Large-scale RT-PCR recovery of full-length cDNA clones

Jia Qian Wu, Angela M. Garcia, Steven Hulyk, Anna Sneed, Carla Kowis, Ye Yuan, David Steffen, John D. McPherson, Preethi H. Gunaratne, and Richard A. Gibbs

INTRODUCTION

Direct study of cDNAs has an important role for the analysis of genome sequences (1,2). The most significant contributions for discovering novel transcripts and precisely defining the structure of known genes have been from expressed sequence tag (EST) and full-length cDNA sequence analysis. When a cDNA sequence is known with high confidence, it can also be used for validating the genomic sequence and for characterizing alternative gene structures and heterogeneous transcripts that differ from known sequences.

There are multiple efforts that aim to capture the sequence of full-length clones that can be directly obtained from the cDNA libraries made from mammals and other select organisms, such as zebrafish, Drosophila, and Caenorhabditis elegans (3–8). Among these, the Mammalian Gene Collection (MGC; http://mgc.nci.nih.gov) is distinguished by a commitment to extremely high clone and sequence quality, as well as aiming to rescue at least one representative transcript from each human gene that codes for a protein (9,10). Currently, a nonredundant set of >10,000 human and >9000 mouse full-length open reading frame (ORF) clones have been sequenced and verified by the MGC program (10). This is an impressive start. However, there are still a large number of genes missing in the collection, as it is estimated that there are 25,000–30,000 protein-coding genes comprising the mammalian transcriptome. In part, this reflects the chosen approach of relatively shallow sampling from multiple clone libraries. It also reflects the biological reality that many genes are unlikely to be found except at very low levels in obscure cell types.

Many of the cDNA clones missing from the central collections are at least partially characterized in different databases or publications (11,12). Therefore, we elected to develop scalable reverse transcription PCR (RT-PCR)-based methods for cDNA clone rescue utilizing these existing sequences. A particular focus has been those clones that were difficult to recover from previous large-scale efforts. These procedures include simplified reliable PCR primer design, stockpiling RNA and tissue pools, accurate and robust 96-well format RT-PCR, streamlined sequencing methods, and custom data management strategies. From the initial results, we have observed many discrepancies between the sequences of the transcripts that we have recovered and the corresponding sequences in the databases, and these lead to interesting insights into the heterogeneity of the mammalian transcriptome. Overall, the ease and efficiency of the RT-PCR cDNA clone rescue pipeline suggests that this method could economically replace traditional cDNA library cloning in any situation in which sufficient sequence is known to design full-length PCR primers. Therefore, it is especially useful for the discovery of novel genes through gene prediction methods. Establishing this platform is an important step towards the development of a complete set of reagents representing the mammalian transcriptome.

MATERIALS AND METHODS

Target Selection and Large-Scale High-Throughput Primer Design

Target cDNAs, for which there were no publicly available clones, were extracted from GenBank® [thanks to Lukas Wagner, National Center for Biotechnology Information (NCBI), National Institutes of Health (NIH)].
Each of the clones had one form in public databases, which we downloaded for local use. Each of the sequences contains putative complete ORFs, but not necessarily additional untranslated region (UTR) sequences.

Algorithms were developed to design primers within 5′ and 3′ UTR sequences in order to amplify the entire ORF of each gene. When no UTR sequences were available, the primers were positioned at the very beginning of the ORF in order to include the coding region if possible.

The algorithm we designed first attempted to select the primers with the theoretical best characteristics based on Restricted Criteria, the most stringent requirements in terms of melting temperature (T_m), GC content, and secondary structures. If no such primers were available in this region, the program adopted less stringent standards, or Relaxed Criteria, to select the “next best” primers. If both attempts failed, the primer pair was located adjacent to the border of the ORF with minimum requirement of sequence composition based on Forced Criteria. The details of the design criteria are shown in Table 1.

