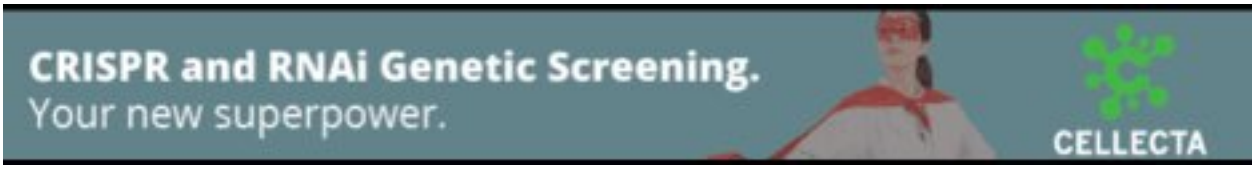
Systematic Recovery and Analysis of Full-ORF Human cDNA Clones

Ágnes Baross, Yaron S.N. Butterfield, Shaun M. Coughlin, et al.

Genome Res. 2004 14: 2083-2092
Access the most recent version at doi:10.1101/gr.2473704



Systematic Recovery and Analysis of Full-ORF Human cDNA Clones

Ágnes Baross, Yaron S.N. Butterfield, Shaun M. Coughlin, Thomas Zeng, Malachi Griffith, Obi L. Griffith, Anca S. Petrescu, Duane E. Smailus, Jaswinder Khattra, Helen L. McDonald, Sheldon J. McKay, Michelle Moksa, Robert A. Holt, and Marco A. Marra[1]

Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia, V5Z 4E6 Canada

The Mammalian Gene Collection (MGC) consortium (http://mgc.nci.nih.gov) seeks to establish publicly available collections of full-ORF cDNAs for several organisms of significance to biomedical research, including human. To date over 15,200 human cDNA clones containing full-length open reading frames (ORFs) have been identified via systematic expressed sequence tag (EST) analysis of a diverse set of cDNA libraries; however, further systematic EST analysis is no longer an efficient method for identifying new cDNAs. As part of our involvement in the MGC program, we have developed a scalable method for targeted recovery of cDNA clones to facilitate recovery of genes absent from the MGC collection. First, cDNA is synthesized from various RNAs, followed by polymerase chain reaction (PCR) amplification of transcripts in 96-well plates using gene-specific primer pairs flanking the ORFs. Amplicons are cloned into a sequencing vector, and full-length sequences are obtained. Sequences are processed and assembled using *Phred* and *Phrap*, and analyzed using Consed and a number of bioinformatics methods we have developed. Sequences are compared with the Reference Sequence (RefSeq) database, and validation of sequence discrepancies is attempted using other sequence databases including dbEST and dbSNP. Clones with identical sequence to RefSeq or containing only validated changes will become part of the MGC human gene collection. Clones containing novel splice variants or polymorphisms have also been identified. Our approach to clone recovery, applied at large scale, has the potential to recover many and possibly most of the genes absent from the MGC collection.

[Supplemental material is available online at www.genome.org and www.bcgsc.ca/bioinfo/MGC.]

A full-open reading frame (ORF) cDNA clone provides the best experimental evidence of transcription, transcript processing, and gene structure. Other approaches for identification of transcribed genomic regions, such as automated gene predictions and expressed sequence tag (EST) data provide useful information; however, computational gene predictions are not fully accurate, and EST data are error-prone and usually only sample a portion of transcripts. The availability of full-ORF cDNA sequence data for all genes within a species would be a key resource for identifying coding regions within the genome, determining the structure of genes including identification of splice variants, and understanding the proteome. Furthermore, access to the cDNA clones would facilitate functional studies of genes and corresponding proteins.

The Mammalian Gene Collection (MGC; http://mgc.nci.nih.gov) project aims to generate a public resource of full-ORF cDNA clones for various species, including human, mouse, and rat (Strausberg et al. 1999, 2002). For human and mouse, significant progress towards these goals has already been achieved via generating and characterizing cDNA libraries originating from a broad variety of tissues and cell lines. In the current paradigm, libraries enriched for full-length cDNAs are constructed, then candidate full-ORF clones are selected based on analysis of 5′ ESTs. This is followed by systematic full-length sequencing of selected cDNA clones (Butterfield et al. 2002; Shevchenko et al. 2002; Strausberg et al. 2002). To date, more than 15,200 human and more than 12,600 mouse full-ORF cDNA clones have been produced by the MGC effort, corresponding to at least 11,100 and 10,100 unique genes, respectively (http://mgc.nci.nih.gov). The systematic EST-based approach for clone identification, however, is decreasingly effective as the number of identified genes increases. There are still thousands of known genes missing from the human and mouse collections, and hundreds of ESTs are now required to identify new cDNAs to add to the MGC collection. Thus, alternative, more efficient strategies are needed for acquisition of cDNA clones representing the genes missed by the systematic EST approach.

We have undertaken an MGC-funded pilot project to develop methods for the "targeted" recovery of cDNAs for genes that remain unsampled by systematic EST analysis. We report here the details of our methods, which rely on a gene-specific reverse transcription-polymerase chain reaction (RT-PCR)-based approach for targeted generation of cDNA clones based on known RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq) sequences. We report also assessment of clone quality based on restriction fragment analysis, full-length sequencing, and a detailed analysis of insert sequences.

RESULTS

Selection of RNA Sources and RT-PCR

We conducted a pilot project in which we targeted for recovery 384 human genes (four sets of 96, Table 1; detailed gene list is shown in Supplemental Table S1). These genes were chosen from a list of 2059 gene targets provided to us by the MGC. These gene

[1]**Corresponding author.**
E-MAIL mmarra@bcgsc.ca; FAX (604) 877-6085.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2473704.

