on the stability and uniqueness of 3'-end subsequences. *Bioinformatics* 21(24):4363–4370
- SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* 95(4):1460–1465
- Lindblad-Toh K, Wade CM, Mikkelsen TS et al (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069):803–819
- Zhou Y, Qu W, Lu Y et al (2011) VizPrimer: a web server for visualized PCR primer design based on known gene structure. *Bioinformatics* 27(24):3432–3434
- Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302(1):205–217
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797
- Andreson R, Möls T, Remm M (2008) Predicting failure rate of PCR in large genomes. *Nucleic Acids Res* 36(11):e66
- Rychlik W (1995) Priming efficiency in PCR. *Biotechniques* 18(1):84–86, 88–90
- Kent WJ (2002) BLAT—the BLAST-like alignment tool. *Genome Res* 12(4):656–664
- Panjkovich A, Melo F (2005) Comparison of different melting temperature calculation methods for short DNA sequences. *Bioinformatics* 21(6):711–722
- von Ahsen N, Wittwer CT, Schutz E (2001) Oligonucleotide melting temperatures under PCR conditions: nearest-neighbor corrections for Mg²⁺, deoxynucleotide triphosphate, and dimethyl sulfoxide concentrations with comparison to alternative empirical formulas. *Clin Chem* 47(11):1956–1961
- Crothers DM, Zimm BH (1964) Theory of the melting transition of synthetic polynucleotides: evaluation of the stacking free energy. *J Mol Biol* 9:1–9

INDEX

A

- Alignment 34, 40, 42–44, 57–71,
75, 91, 93–94, 99, 105, 107, 114, 133, 134,
144–146, 148, 153, 158, 180, 195, 199, 207–209

Allele-specific real time PCR 117–123

Asymmetric PCR 87

C

Consensus primers 74, 76–78

CTX-M beta lactamase 75

D

Degenerate PCR primers 103–115

Degenerate primers 73, 103–115, 143, 144, 203

Different subtypes of one same infectious agent 91

DNA

- isolation 145, 147, 148
- masking 3, 7
- polymerases 17, 18, 24, 29,
43, 47, 141, 142, 145, 149, 152, 159, 200, 201
- purification 145, 147
- repeats 3

E

Entropy 41, 59, 64–71, 135

F

Free on-line software 175

G

Gene Runner 144, 146

Genotyping 2, 10, 43, 44, 117, 118

H

Hierarchical clustering 112

Hybridization probes 82–84

Hydrogen bonding 27

I

Inclusivity 58, 69

Influenza 57–71

Initiator primers 73–88

L

LATE-PCR 74, 76, 80, 82, 84, 86, 87

M

Microarrays 39, 152, 180

Mixed population PCR 103

Molecular biology 19, 42,
69, 76, 141, 144, 151

Molecular diagnostics 73

Multiple pathogens 91–101

Multiplex PCR 18, 19, 49, 91–101, 126,
163, 165, 169, 203

N

Nucleic acid sequence diversity 73

O

Oligonucleotide 18, 39, 41, 43,
58, 76, 92, 98, 117, 118, 126, 171, 180, 142, 143

Oligo primer analysis software 91

P

PCR primer 1–3, 5, 6, 8,
13, 14, 20, 26, 28, 29, 38, 39, 76, 91–101, 126,
127, 131, 136, 141–149, 201–203, 205, 210

Polymerase chain reaction 17–29,
57–71, 91–101, 117–123, 141–149, 199–210

Population stratification 103

Primer-BLAST tool 92, 94, 96, 99, 113

Primer design 1, 2, 5, 6, 8,
9, 13, 17–29, 34, 37–40, 42, 45, 50, 77–80, 83, 84,
86, 92, 98–100, 103–115, 118, 122, 126,
131–132, 142–146, 148, 151–161, 164, 172–175,
180–185, 188–190, 193, 195

Primer express 19, 38,
41, 151–161

Primer 3 19, 24, 98–99, 146, 172

Probe 31–53, 74, 76,
80–82, 84, 88, 154, 155, 171, 180–183, 186

Q

Quantitative real-time PCR (qPCR) 31–53, 57,
69, 151, 152, 159, 160, 171–177

R

- RAPD 145, 147–148
Reverse transcription real-time PCR 50
RNA 17, 18, 31, 32, 49–52, 59,
75, 76, 99, 101, 146, 151, 152, 161, 195, 204

S

- Short tandem repeat (STR) 18, 19,
21–23, 27, 28, 165
Simultaneous detection 91–101
Single nucleotide polymorphisms (SNP) 2, 18,
19, 21, 25–27, 41, 42, 117, 119, 163–168
Sizes of amplicons 91
SNaPshot 154, 155, 157,
163–169, 183–187, 190–192
SNP. *See* Single nucleotide polymorphisms (SNP)
Software and databases 31

STR. *See* Short tandem repeat (STR)

SYBR Green primers 171–175

T

- Targeted resequencing 103
 T_m values of primer sets 96, 97, 100, 210

U

Urate transporter 117–123

V

- Validation 32, 40, 42,
44–49, 51–53, 59, 61–64, 70, 122, 141–149, 182
Viral genome amplification 104

W

Web Primer 143–146