### Table 1. Primer Design Criteria

<table>
<thead>
<tr>
<th>Primer Design Criteria</th>
<th>Restricted Criteria</th>
<th>Relaxed Criteria</th>
<th>Forced Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primer location</strong></td>
<td>≥80 bp away from the ORF</td>
<td>≥80 bp away from the ORF</td>
<td>Force primers to flank the ORF¹</td>
</tr>
<tr>
<td><strong>Whole ORF</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>All or most of the ORF</td>
</tr>
<tr>
<td><strong>GC content</strong></td>
<td>35%–75%</td>
<td>30%–75%</td>
<td>N.A.</td>
</tr>
<tr>
<td><strong>Length</strong></td>
<td>25–30 bp</td>
<td>20–30 bp</td>
<td>N.A.</td>
</tr>
<tr>
<td><strong>T_m range</strong></td>
<td>65°–80°C</td>
<td>60°–80°C</td>
<td>60°–80°C</td>
</tr>
<tr>
<td><strong>Hairpin structure</strong></td>
<td>Avoid HP ≥ 3 bp</td>
<td>Avoid HP ≥ 4 bp</td>
<td>N.A.</td>
</tr>
<tr>
<td><strong>Homology with other sequences</strong></td>
<td>Avoid</td>
<td>Avoid</td>
<td>N.A.</td>
</tr>
<tr>
<td><strong>Homopolymeric runs</strong></td>
<td>Avoid ≥ 3 bp</td>
<td>Avoid ≥ 4 bp</td>
<td>N.A.</td>
</tr>
<tr>
<td><strong>GC clamp</strong></td>
<td>Avoid</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

ORF, open reading frame; T_m, melting temperature; HP, hairpin; N.A., not applicable; UTR, untranslated region.

¹Primers are located 80 bp to the ORF, or position 1 to 27 of the Refseq sequence if 5′ UTR < 80 bp, or the last 27 bp of the RefSeq sequence if the 3′ UTR < 80 bp.

Each of the clones had one form in public databases, which we downloaded for local use. Each of the sequences contains putative complete ORFs, but not necessarily additional untranslated region (UTR) sequences.

Algorithms were developed to design primers within 5′ and 3′ UTR sequences in order to amplify the entire ORF of each gene. When no UTR sequences were available, the primers were positioned at the very beginning of the ORF in order to include the coding region if possible.

The algorithm we designed first attempted to select the primers with the theoretical best characteristics based on Restricted Criteria, the most stringent requirements in terms of melting temperature (T_m), GC content, and secondary structures. If no such primers were available in this region, the program adopted less stringent standards, or Relaxed Criteria, to select the “next best” primers. If both attempts failed, the primer pair was located adjacent to the border of the ORF with minimum requirement of sequence composition based on Forced Criteria. The details of the design criteria are shown in Table 1.

### cDNA Synthesis from the Pooled Tissues

For use in human cDNA clone rescue, equal amounts of total RNA were pooled from 20 human tissues, including adrenal gland, bone marrow, brain cerebellum, brain (whole), fetal brain, fetal liver, heart, kidney, liver, lung, placenta, prostate, salivary gland, skeletal muscle, spleen, testis, thymus, thyroid gland, trachea, and uterus (all from Human Total RNA Master Panel II; BD Biosciences Clontech, Palo Alto, CA, USA). A second human total RNA “supplemental” pool containing samples from human breast, colon, pancreas, and stomach (all from Ambion, Austin, TX, USA) was also used in later experiments. For mouse cDNA clone rescue, equal amounts of total RNA were pooled from 15 mouse tissues including an 11-day embryo, a 15-day embryo, a 17-day embryo, brain (whole), heart, liver, lung, prostate, salivary gland, smooth muscle, spleen, stomach, testis, thymus, uterus (all from Mouse Total RNA Master Panel; BD Biosciences Clontech). Reverse transcription reactions (13) were performed in bulk and then distributed for the subsequent PCR amplifications. A total of 4 μg RNA was used in a final volume of 20 μL reverse transcription reaction (200 ng/μL). Reverse transcription reactions were primed by oligo(dT) using 200 U SuperScript™ II Reverse Transcriptase in a 20-μL reaction (Invitrogen, Carlsbad, CA, USA). In order to recover longer transcripts efficiently, we included the proof reading enzyme Pfu (5 U) and 10% dimethyl sulfoxide (DMSO) in the reverse transcription reaction (14).
**RT-PCR Amplification of Specific Genes**