Table 1. Progress Summary of Clone Recovery

	1st set		2nd set		3rd set		4th set		Total	
Genes targeted	# Genes	%	# Genes	%	# Genes	%	# Genes	%	# Genes	%
Attempted by RT-PCR	96	*100*	96	*100*	96	*100*	96	*100*	384	*100*
Expected size amplicons	89	*93*	87	*91*	78	*81*	92	*96*	346	*90*
Clones available	88	*92*	87	*91*	78	*81*	91	*95*	344	*90*
Acceptable clones found[a]	75	*78*	67	*70*	55	*57*	62	*65*	259	*67*
No acceptable clones[b]	13	*14*	17	*18*	20	*21*	17	*18*	67	*17*
Clones pending analysis	0	*0*	3	*3*	3	*3*	12	*13*	18	*5*

[a]Acceptable clones contain ORFs with identical sequences to RefSeq, or contain only validated changes.
[b]These clones were not acceptable based on the current MGC criteria. However, 17 of them may represent real polymorphisms or splice variants. These clones are further categorized in Figure 7.

targets represent well characterized genes that were not represented in the MGC collection. An overview of our process is shown in Figure 1. The results from this study are described below and demonstrate the feasibility of our approach to produce full-ORF cDNA clones.

We selected groups of 96 genes based on similarity in coding sequence (CDS) length, and expression profile as determined using transcription data found in UniGene (NCBI; http://www.ncbi.nlm.nih.gov/UniGene). These criteria for gene selection were designed to increase the efficiency of subsequent RT and PCR reactions, as the RNA source and PCR cycling conditions were consistent for genes within a group. cDNAs were synthesized from mRNAs originating from various human tissues (Fig. 1; Methods). Based on our analysis of UniGene, we chose 13 tissues from which expression of the majority of the 2059 genes were detected. These RNA sources were brain, colon, heart, kidney, liver, lung, muscle, ovary, pancreas, placenta, stomach, testis, and uterus. Prior to amplification, the quality of cDNA was verified on agarose gels. The reverse transcription method resulted in high-quality cDNA with strong, continuous smears on the gels visible up to 12 kb or more (Fig. 2). These cDNAs were

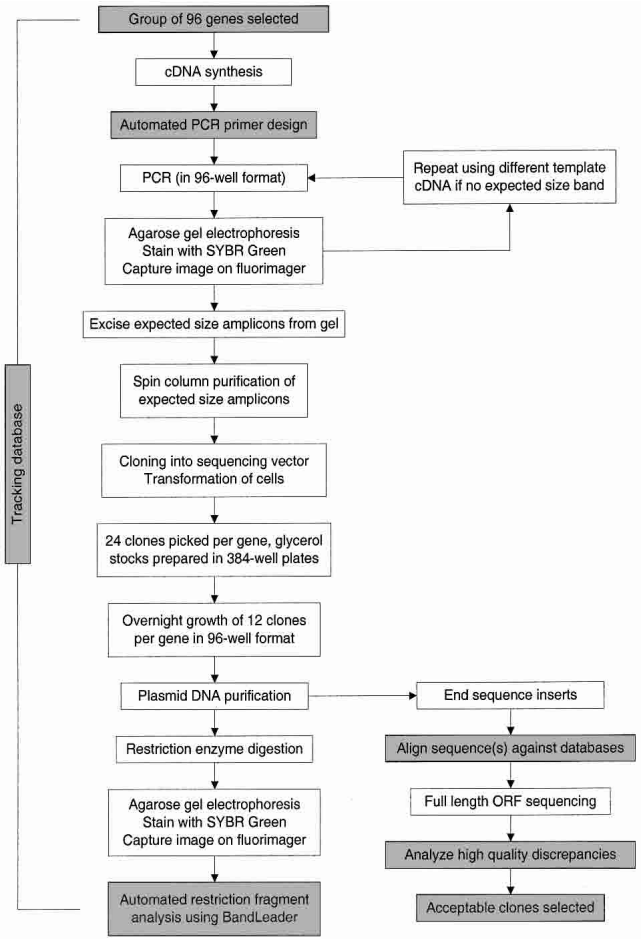


Figure 1 Overview of the targeted clone recovery process. "Wet lab" experimental approaches are shown on white background, and bioinformatics methods are shown on gray background.

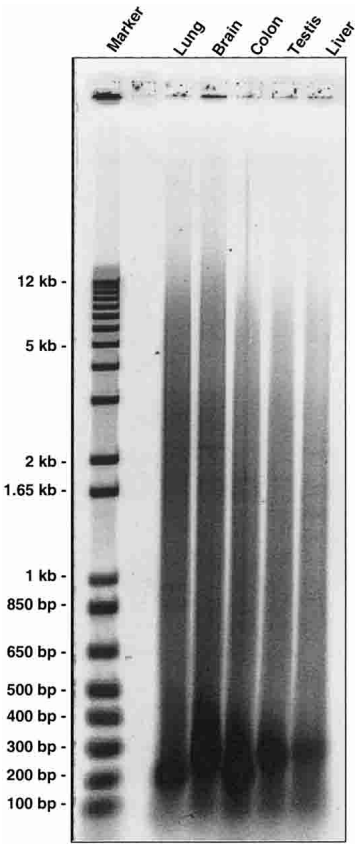


Figure 2 Agarose gel electrophoresis of double-stranded cDNA. The sources of RNA are shown at the *top* of the gels. cDNA was synthesized from 1 µg high-quality mRNA per sample, and 1 µL of the resulting 20 µL cDNA per sample was loaded in the five sample wells of a 1% agarose gel.

then used as templates in gene-specific PCR amplification reactions.

Following the RT reaction, gene-specific primers were designed using an automated method we developed (Fig. 1). Perl programs were written to obtain Fasta sequences for given genes from RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq), and Primer3 (http://www.broad.mit.edu/cgi-bin/primer/primer3_www.cgi; Rozen and Skaletsky 2000) was used to design PCR primers that flank the ORFs. Primers were chosen by Primer3 such that they conformed to a number of parameters (Methods), including GC content, length, melting temperature (Tm), potential for forming hairpins, and self annealing.

Transcripts were amplified by PCR in 96-well format using a high-fidelity DNA polymerase (Methods) and gene-specific primers (Fig. 1). The primer pairs used for each gene are shown in Supplemental Table S1. PCR products were then subjected to agarose gel electrophoresis, and the sizes of amplicons were manually estimated (Fig. 1). An example of a PCR gel is shown in Figure 3 with amplicons of the expected size indicated with arrows. Often there were PCR products in addition to the expected size amplicons, possibly due to splice variants of the targeted gene or nonspecific amplification of other transcripts or transcript fragments. Elimination of the latter would require further individual optimization of the PCR parameters for each transcript. This would not be amenable to high-throughput methodology and was therefore not pursued. Indeed, for purposes of acquiring full-ORF clones for well characterized genes, elimination of the extra bands through PCR optimization was not necessary, as long as the expected-size amplicons were well separated from the other bands on agarose gels.

