PrimerBank: a resource of human and mouse PCR primer pairs for gene expression detection and quantification

Athanasia Spandidos\textsuperscript{1,2}, Xiaowei Wang\textsuperscript{1,2}, Huajun Wang\textsuperscript{1,2} and Brian Seed\textsuperscript{1,2,*}

\textsuperscript{1}Center for Computational and Integrative Biology, Massachusetts General Hospital and 
\textsuperscript{2}Department of Genetics, Harvard Medical School, 185 Cambridge Street, Boston, MA 02114-2790, USA

Received July 27, 2009; Accepted October 16, 2009

ABSTRACT

PrimerBank (http://pga.mgh.harvard.edu/primerbank/) is a public resource for the retrieval of human and mouse primer pairs for gene expression analysis by PCR and Quantitative PCR (QPCR). A total of 306,800 primers covering most known human and mouse genes can be accessed from the PrimerBank database, together with information on these primers such as \( T_m \) location on the transcript and amplicon size. For each gene, at least one primer pair has been designed and in many cases alternative primer pairs exist. Primers have been designed to work under the same PCR conditions, thus facilitating high-throughput QPCR. There are several ways to search for primers for the gene(s) of interest, such as by: GenBank accession number, NCBI protein accession number, NCBI gene ID, PrimerBank ID, NCBI gene symbol or gene description (keyword). In all, 26,855 primer pairs covering most known mouse genes have been experimentally validated by QPCR, agarose gel analysis, sequencing and BLAST, and all validation data can be freely accessed from the PrimerBank web site.

INTRODUCTION

Quantitative Polymerase Chain Reaction (QPCR) has become a commonly used method for precise determination of gene expression and evaluating DNA microarray data (1,2). The main advantages of this technique are its unparalleled dynamic range, being able to detect \( >10^7 \) fold differences in expression, and the potential to amplify very small amounts of DNA template, down to a single copy (3–5). QPCR products can be detected by two general methods: one utilizing various types of fluorescence containing hybridization probes (6–20) and the other utilizing SYBR Green I dye fluorescence (21–23). Hybridization probes are designed to be target specific and can thus minimize nonspecific amplification, but can be difficult to design and costly (5). The SYBR Green I method is the most simple and inexpensive QPCR method and has become the most commonly used for gene expression analysis (21,22). SYBR Green I dye intercalation into double-stranded DNA, such as PCR products, results in detectable fluorescence, corresponding to the amount of PCR product generated in each cycle (23). QPCR amplification plots provide information for relative quantification between samples and on the amount of initial DNA template (24–26). Dissociation curves, generated after the QPCR step, can give information on the specificity of the reaction (27).

We have developed a database, named PrimerBank (http://pga.mgh.harvard.edu/primerbank/), for the retrieval of human and mouse primer pairs for gene expression analysis by PCR and QPCR. PrimerBank primers can work with SYBR Green I detection methods and the primer design was based on an algorithm that had been previously used for oligonucleotide probe design for DNA microarrays (28). Nonspecific amplification of nontarget sequences is a common problem encountered in PCR and QPCR experiments. So, for the PrimerBank primer design, various filters for cross-reactivity were used to reduce nonspecific amplification (29). Furthermore, all primers have been designed to work under a high annealing temperature of 60°C. At least one primer pair represents each gene, and in many cases alternative primer pairs have been designed. See Table 1 for information on primers contained in PrimerBank. In addition, we have previously experimentally validated 26,855 primer pairs, which cover most known mouse genes (30). We found that PrimerBank primers can amplify specifically the genes for which they have been designed (82.6% success rate based on...
Table 1. Primers that can be retrieved from PrimerBank

<table>
<thead>
<tr>
<th>Number of primers</th>
<th>Number of genes covered</th>
<th>Organism</th>
<th>Number of experimentally validated primer pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>306800</td>
<td>61425</td>
<td>All organisms</td>
<td>26855</td>
</tr>
<tr>
<td>138918</td>
<td>27684</td>
<td>Mus musculus</td>
<td>26855</td>
</tr>
<tr>
<td>167882</td>
<td>33741</td>
<td>Homo sapiens</td>
<td></td>
</tr>
</tbody>
</table>

Primers stored in the PrimerBank database have been designed to represent most known human and mouse genes. A total of 26855 primer pairs, representing 27684 mouse genes, have been experimentally validated by QPCR, agarose gel electrophoresis, sequencing and BLAST and all experimental validation data can be viewed from the PrimerBank web site.