The reverse transcription reactions were followed by touchdown PCR (15) combined with an “autosegment extension” PCR program. One micro-liter reverse transcription reaction from the above was used in 50 μL of PCR. Annealing temperatures were initially set at approximately 3°–10°C above the calculated primer Tₘ. During the following 5–7 cycles, the annealing temperature was reduced by 2°C each cycle, until it reached a temperature approximately 2°–5°C below the calculated primer Tₘ. The rest of the amplification cycles use this temperature. Using the autosegment extension PCR program, PCR extension times were extended by 15 s each cycle, starting at the 15th cycle.

Specific primers were used to amplify the entire coding region of each gene. Twenty-eight cycles of PCR were performed using first strand cDNA as template and the Advantage™ 2 PCR Enzyme System (BD Biosciences Clontech). To reduce mutations introduced by polymerase misincorporation, the majority of the PCR products after 28 cycles were separated and saved for cloning. A small aliquot of the PCR products (10 μL) was amplified for 35 PCR cycles to produce an adequate amount of DNA to be visualized on an agarose gel. Kodak® 1D Image Analysis Software v. 3.0.1 (Eastman Kodak, Rochester, NY, USA) was used to estimate the product sizes. The observed sizes were compared with the expected sizes, and data were recorded into a database. For genes that failed to yield a PCR product, RT-PCRs were repeated either using alternative tissue sources, by increasing the number of PCR cycles to 45, or by using redesigned primers.

**Cloning PCR Products and Prescreening for the Inserts Size and Identity**

All PCR products were cloned utilizing the PCR-Script® Amp Cloning Kit (Stratagene, La Jolla, CA, USA) as follows. PCR products were purified with the QIAquick® 96-well PCR purification kit (Qiagen, Valencia, CA, USA). The purified PCR products were then “end-repaired” and inserted into the PCR-Script Amp cloning vector. Transformation was performed with XL10-Gold® Kan ultracompetent cells (Stratagene) in 96-well format. β-Galactosidase activity screen was used to distinguish vectors without insert. A maximum of 12 white colonies of each gene were picked for DNA preparation. The DNA of each subclone was digested with EcoRI and NotI restriction endonuclease and analyzed by agarose gel electrophoresis in order to determine the approximate size of the insert. The digestion data were captured in a database established for clone management and internal tracking of clones in the pipeline. Band calling and matching software were integrated to develop lists of genes and subclones to be passed onto the next step. Six subclones with inserts of expected sizes were then end-sequenced to verify their identity before attempting full-length sequencing.

**Full-Length Sequencing by Primer Walking**

Standard oligonucleotide walking techniques were used to sequence multiple subclones obtained from the RT-PCR procedure (16). Software has been developed to manage the task of intelligent sequencing primer design, and the primers were synthesized based upon the expected sequence from the database. Forward and reverse primers were designed every 350 bp to cover the entire length of the cDNAs. When a complete and “acceptable” (see definition of acceptable in the section entitled Extensive Transcript Heterogeneity and Clone Acceptability) subclone was identified, sequencing on all remaining subclones from that targeted gene was terminated. The resulting sequence reads were assembled using the Phred/Phrap computer programs (17) and edited via the Consed editor (18). The known sequence from the target gene was incorporated to facilitate alignment of the individual subclone reads during assembly.

**Finishing and Sequence Analysis**

After assembly, additional sequencing and editing were required to ensure the accuracy of the data. Low-quality regions or unsure bases were confirmed through a minimal number of primer walks, unless an obvious error was found within the ORF, in which case the sequencing of that subclone was terminated. Each subclone was sequenced to the standard of an accuracy of less than one sequencing error expected per 10,000 bases.

**RESULTS**

We have developed a pipeline for the rescue of functional cDNA clones from mammalian genes that were difficult to recover using other methods. The overview for the high-throughput RT-PCR rescue is shown in Figure 1. The primary areas developed to enable the pipeline are primer design, RT-PCR, cloning, and sequence evaluation. We have generated more than 570 cDNA clones
from humans and mice, and the recovered transcripts show extensive transript heterogeneity when compared to RefSeq and genomic sequences.