When the RNA template was chosen using available EST expression information from UniGene, the PCR reaction yielded expected-size bands for 65 (67%) of the 96 targeted genes. The PCR was highly reproducible when repeated from the same template; thus we did not routinely repeat it the same way unless there was an obvious technical error. However, we found that repeating the PCR step on the failed attempts using different RNA sources (Fig. 1) significantly increased the success rate to 91% (third set of genes, Table 1). So far we have obtained expected-size amplicons for a total of 346 (90%) of the 384 targeted genes in Table 1 and Supplemental Table S1. This was achieved using 5–10 different RNA sources (or mix of RNA sources) per set of 96 genes. When expression information was available for the set of 96 transcripts, the first PCR attempt was performed using a common known RNA source (e.g., brain RNA if all 96 transcripts were detected in brain). Subsequent PCR attempts were performed using cocktails of cDNAs as templates originating from the 13 tissues listed above (e.g., mix of lung, colon, and testis cDNA used in the second PCR attempt, mix of liver, uterus, and kidney cDNA used in the third PCR attempt, etc.). The RNA sources that yielded expected-size RT-PCR amplicons for given genes are listed in Supplemental Table S1. For the third set of genes we targeted (Table 1), there was no evidence of expression available in UniGene in our 13 chosen tissues. However, using our 13 RNA sources as templates, we obtained expected-size amplicons for 81% of these 96 genes. Of the 384 genes we targeted, there were only 38 (10%) for which we have not yet been able to obtain expected-size PCR amplicons (Table 1 and Supplemental Table S1).

Amplicon Cloning

After amplification, visible bands of expected size were excised from agarose gels (Fig. 1). Gel fragments were purified using a spin column-based method (Fig. 1; Methods) and ligated into a sequencing vector containing M13 forward (−20) and M13 reverse priming sites (Fig. 1; Methods). The cloning method proved to be efficient, and we successfully generated clones from 344 of the processed 346 amplicons (Table 1; Supplemental Table S1).

Plasmid vectors containing PCR amplicon inserts were electroporated into bacterial cells (Fig. 1), and recombinant clones were selected on agar plates containing appropriate antibiotics (Methods). Glycerol stocks were prepared from 24 individual clone isolates per amplified gene and stored in 384-well plates (Fig. 1). Plasmid DNA was prepared (Fig. 1; Methods), and then plasmids were digested with EcoRI and analyzed by agarose gel

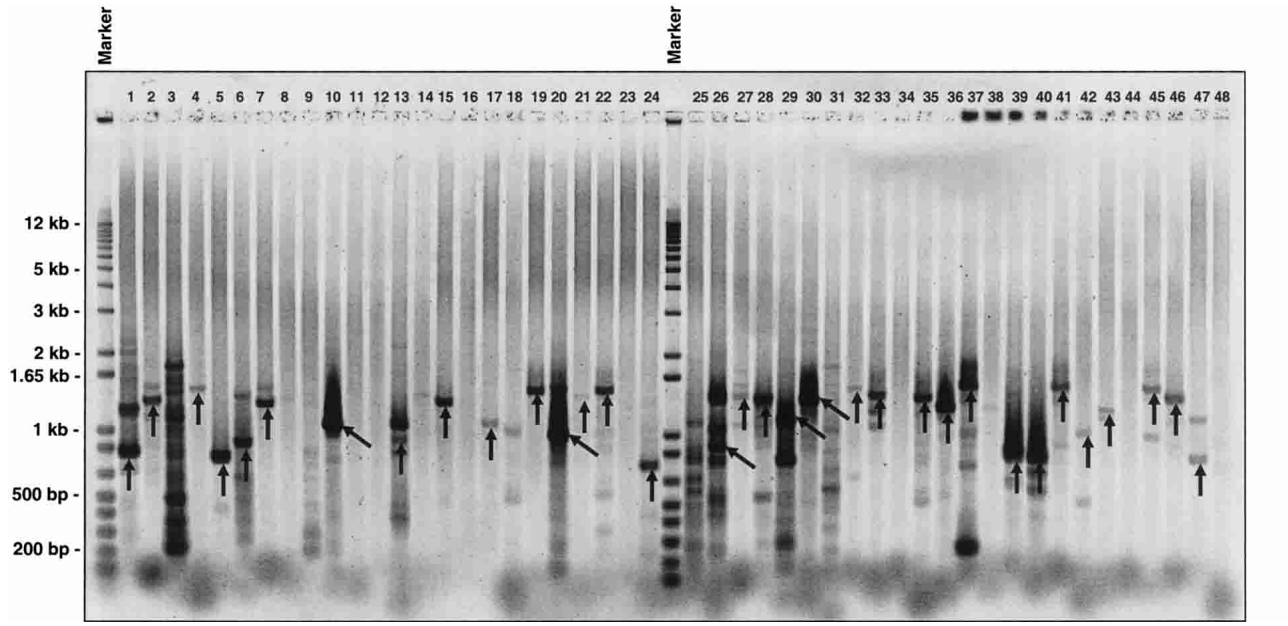


Figure 3 Electrophoretic analysis of PCR-amplified ORFs. PCR amplification was performed using lung cDNA template and gene-specific primers for 96 target genes. The results of 48 amplifications are shown here. Ten µL of a 25-µL reaction for each sample was loaded on a 1% agarose gel. Expected-size amplicons of target genes are indicated with arrows.

electrophoresis to screen for nonrecombinant clones (Fig. 1; Methods). An example of clone restriction fragments separated on an agarose gel is shown in Figure 4. The resulting restriction fragments were compared with restriction fragments computationally predicted from the RefSeq sequence for each gene (i.e., in silico digests; Fig. 1). Clones with the expected pattern were sequenced for quality assessment (Fig. 1).