visualization of one band of the expected size by agarose gel analysis). The reproducibility of the QPCR technique and the uniformity of amplification using PrimerBank primers were also analyzed (30). Furthermore, the amplification efficiency of the QPCR using PrimerBank primers was determined and it was found that for 13 primer pairs tested it ranged from 79% to 96%, using an analytical method. The same 13 PrimerBank primer pairs as above were used and one-way ANOVA (ANalysis Of VAriance) analysis was done using each primer pair in a series of titration QPCRs of template DNA, in order to determine if amplification efficiencies were similar between different PrimerBank primer pairs. The efficiencies were found to be similar between these primer pairs ($P = 0.7338$ i.e. $P > 0.05$) (30). Since PrimerBank primers have been designed to be used under the same annealing temperature, high-throughput QPCR in parallel is facilitated, as an alternative approach to DNA microarrays (31,32) for the study of gene expression.

PRIMER DESIGN

Oligonucleotide probe sequence design for DNA microarrays has become the subject of many studies using a number of algorithms (33–46). Most of these algorithms use BLAST (47) to identify regions of the gene from which oligonucleotide sequences can be selected (48). PCR primer design can be based on these algorithms, since BLAST is used in design of both DNA microarray probes and PCR primers. The PrimerBank primer design was based on a successful approach that had been previously used for the prediction of oligonucleotide probes for DNA microarrays (28). However, the PrimerBank primer design differs by the addition of filters that are considered to be important for primer specificity (29).

All gene sequence information was obtained from the NCBI protein database (http://www.ncbi.nlm.nih.gov/entrez/) (49). DNA-coding sequences were retrieved and the redundant sequences were clustered using the DeRedund program (28). Low-complexity regions, which may contribute to primer cross-reactivity (50), were excluded using the program DUST (51). If primers contained six or more identical contiguous bases they were rejected, so that more complicated sequences could be chosen. Furthermore, no primers were selected from low-quality regions of sequence (29). Primers were designed to represent at least once each gene, and most known human and mouse genes were covered. See Table 2 for the statistics of primer pair design with respect to gene representation.

In many cases, coding regions were scanned from the 5'- to the 3'-end until three suitable primer pairs were found (in these cases the PrimerBank IDs of the primers contain ‘a1’, ‘a2’ or ‘a3’, the ‘a1’ primer pair being most 5’ and the ‘a3’ being most 3’). Two general methods can be used for cDNA library preparation: the oligo(dT) and random priming methods. Oligo(dT) priming during cDNA preparation can result in reduced coverage of the 5'-end of sequences, since some 3' UTRs can be very long (3). Random priming can result in the highest coverage of the 5'-end and this method was used for our cDNA preparations (3,30). Because of this higher coverage at the 5'-end, the most 5'primers were experimentally validated (see ‘Database generation and content’ section below). Also, primers were designed irrespective of their location on exons. In order to prevent any nonspecific amplification of any contaminating genomic DNA, primers can be designed to be located on exon boundaries; however, in many cases it was not possible to design primers located on exon boundaries that fulfilled all the design criteria, since some transcripts consist of a single exon. See Table 3 for the statistics of primer location with respect to exons.

PrimerBank primers have been designed to have uniform length and GC% properties (29). All PrimerBank primers are 19–23nt, with a preferred length of 21nt. This length is optimal for gene-specific sequences and minimizes cross-reactivity. Also, this length is optimized to reduce costs if primers are synthesized in large sets. Primers have similar GC% from 35% to 65% in order to ensure uniform priming. The algorithm used for primer design also evaluated the ΔG value for the last

Table 2. PrimerBank mouse primer pair design and validation

<table>
<thead>
<tr>
<th>Mouse primer pairs or genes</th>
<th>Number of mouse genes/primer pairs validated</th>
<th>Percentage (%) of total primer pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of primer pairs</td>
<td>26855</td>
<td>100</td>
</tr>
<tr>
<td>Total number of genes represented</td>
<td>27684</td>
<td></td>
</tr>
<tr>
<td>Total number of genes not represented</td>
<td>1165</td>
<td></td>
</tr>
<tr>
<td>Primer pairs with no redundancy</td>
<td>23700</td>
<td>88.2</td>
</tr>
<tr>
<td>Primer pairs with two target genes</td>
<td>2534</td>
<td>9.4</td>
</tr>
<tr>
<td>Primer pairs with &gt;2 target genes</td>
<td>621</td>
<td>2.3</td>
</tr>
<tr>
<td>Total number of successful primer pairs based on all validation criteria</td>
<td>17483</td>
<td>65.1</td>
</tr>
<tr>
<td>Total number of successful primer pairs based on agarose gel electrophoresis</td>
<td>22189</td>
<td>82.6</td>
</tr>
<tr>
<td>Total number of successful primer pairs based on BLAST analysis</td>
<td>19453</td>
<td>72.4</td>
</tr>
<tr>
<td>Total number of failed primer pairs by QPCR (due to no amplification)</td>
<td>1745</td>
<td>6.5</td>
</tr>
</tbody>
</table>