**Automated High-Throughput Primer Design**

An automated program we developed has designed primers for over 10,000 genes. Efficient primer design at this scale was a significant challenge at the initiation of this project (see Discussion). In general, the task of primer design is a balance between the placement of the primer position relative to the ORF and the constraints imposed by the composition of the underlying sequence. The software is now able to generate PCR primer designs for 99% of the candidate sequences. Our primers were located in the 5’ or 3’ UTR adjacent to the ORF when possible to amplify complete ORF (Figure 2A). The primers were generated using three different stringency levels of design criteria resulting in a variation in the characteristic of primers. As expected, the restrict/restrict combination appeared to be the most successful. Only minimal differences were observed, however, between the efficiencies of primer pairs produced by other combinations of the three-design criteria (Figure 2B).

**Reverse-Transcription Reaction and PCR Amplification**

Our gene rescue targets were those known genes not recovered via previous cDNA sequencing efforts. In order to increase the chance of recovering these genes, our RNA source was a mixture of 20 human or 15 mouse tissues, which represented the expression spectrum of the majority of gene targets. PCR conditions were optimized to increase yield in a 96-well format without introducing extra mutations due to polymerase infidelity. This included adjusting the concentration of various reagents, using minimized denaturing times, and varying annealing temperature, extension temperature, and extension time. Combinations of commercially available reverse transcription and PCR enzymes were procured and tested. The yields ranged from 15% to 50%. We chose SuperScript II reverse transcriptase and Advantage 2 PCR Enzyme System for their high yield and relative accuracy. Using these optimized conditions, RT-PCR was at-

---

**Figure 1. High-throughput RT-PCR gene recovery pipeline.** At several points in the pipeline, a failure invokes an alternative protocol to provide materials to reenter the pipeline. RT-PCR, reverse transcription PCR.
tempted on a total of more than 2800 genes, approximately 65% (1874/2889) of which yielded a PCR product of any kind, and about 51% (1464/2889) generated visible products of the expected size (this is the yield we refer to) when analyzed using agarose gel electrophoresis (Figure 3, A and B).

In our initial studies, we observed a pronounced lowering of the efficiency of the amplification of longer fragments. Therefore, we sorted target genes into different size fractions and established optimal PCR conditions for genes of different lengths. We were able to amplify transcripts up to 10 kb by adding Pfu and DMSO to the reverse transcription reaction and by using longer PCR extension times up to 6 min. With the tailored conditions, we found that size has less impact on the efficiency of recovery. The current yield for fragments in the 0–1.2 kb range is approximately 60%, and for longer fragments up to 10 kb, the value is approximately 50%.

In an effort to recover additional clones, experiments focusing on the RNA source, number of PCR cycles, and primer redesign were performed. The genes that were not isolated using the initial Human Master Panel RNA source were processed for a second round of RT-PCR utilizing a supplemental panel of four tissues (see Materials and Methods). The expected products covered the full range of sizes of 0–3 kb. These tissues were selected on the basis of the fact that a large fraction of the PCR-negative genes in the first round were expected to be expressed in these four tissues (data not shown). Approximately 30% of genes that failed to yield a product from the initial RT-PCR attempt (PCR negative) were rescued through a second round of PCR using these alternate tissue sources. Additionally, increasing PCR to 45 cycles resulted in a 25%–30% yield among PCR-negative genes. Finally, a set of 192 PCR-negative genes, which were expressed in the initial Human Master Panel RNA source, were subjected to a second round of primer design. The second round PCR yield with new primers was as low as 7%. In summary, the accumulative yield reached 75% after the above PCR experiments.