Clone Sequence Analysis

As clone inserts were generated via RT-PCR, we anticipated the possibility of reverse transcription- and PCR-induced artifacts. From the manufacturers' estimates of the error rates of reverse transcriptase and high-fidelity DNA polymerase used (Methods), we calculated the probability of finding an error-free clone for a transcript for set numbers of sequenced clones (Methods). Figure 5 shows the number of clones that would need to be sequenced as a function of PCR cycle number, in order to find at least one error-free clone with probabilities 50%, 75%, 90%, and 99%. The probability of finding at least one acceptable clone was calculated for PCR cycle numbers between 20 and 70 for 1-kb ORF length, and PCR cycle numbers between 20 and 50 for 2-, 3-, 5-, and 10-kb ORF lengths. The number of PCR cycles currently used in our protocol is between 20 and 30 (10 touchdown cycles followed by 20 cycles at standard annealing temperature). For 2-kb ORFs amplified using up to 25 PCR cycles, we calculated that sequencing 12 clones would be sufficient to find one acceptable clone 99% of the time on average (Fig. 5). For 2-kb ORFs that are amplified with 30 cycles, there is still a greater than 90% chance of finding an acceptable clone among 12, based on the above prediction. However, in some cases sequencing of additional clones will likely be necessary. The numbers of RT-PCR errors, and consequently the predicted numbers of clones to be sequenced, increase exponentially with ORF length. For 3-kb ORFs and 30 PCR cycles, sequencing 12 clones is sufficient to identify an acceptable clone with 75% probability, and this probability increases to over 90% when 24 clones are sequenced. In the case of 5-kb ORFs and 20–30 PCR cycles, the predicted chance for finding an acceptable clone is between 50% and 90% among 24 sequenced clones, depending on the exact number of PCR cycles. For 10-kb ORFs, the chances of finding an acceptable clone are less than 50% even if 50 clones per gene were sequenced. The genes we report on here were up to 2 kb, and we sequenced up to 12 clones per gene to full length in the first round. In cases where acceptable clones have not yet been recovered (Table 1), we plan to process an additional 12 (Fig. 6). However, by sequencing only up to 24 clones, the efficiency of gene rescue would decrease markedly for ORF sizes greater than 5 kb, even if the RT-PCR-based clone acquisition itself was successful. For ORFs less than 1 kb, our experience has been that fewer than 12 clones are sufficient.

Sequence end reads were compared with RefSeq and genomic sequence (Fig. 1; Methods). If the end reads were of high quality and matched the target gene, we proceeded to obtain full-length sequences (Methods). According to the standards established by the MGC for our pilot project, clones with full-ORF sequences identical to RefSeq were considered rescued. A complete match of the UTR sequence was not required to consider the gene rescued. In cases of sequence discrepancies, we evaluated each change as follows according to the standards established in the MGC pilot project for clone recovery. Base changes that did not result in an amino acid substitution were acceptable and did not require validation. Sequence changes that were confirmed from the genomic sequence, dbSNP (http://www.ncbi.nlm.nih.gov/SNP) stored in Ensembl (http://

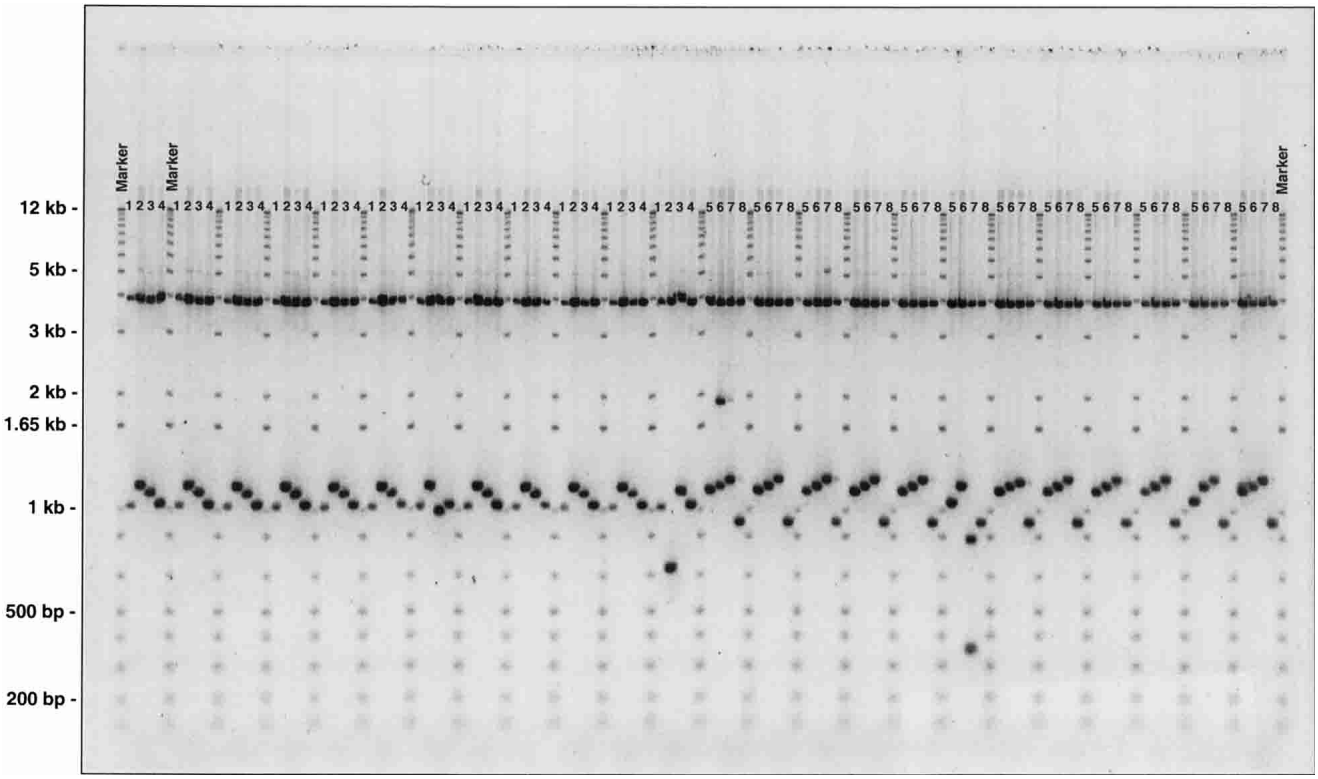


Figure 4 Agarose gel electrophoresis of EcoRI restriction-digested clones. The gel contains digests of 12 clones from each of eight PCR-amplified ORFs. One µL of plasmid DNA per clone was cut with EcoRI in a 96-well plate and loaded on a 1.2% agarose gel. Due to the difference in spacing between the gel and the multichannel pipetter used for loading, clones for the same gene are located in every fifth well. DNA marker is loaded in every fifth lane.