A total of 26855 primer pairs were synthesized, which correspond to a higher number of 27684 mouse gene targets since some of these primer pairs amplify the same sequence from two genes or gene isoforms. Primers were not designed for another 1165 mouse genes, mainly because of low sequence quality. The average mouse gene has 1293 bp; however, the average length for these genes is 435 bp and most are ‘unknown’ or RIKEN sequences.
its complementary sequence (29). A BLAST similarity search for the primer sequence was carried out on the complementary strand and the score was required to be <18. To prevent primer dimers, primers were rejected if the four residues at the 3′-end of the primer were found in its complementary sequence. Complementarity in the forward and reverse primers in a pair was also evaluated in order to prevent heterodimer formation.

The filters used for primer selection were very stringent and this was reflected by the fact that 99.5% of primers were rejected. Of the rejected primers, 50.7% had too high or too low T_m values, 28.7% cross-hybridized to nontarget sequences, 19.8% were rejected because of sequence self-complementarity, 0.5% were from low-complexity regions and 0.3% were rejected because of other properties (GC content and end stability).

### DATABASE GENERATION AND CONTENT

We have stored the primer sequences in the PrimerBank database (http://pga.mgh.harvard.edu/primerbank/) together with other information about the primers such as T_m, location on the transcript and expected amplicon size. Furthermore, we have experimentally validated 26855 of these primer pairs, corresponding to 27684 mouse transcripts, by QPCR, agarose gel electrophoresis, sequencing and BLAST analysis (30). Random priming was used to prepare the cDNA library from a commercial universal mouse composite total RNA preparation, which contains RNA from a panel of 11 different mouse cell types for a good representation of the majority of mouse genes. The cDNA library prepared was used as the template in all QPCRs for the high-throughput validation and the same PCR conditions were used for all reactions. QPCR amplification plots and dissociation curves were analyzed and the presence of a single band of the expected size was determined by agarose gel analysis. Sequences obtained from these PCR products were BLAST analyzed as batch sets against the NCBI database (47). For the identification of successful samples, the main parameters considered were the alignment length of the PCR product sequence to the expected, the expected sequence match position to the sequence returned by NCBI BLASTn and the percent identity of the two sequences. The success rate of the high-throughput PrimerBank primer validation experiments was high, observed both from agarose gel and BLAST analysis, and 17483 primer pairs were found to be successful based on all validation criteria. See Table 2 for data obtained from the high-throughput experimental validation procedure.

Because of no amplification, 1745 samples (6.5%) failed. Several were found to belong to olfactory receptors, vomeronasal receptors, transcription factors and low abundance transcripts while others were of unknown function or RIKEN sequences. To determine whether the templates for the failed primer pairs were present in the cDNA sample used, we tested these primers using genomic DNA template (30). From these experiments, we found that these primer pairs had originally failed in most
cases because their respective templates were not present or present in very low amounts in the source of DNA template used and not because of poor primer design. Specific tissues may be used as sources of cDNA templates where expression of the genes of interest is known, for increased amplification success. Furthermore, we determined the uniformity of amplification using fully validated PrimerBank primer pairs i.e. primer pairs that had been successful in all steps of the validation procedure and found that amplification using PrimerBank primers was relatively uniform (30).

WEB INTERFACE

Search tools

The PrimerBank database can be searched for primers for a gene of interest using any of the following search terms: GenBank accession number, NCBI protein accession number, NCBI gene ID, PrimerBank ID, NCBI gene symbol or keyword (gene description). Search results show the primer sequences and some information on the primers, such as the expected amplicon size and $T_m$. The cDNA and amplicon sequences as well as the experimental validation data can be viewed from the PrimerBank search result web pages, by clicking on the appropriate links. All validation data can be retrieved from the PrimerBank web site, since the criteria of the users for success or fail may be different from our validation criteria. In addition, users can use a BLAST tool that can be found on the PrimerBank homepage, to find any primers contained in the PrimerBank database that would amplify their sequence of interest. Users can also BLAST analyze the sequence obtained from the PCR product generated during the high-throughput experimental

![PrimerBank search results for beta-actin primers](image-url)

Figure 1. PrimerBank search results for beta-actin primers. The primer search function can be found on the PrimerBank homepage. The database was searched for mouse beta-actin primer pairs by PrimerBank ID (6671509a1) and the search results obtained are shown here. The primer sequences, lengths, $T_m$s, location of primers on the amplicon and expected amplicon size can be seen.
The following primer pair is found for 6671509a1

**Gene Descriptions:**
- **GenBank Accession**: NM_007393
- **NCBI Protein Accession**: NP_031419
- **Species**: Mouse
- **Coding DNA Length**: 1128
- **Gene Description**: actin, beta, cytoplasmic; A-X actin-like protein; melanoma X-actin [Mus musculus].