Cloning, Prescreening for Full-length Sequencing, and Sequence Evaluation

A major improvement during the blunt-end PCR cloning procedure was to convert to a 96-well format. This
modification expedited high-throughput cloning and resulted in an 80% cloning efficiency according to β-galactosidase activity screening. Through prescreening subclones, the correct insert sizes were obtained for approximately 47% (1356/2889) of all the attempted genes. Laboratory information management systems (LIMS) software was developed to effectively handle the large number of subclones. Subclones with insert sizes different from the RefSeq record were saved for future alternative transcript analysis. An average of 3.9 subclones per gene (5398 subclones for 1356 genes) were rearrayed and sequenced in order to obtain clones suitable for submission. For sequencing, we found it was most efficient to use a primer walking strategy, where only one set of primers needs to be synthesized for each gene.

The final step in subclone processing is to compare the finished sequence to the targeted RefSeq sequence and to other databases in order to determine if the sequence is acceptable (see the section entitled Extensive Transcript Heterogeneity and Clone Acceptability). Software has been developed to automate the pipeline for the selection of the desired clones representing the targeted genes. Of the genes that have fully sequenced subclones, 90% (576/633) were successfully rescued. Out of these, about 40% perfectly match the established RefSeq sequence, 30% have alterations outside the ORF as well as synonymous nucleotide changes, and 20% are genomic matches (including the ones matching EST, mRNA, and SNP databases). The other 10% of the genes have nucleotide discrepancies in both RefSeq and the genome in all of the sequenced subclones (Figure 4). The discrepancy rate is about 1.5 bases/kb with the substitution to insertions/deletions ratio close to 1:1 based upon approximately 1018 kb of sequence.

**Extensive Transcript Heterogeneity and Clone Acceptability**

The cDNAs generated in this study are each expected to have a corresponding sequence in the RefSeq collection with an already defined ORF. Reasons for possible differences between the rescued clones and preexisting data include errors in RefSeq, splice variants, the accuracy of the ORFs that are nominated, and polymorphisms. Therefore, besides the RefSeq database, other biological evidence is needed to determine whether rescued sequences represent sequences found in nature. The current guidelines as to what are acceptable clones are as follows.

The rescued cDNA clone may have single DNA base differences to the RefSeq, which do not change the ORF (e.g., either outside the ORF or silent changes). Natural variations supported by high-quality sequences are also acceptable. For example, differences to RefSeq but matches to the genome after an appropriate alignment of the rescued clone sequence to the finished human genomic sequence are acceptable. Likewise, similar matches of

![Figure 4. Categorization of finished genes by correspondence to database sequences.](image-url)
murine cDNAs to the finished regions of the mouse genome are acceptable. Many such cases have been observed, for example, there were 116 genes that had subclones with mismatches to RefSeq sequence but a perfect match to the available reference genome sequence. Additionally, due to the likelihood of polymorphisms, it is likely that some clones will have a mixture of RefSeq sequence and genomic sequence where the two differ. Also, when a discrepancy perfectly matches existing EST, mRNA, and SNP databases, this is acceptable. The occurrence of insertion/deletions follows essentially the same rule set as single base differences. For example, if the ORF is unchanged, then the clone is acceptable. In-frame deletions are not acceptable. Polymorphic deletions are acceptable. The following are two illustrative examples of interesting transcripts that differ from RefSeq but match genome sequences.

Contig 337 (NM_006191). This transcript has two nucleotide substitutions inside the ORF and substitutions as well as a deletion outside of the ORF relative to the RefSeq sequence. One substitution inside of the ORF leads to an amino acid change (Ala to Pro). Theoretically this transcript is unacceptable. However, we compared contig 337 to the genomic sequence and observed no discrepancy. In addition, contig 337 maintained the same ORF structure as the RefSeq sequence.

Contig 523 (NM_006418). This transcript showed extensive differences with RefSeq sequence at both the nucleotide and translated amino acid level when analyzed using Blast2 (http://www.ncbi.nlm.nih.gov/gorf/bl2.html) (Figure 5A). However, it showed no discrepancy with genome sequence except one nucleotide substitution outside of the ORF (Figure 5B). Most interestingly, there was a 257-bp fragment of NM_006418 missing from the genome sequence in addition to multiple substitutions. And this 257-bp fragment matches a fragment adjacent to this locus on the opposite DNA strand, therefore, forming a palindrome structure in the RefSeq sequence. There were several transcript variants documented at this locus. Some resembled the RefSeq sequence (such as AF097021), and some were similar but not identical to the transcript we have recovered (such as AK000683). The gene structure of the transcript we have recovered could be represented by the combination of several documented mRNAs/ESTs. Furthermore, NM_006418 has an ORF of 657 bp, but our recovered transcript has a longer ORF of 1104 bp.