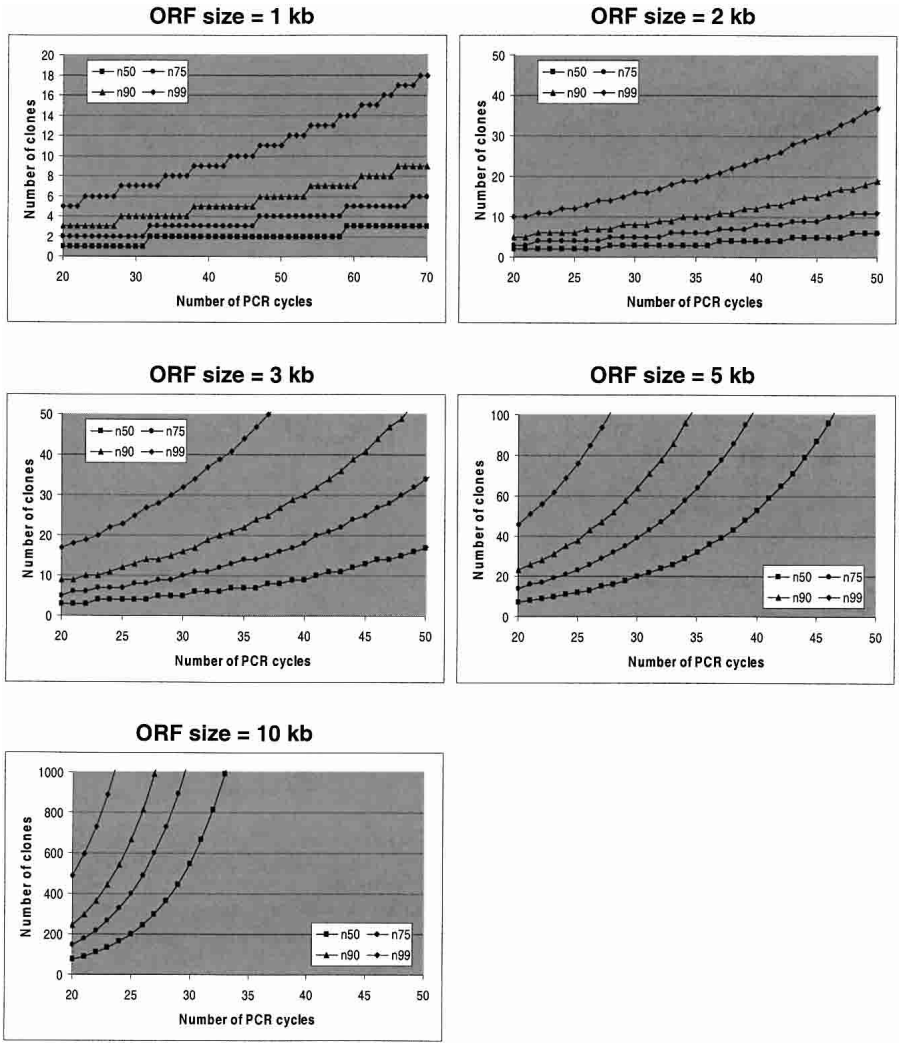


Figure 5 Estimated numbers of RT-PCR-generated clones required on average to identify at least one acceptable clone of the indicated length (as a function of PCR cycle number). This is based on 1/15,000 error rate of the reverse transcriptase, and 1/50,000 error rate of the high-fidelity DNA polymerase used in the clone acquisition process. n50, n75, n90, and n99 indicate the predicted numbers of clones that need to be sequenced in order to find an acceptable clone with probabilities of 50%, 75%, 90%, and 99%, respectively, based on the above error rates.

www.ensembl.org; Hubbard 2002; Hubbard et al. 2002; Birney et al. 2004; Kasprzyk et al. 2004) or dbEST (http://www.ncbi.nlm.nih.gov/dbEST; Boguski et al. 1993) databases were also acceptable. Clones containing confirmed changes within the ORFs were considered rescued. The full list of analyzed clones with corresponding Phred scores and sequence analysis results is shown in Supplemental Table S2. A few examples of rescued clones are listed in Table 2. If more than one acceptable clone was found for a gene, the acceptable clones were ranked based on the number of sequence discrepancies compared to RefSeq. The best clones for a given gene were defined as having the least number of validated discrepancies within the ORF, and the least number of overall changes outside the ORF. These acceptable clones will be submitted to MGC.

To date, acceptable clones have been found for 259 genes, which is 67% of the 384 genes targeted and 75% of the 346 genes that resulted in an expected-size PCR product (Table 1). In order to find these, 4718 clones were sequenced (Supplemental Table S2). For 67 genes (17%) of the 384, clones have been generated from expected-size amplicons; however, after sequence analysis these genes did not meet the requirements described above and shown in Figure 6. The reasons for failure and numbers of genes failing are summarized in Figure 7. Thirty-eight genes of the 384 (10%) were declared failures due to the lack of PCR product of expected size. For two genes, cloning of the PCR products was unsuccessful. Among 67 nonrescued genes where clones were generated and sequenced, 40 matched RefSeq sequences of genes other than the targeted ones, of which 32 contained the PCR primer sequences used for the target gene and eight did not (Fig. 7).

Clones for 27 of the 67 nonrescued genes matched the correct RefSeq sequence, but failed due to various nonvalidated discrepancies (Fig. 7). Interestingly, in this category the clones for a given gene often shared a common error that resulted in failure. In 17 of the 27 genes in this category, the same error failed 50%–100% of the clones. Five of these common errors were deletions, three were insertions, and five were base substitutions. Four of these errors involved both insertions and deletions. We believe that these kinds of "errors" are not likely to be artifacts of our procedure (e.g., introduced by RT-PCR), but rather originate from the template RNA and are probably biologically valid polymorphisms even if they had not been found before.