**Primer Pair Descriptions:**
- **PrimerBank ID**: 6671509a1
- **Validation Results**: [Click here to view experimental validation data: amplification plots, dissociation curves, 2% agarose gel analysis, sequencing and BLAST data]

**Amplicon Size**: 154

**Location in Coding Sequence** (primers and amplicon highlighted)

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence (5' -&gt; 3')</th>
<th>Length</th>
<th>Tm</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward Primer</td>
<td>GGCTGTATCCCCCTCCATCG</td>
<td>20</td>
<td>61.8</td>
<td>84-103</td>
</tr>
<tr>
<td>Reverse Primer</td>
<td>CCAGTTGTTAACAATGCCATGT</td>
<td>22</td>
<td>61.1</td>
<td>237-216</td>
</tr>
</tbody>
</table>

**Figure 2.** cDNA and amplicon sequence for beta-actin primers. The full cDNA and amplicon sequences as well as the highlighted location of primers on the amplicon are seen here and can be viewed from PrimerBank by clicking on the cDNA and amplicon sequence link found on the primer information web page (seen in Figure 1).
validation procedure by using a BLAST tool to query the NCBI database, which can be found on the validation data web page.

Sample search results

The results obtained from a primer pair search (mouse beta-actin primer pair; PrimerBank ID: 6671509a1) can be seen in Figures 1–3, as an example. Primer sequences for mouse beta-actin can be viewed together with information on the primer Tm’s and expected amplicon size (Figure 1). Users can click on the cDNA and amplicon sequence link shown on the primer information web page to view the full cDNA sequence and highlighted location of primers on this (Figure 2). In order to view the experimental validation data, users can click on the validation results link shown on the primer information web page and on the cDNA and amplicon sequence web page. On the experimental validation web page, the QPCR amplification plot, followed by the dissociation curve and agarose gel data can be seen (Figure 3). The sequence obtained follows and it is possible for users to scroll through this to view it in its entirety. A summary of the BLAST results obtained is shown below the sequencing result, including the percent (%) identity of the sequence obtained to the expected sequence, the match length of the two sequences and the match position of the expected sequence out of the total number of sequences to which the queried sequence matched to (Figure 3).

Primer statistics

Users can click on the ‘Primer Statistics’ tab found on the PrimerBank homepage to view some statistics of the primers currently contained in the database.

Protocols

The QPCR and reverse transcription protocols used for the high-throughput primer validation procedure can be found on the PrimerBank web site, as well as a troubleshooting guide, under the ‘PCR Protocol’ tab. These protocols may be used with all PrimerBank primer pairs, unless specific protocols are required.

Comments

Users can provide their comments on the PrimerBank web site or primer design by clicking on the ‘Comments’ tab seen on the homepage and filling out the ‘Comments’ box. Also, users can input information regarding their experimental use of the validated primer pairs and add any comments in a feedback table that can be seen on the validation data web page.

Primer submission

Users can recommend their own primer pairs for human and mouse genes by clicking on the ‘Primer submission’ tab.
tab seen on the homepage. In order to do this, users must provide their name, name of institution, email address and a password (optionally) the first time when they submit primers. If the submitted primers conform to the PrimerBank standards they will be added to the database.

DISCUSSION

A large number of tools and resources are available that can be used for designing PCR primers for various applications (56–67). In addition, databases exist that contain primers for PCR and QPCR, which have been submitted by researchers, but only a few thousand of these primer pairs have been experimentally validated (68,69). Also, primers contained in these databases have not been designed to work under the same PCR conditions. Therefore, it would be required to validate the primers for the gene(s) of interest and optimize the PCR conditions in order to use these primers. The PrimerBank database contains experimentally validated primers for most known mouse genes (17 483 primer pairs were successful based on all validation criteria) and all primers work under the same PCR conditions. PrimerBank thus provides a resource for researchers who are interested in validating DNA microarray data or high-throughput PCR.

FUNDING

National Institutes of Health Program for Genomic Applications (grant U01 HL66678). Funding for open access charge: Center for Computational and Integrative Applications (grant U01 HL66678). Funding for open access charge: Center for Computational and Integrative Applications (grant U01 HL66678).

REFERENCES