**DISCUSSION**

RT-PCR gene amplification is a familiar method for experimentation on a small scale. The large-scale process of amplifying and efficiently rescuing hundreds of clones of a genome provides additional challenges. We have overcome practical difficulties in developing robust and efficient protocols for high-throughput RT-PCR cloning of transcripts of various sizes, integrating a number of computational tools, and assessing overall cost/errors/speed and scalability.

The gene targets in this study were selected from those that were previously documented in databases and publications, but did not appear in the MGC clone set after sequencing more than 100 cDNA libraries constructed for the program. Therefore, an improved method for obtaining representative cDNA clones has become essential. The method we have developed is shown generally suitable for gene recovery. It could also be implemented to target genes expressed at specific locations, time and levels, as well as to discover new genes suggested by gene prediction programs (Wu et al., unpublished data). The various factors affecting the efficiency of the pipeline are discussed below.

Most existing primer design programs, such as primer 3, require customization to locate the primer sets closely flanking the ORF in a high-throughput manner. Primer selection is often further complicated by the limited length and sequence composition of documented 5' or 3' UTR regions. We therefore developed an automated primer selection system that is suitable for the RT-PCR rescue pipeline. In general, the primer design does not appear to be the limiting factor in our pipeline.

The important limiting step lies in optimizing RT-PCR conditions in a scalable manner and improving the success rate. The major challenges for rescuing genes via PCR-based methods required the following considerations: (i) adequate sensitivity for amplifying genes that have not been isolated by other means; (ii) fidelity for reduced mutation rate; and (iii) specificity in amplifying the gene of interest. We have demonstrated a functional pipeline that addressed each. In terms of sensitivity and fidelity, we chose the Advantage 2 polymerase mixture because it is relatively accurate and capable of efficiently amplifying the cDNA transcripts expressed at very low levels. Although some polymerases have
a greater fidelity than Advantage 2 polymerase, our decision was based on the overall robustness and precision of the enzyme. In order to improve the specificity of PCRs, a critical decision in the protocol was to use touchdown PCR, which significantly simplifies the process of determining optimal annealing temperatures. This is particularly useful in our large-scale project, because the choice of primer T_m could be limited in some UTR regions flanking the ORF. Additionally, we increased the PCR extension time incrementally in order to increase the yield of PCR products, which is crucial for increasing the insert to vector ratio to achieve high efficiency during the blunt-end cloning step. Finally, from the additional 30% yield obtained when utilizing a supplemental panel of four tissues (see Results section), it appears that altering the composition of the RNA pool is likely to be very productive for increasing sensitivity in recovering genes that failed initially. Many genes are missing from the current collection because they are only expressed in specific tissues or during specific developmental stages.

All genes for which PCR was attempted were passed through the cloning pipeline, irrespective of whether a PCR product was detected by agarose gel electrophoresis or not. This decision was based on the observation that approximately 15% of the PCR-negative samples yielded subclones of the expected size that were subsequently rescued. Apparently the desired fragments were present but not visible on the agarose gels. Furthermore, we have observed a trend toward a decrease in subclone yield per gene with increasing insert size. Cloning systems designed specifically for longer fragments are currently being tested. In-Fusion™ Cloning Kit (BD Biosciences Clontech) has shown an efficient cloning yield for PCR fragments up to 8 kb.

The RT-PCR rescue method requires a postsequencing decision of whether a PCR product was detected by agarose gel electrophoresis or not. This decision was based on the observation that approximately 15% of the PCR-negative samples yielded subclones of the expected size that were subsequently rescued. Apparently the desired fragments were present but not visible on the agarose gels. Furthermore, we have observed a trend toward a decrease in subclone yield per gene with increasing insert size. Cloning systems designed specifically for longer fragments are currently being tested. In-Fusion™ Cloning Kit (BD Biosciences Clontech) has shown an efficient cloning yield for PCR fragments up to 8 kb.