PCR bands of the amplicons processed to date were sorted into three categories in a semiquantitative manner based on band intensity: "faint" bands, "medium" bands, and "intense" bands. Once the clone acquisition process was completed and clone insert sequences were analyzed, we compared the rates of successfully rescued genes within these categories. For clones generated from faint PCR bands, the ratio of successful rescue versus failure to obtain an acceptable clone for the targeted gene was about 1:1. For medium-strength bands, this ratio increased to about 2:1, and was far better for the intense bands, about 12:1. Thus, recovery was most efficient when clones were generated from a large quantity of PCR product. Although less efficient, we chose to process faint bands, as we still achieved 50% recovery of the targets.

As mentioned earlier, we often observed multiple bands on the PCR gel instead of the expected full-ORF amplicon size alone (Fig. 3). To test whether some of these represented additional splice variants of the target gene, we cloned 37 of these extra bands for 31 genes, and processed them as described above (Fig. 1). An agarose gel containing a subset of these amplicons is shown in Figure 8. Sequence analysis revealed that 23 of these fragments of unexpected size contained potential alternative splice forms of the targeted gene (Supplemental Table S3). Six bands were not splice variants (these resulted from mispriming events within the targeted ORF or other transcripts), and clones

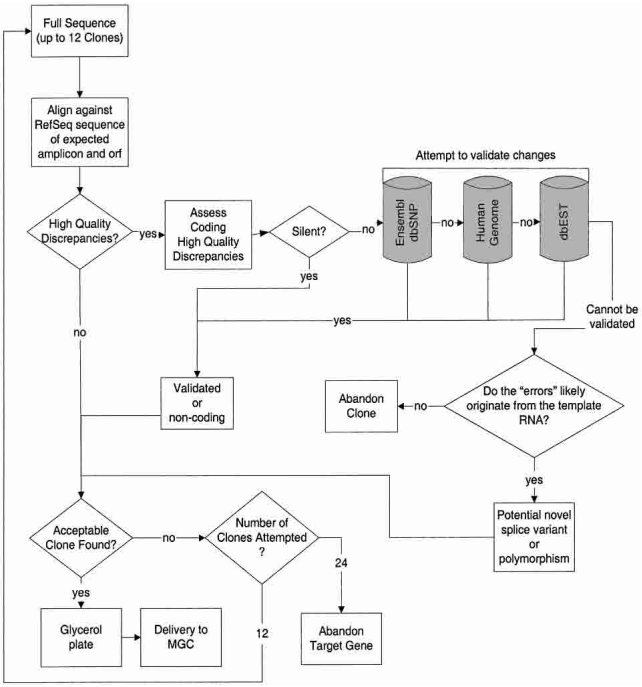


Figure 6 Bioinformatics sequence analysis pipeline. Databases used for validating clone sequence versus RefSeq discrepancies are shown on grey background.

from eight bands are still being analyzed. In some instances, one isolated PCR band yielded more than one splice variant (Table 3; Fig. 9). In these cases, the sizes of splice variants were very similar, and could not be resolved by our agarose gel electrophoresis conditions.

DISCUSSION

Using our approach (Fig. 1), we successfully generated cDNA clones from expected full-ORF size RT-PCR amplicons (based on RefSeq Fasta sequences) for 344 (90%) of 384 targeted human genes (Table 1; Supplemental Table S1). Agarose gel electrophoresis of PCR products (Fig. 3) and EcoRI restriction fragment analysis (Fig. 4) provided an initial indication of whether a given amplicon represented the gene of interest. The ultimate method for determining whether clones were acceptable was full-length sequencing (Fig. 1) and sequence analysis (Fig. 6). For 259 genes (67% of the 384) we have found at least one acceptable clone according to current MGC criteria, such that the full-ORF sequence was identical to RefSeq, or the changes were either silent or represented known polymorphisms validated by dbSNP or dbEST databases.

Categories of nonrescued genes are shown in Figure 7. The cases where clone insert sequences did not match the targeted gene, and included the correct PCR primers, indicate that other transcripts or transcript fragments of similar size to the target gene and containing sequences very similar to the primer annealing sequences were cloned. Generally this is more likely to occur for paralogous gene family members. To reduce this kind of error in the future, we plan to further refine our automated primer design process using the publicly available electronic PCR tool (ePCR; Schuler 1997, 1998) which can be used to check for DNA sequences that have sufficient similarity to the primer sequence to initiate the amplification of a product of comparable size. The PCR primers will be screened against a database of cDNAs including RefSeq, MGC, and the Ensembl transcript set, as well as the human genome data (for the latter, splicing information will be indispensable in determining the real size of the expected transcript).

In cases where another transcript was cloned and did not contain the expected PCR primer sequences (Fig. 7), we did detect primer sequences of other target genes from the same set of 96. This suggests that these resulted from laboratory error, possibly cross-contamination of wells during loading of agarose gels prior to gel purification. To reduce the chances of this, we now load our PCR products in every second well on the gels instead of adjacent wells. In all of these cases of failed rescue attempts, we are repeating the clone acquisition process starting from the PCR step to overcome the lab error that presumably occurred during the first attempt.

The cases where clones corresponded to a targeted gene, but could not be rescued due to a common "error" in the recovered clones, likely correspond to biologically valid expressed sequences. We argue that these changes are not artifacts, but instead represent sequence polymorphisms (Fig. 7). Even if an error was introduced during the reverse transcription step or one of the early steps of PCR, it would be very unlikely to affect 50% or more of the clones, as we do not start our process from a single RNA/cDNA molecule but from many millions. Thus, these errors are more likely to be novel splice variants or polymorphisms not included in the current databases.

In the 384 genes attempted to date (Table 1; Supplemental Table S1), we had targeted CDS lengths of up to 2 kb. We have recently started experimenting with an additional set of 96 genes (data not shown) that includes transcripts varying from 2 kb to 17 kb in size. Based on preliminary data (not shown), we were able to obtain expected-size amplicons for ORFs up to 4 kb using

Table 2. Examples of Acceptable Clones Found

Target RefSeq ID	Gene name	PCR band size	Contig size	Comments
NM_000871	*HTR6*	~1400 bp	1401 bp	Matches ORF perfectly.[a]
NM_000910	*NPY2R*	~1200 bp	1203 bp	T627C[b] I−>I (silent),[c] T978C I−>I (silent)
NM_001971	*ELA1*	~800 bp	822 bp	T541C Y−>Y (silent), C680G L−>V (Genomic validation[d] − BLAT)
NM_003308	*TSPY*	~1000 bp	978 bp	A161G V−>V (silent), CCCG301GGGC PR−>RA (Genomic validation − BLAT). G458C H−>Q validated by dbEST.
NM_004347	*CASP5*	~1300 bp	1320 bp	G975C V−>L validated by Ensembl dbSNP as an SNP.

[a]Based on RefSeq sequence.
[b]Base change in clone insert vs. RefSeq.
[c]Effect of sequence change on protein coding.
[d]Validation means confirming the change as known polymorphism, based on current databases.

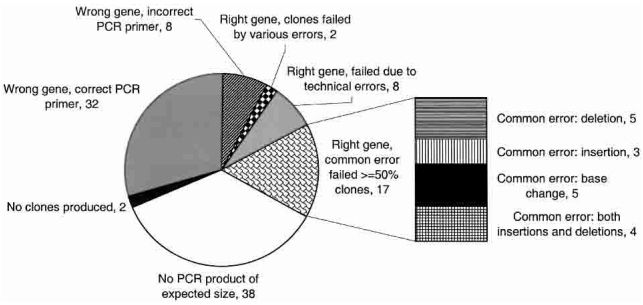


Figure 7 Summary of failed rescue attempts from RT-PCR-based clone recovery. Of 107 genes nonrescued to date, 38 were declared failures due to the lack of expected-size PCR amplicons. The cloning process failed for two genes. For 67 genes, clones representing expected-size amplicons were generated. Of these, clone insert sequences of 40 matched a RefSeq sequence other than the targeted gene; 32 of these contained sequences for the correct PCR primers used, whereas the remaining eight did not. Of 27 genes where the clone sequences matched the targeted gene, two failed due to various nonvalidated errors. Eight failed due to technical errors, such as primers amplifying within the ORF. For 17 genes, however, at least half of the clones could not be rescued due to a common unvalidated change. We suggest that clones in the last category may not be true failures, but rather novel splice variants or real polymorphisms that should be considered biologically valid.

the protocol described here. The distribution of CDS for 2059 genes provided to us by the MGC for targeted clone recovery is as follows: A total of 63% of these ORFs are below 2 kb, and 88% of ORFs are below 4 kb. Hence, assuming that the size distribution of these transcripts is a representative sample of all genes, our RT-PCR approach can be used to target approximately 88% of genes. In order to target genes with ORFs longer than 4 kb, we are currently experimenting with long PCR (Barnes 1994; Cheng et al. 1994). Using this method, we have generated a few clones for transcripts between 4 and 6.5 kb CDS length (data not shown). However, it will require more clones per gene to achieve the same rate of rescue in this and the higher size range compared to ORFs less than 4 kb, due to the increased number of RT-PCR errors introduced into longer sequences (Fig. 5).

Although only human genes have been targeted to date, the described approach can be easily applied to other species. Further, the genes we have worked with are known and relatively well characterized. Our current RT-PCR protocol could also be applied with some modifications to target putative or predicted genes, whose existence has not been conclusively proven by experimental data. For these "hypothetical" genes, only genomic sequence is available. Thus, PCR primer design would be based on genomic sequence flanking terminal exons and cross-species genomic sequence comparisons to define exons and conserved noncoding flanking sequences. This is expected to work well for genes where the predicted sequence and structure are accurate. Appearance of the transcript structure in rescued clones should provide assurance of the biological validity of the predicted transcript. However, for ab initio gene predictions, only genomic sequence alignments will be possible with clone sequences, and validation of possible high-quality noncoding discrepancies may not be possible due to the absence of RefSeq and EST data.

The described RT-PCR-based clone acquisition process is a scalable approach for efficient targeted recovery of full-ORF cDNA clones, and can also be applied for isolation and identification of previously unknown splice variants. Using this method, full-ORF cDNAs that could not be obtained previously by systematic EST analysis can now be added to the existing full-ORF cDNA collections. These will provide a valuable addition to the annotation of human and other genomes, as full-ORF cDNA collections move towards completion.

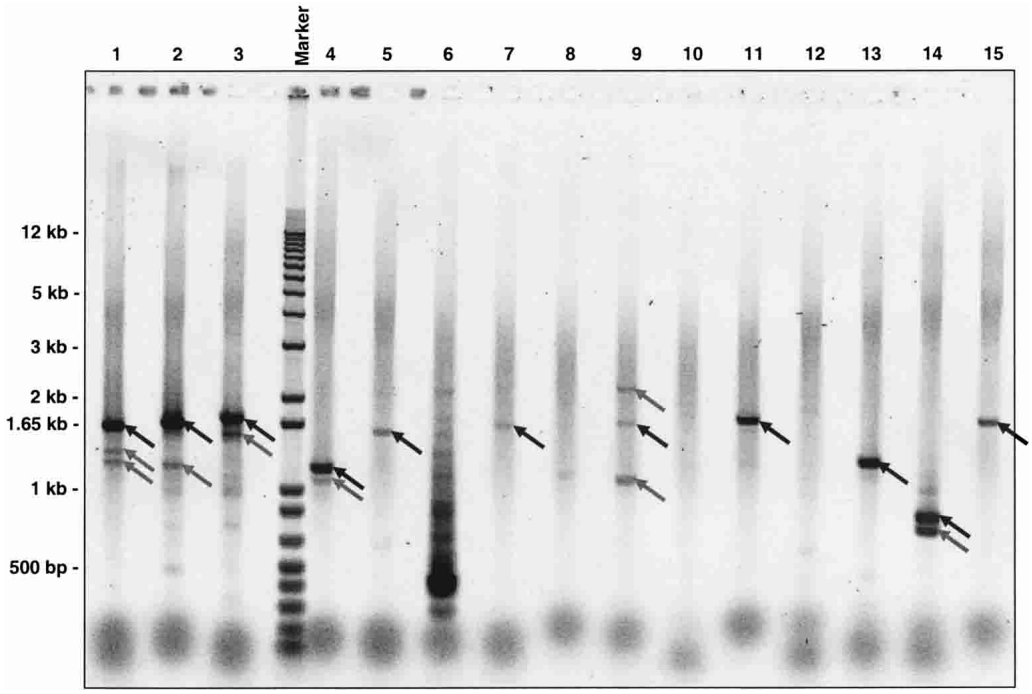


Figure 8 Electrophoretic analysis of PCR-amplified ORFs. PCR amplification was performed using brain cDNA template and gene-specific primers for 96 target genes. The results of 15 amplifications are shown here. Ten µL of a 25-µL reaction was loaded in each well on a 1% agarose gel. Expected-size amplicons of target genes are marked with black arrows. Amplicons different from expected size and isolated as potential splice variants are indicated with gray arrows.