The RT-PCR rescue method requires a postsequencing decision of whether a PCR product was detected by agarose gel electrophoresis or not. This decision was based on the observation that approximately 15% of the PCR-negative samples yielded subclones of the expected size that were subsequently rescued. Apparently the desired fragments were present but not visible on the agarose gels. Furthermore, we have observed a trend toward a decrease in subclone yield per gene with increasing insert size. Cloning systems designed specifically for longer fragments are currently being tested. In-Fusion™ Cloning Kit (BD Biosciences Clontech) has shown an efficient cloning yield for PCR fragments up to 8 kb. An important observation from our study was that discrepancies between RefSeq and genome sequences are frequent (see Contig 337 and Contig 523 sections). Therefore, we have decided the sequences different from RefSeq but matching the finished human or mouse genomic sequence are acceptable. More interestingly, contig 523 is supported by combinations of mRNA/EST sequences. The information from haplotype study of cDNAs will be very useful in assessing these transcripts in the future. Noticeably, the discrepancy rate measured using the direct sequencing approach from our experiment (approximately 1.5 base/kb) is much higher than the error rate measured using codon reversion and colony selection by suppliers. One of the likely explanations is that some discrepancies are actually true polymorphisms. Therefore, we are currently engaged in resequencing a set of human DNA samples to identify possible polymorphisms among these observed discrepancies.

Previous EST-based studies suggested a major percentage (30%–60%) of the human and mouse genes have alternative spliced transcripts (21–24). This adds another layer of complexity to the mammalian transcriptome. In our approach, the longest form of the RefSeq sequence of each gene is targeted. If the alternative forms involve only internal exons it may be possible to amplify several possible forms with a specific primer pair in one PCR. In cases where alternative terminal exons may be involved, targeting the longest PCR product may miss some forms. We have observed subclones with large sequence deletions or insertions compared to the RefSeq sequences. We are in the process of assessing the possibility that they are alternative transcripts by comparing the sites of deletion or insertion to the splice site information in UCSC database (http://genome.ucsc.edu). If the splice site exists at the position of deletion or insertion, it is likely that the sequence is a potential variant transcript. In some cases, it still will be challenging to distinguish functional transcripts from pseudogenes.

In conclusion, we have addressed the difficult issues involved in obtain-
ing functional clones for transcripts that have not been isolated by other means by improving RT-PCR amplification efficiency, clone fidelity, and clone validation where sequences differ from the existing reference database. A functional format scalable to a 96-well format has been established, in which several thousand unique human and mouse genes are undergoing complete analysis. The sequences of acceptable subclones along with a glycerol stock will be submitted to a public repository (clone information can be found at MGC web site: http://mgc.nci.nih.gov). The submitted cDNA clones will be made freely accessible to the research community.

ACKNOWLEDGMENTS

Thanks to the Mammalian Gene Collection (MGC) for programmatic support. Thanks to the Baylor College of Medicine Human Genome Sequencing Center for production support. We express our gratitude to the following individuals for assistance: Donna M. Muzny, Anne Hodgson, Ryan Martin, Seema Nair, Amit Nanavati, Hermela Lounseged, Edwin Fuh, Kim Haebeler, Rui Chen, Jing Liu, Feng Xue He, Graham B. Scott, and Peter R. Blyth. Thanks to Dr. M. Metzker, Dr. Z.D. Zhang, I. Yakub, and F.L. Yu for useful discussions. This project was partially supported by the grants from the National Cancer Institute (NCI)/Science Applications International Corporation (SAIC) (20XS182A).

REFERENCES


Received 5 December 2003; accepted 9 February 2004.

Address correspondence to: Jia Qian Wu
Human Genome Sequencing Center
Department of Molecular and Human Genetics
Baylor College of Medicine
Houston, TX 77030, USA
e-mail: jwu126640@bcm.tmc.edu

DRUG DISCOVERY AND GENOMIC TECHNOLOGIES