Table 3. Examples of Novel Splice Variants Found

Target RefSeq ID	Gene name	PCR band size	Contig size	Gene structure
NM_000877	*IL1R1*	~1800 bp	1800 bp	Perfect match to RefSeq – E2[a] – 11 of Ensembl
		~1200 bp	1160 bp	Deletion of: part of E3, E4, E5, E6, and 19 bp of E7
NM_001094	*ACCN1*	~1600 bp	1625 bp	One validated discrepancy – Ensembl match: E2–E10
		~1300 bp	1345 bp	E4–E10
		~1200 bp	1192 bp	E4, E5, E7–E10
NM_003160	*STK13*	~900 bp	916 bp	Passed – perfect match – E1–E7
		~1100 bp	1111 bp	E2, E3, E4 (+ 5 bp from intron 5-6), E5, E6, I6-7, E7
		~800 bp	769 bp	E2, E3, E4, E6, E7
		~800 bp	779 bp	E2, E3, E5, E6, E7
NM_006212	*PFKFB2*	~1600 bp	1583 bp	Passed – some validated discrepancies – E1–E6
		~1400 bp	1390 bp	E1 – E11, E14, E15
NM_015310	*EFA6R*	~1650 bp	1667 bp	Passed – perfect match – E1–E13
		~1200 bp	1173 bp	E1, E6, E8 – E13
		~1200 bp	1216 bp	E1, E6 – E13

[a]E and I indicate exons and introns found in the clone inserts. For example, "E1–E11, E14, E15" means the presence of exons 1 to 11, with exon 14 and exon 15 (exons 12, 13, and 16 were not found).

METHODS

cDNA Synthesis

mRNA from various human tissues was obtained from commercial sources (Ambion and Clontech). Double-stranded cDNA was synthesized from 1 µg of mRNA per reaction using oligo-d(T)$_{12\text{-}18}$ primers and the SuperScript Choice System for cDNA Synthesis (Invitrogen) following the manufacturer's protocol. After the second strand synthesis, the reaction was diluted with dH_2O to 200 µL volume, and purified as follows. Two extraction steps were performed in Phase Lock Gel tubes (Eppendorf) using equal volumes of phenol/chloroform. cDNA was precipitated by addition of 133 µL of 7.5 M NH_4OAc, 3 µL glycogen, and 777 µL of 100% ethanol, and overnight incubation at −20°C. The sample was centrifuged at maximum speed in a microcentrifuge at 4°C for 30 min. The pellet was washed twice with 75% ethanol, then resuspended in 20 µL of TE buffer. cDNA quality was inspected by running 1 µL of the resulting 20 µL cDNA solution on a 1% agarose gel, staining with SYBR Green (Mandel), and visualization using a Typhoon 9400 Variable Mode Imager. For subsequent PCR reactions, 0.5 µL of 10-fold cDNA dilution was used as template.

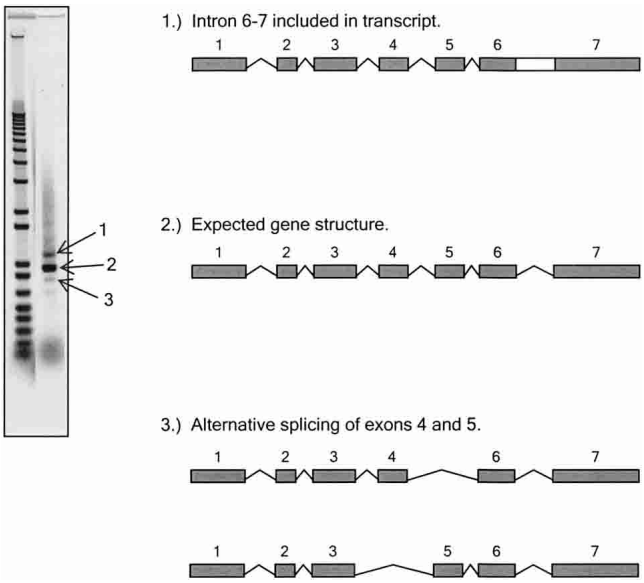


Figure 9 Splice variants found for the aurora kinase C (*AURKC*) gene. Three PCR amplicons that were isolated and cloned yielded four different splice forms. "2" corresponds to the expected gene structure (from RefSeq) of seven exons. "1" includes an extra sequence previously known as an intron between exons 6 and 7. Clones generated from PCR amplicon "3" yielded two different splice forms of similar size, one without exon 5, and one without exon 4.

PCR Primer Design

For a list of target genes, mRNA Fasta sequences and start and stop codon locations were obtained from the Reference Sequence (RefSeq) database (http://www.ncbi.nlm.nih.gov/RefSeq). Gene-specific PCR primers were designed using an automated method utilizing Primer3 (http://www.broad.mit.edu/cgi-bin/primer/primer3_www.cgi; Rozen and Skaletsky 2000), with the following conditions. The optimal melting temperature (Tm) was 60°C, minimum Tm was 55°C, and the maximum Tm was 67°C. The size range was set between 18 and 27 base pairs. The GC content varied between 20% and 80%. The maximum allowable local alignment scores for self-complementarity and 3′ self-complementarity were 8.0 and 3.0, respectively. The primers were within 100 bp nucleotides flanking the ORFs, to ensure that full-ORFs would be amplified.

PCR

Transcripts were amplified by PCR in 96-well plates using gene-specific primers (listed in Supplemental Table S1) flanking the ORFs. PCR was performed in 25 µL reactions, containing 0.5 µL of cDNA (~4 ng) prepared as described above, 5 pmol of each primer, 2.5 µL of 10X HF 2 dNTP mix (Clontech), 1.2 µL of DMSO, 2.5 µL of 10X HF 2 PCR buffer (Clontech), and 0.5 µL of 50X Advantage-HF 2 Polymerase mix (Clontech). The cycling conditions were 95°C for 5 min, followed by 10 touchdown PCR cycles starting with 95°C for 15 sec, 65°C (decreased by 1°C in each subsequent cycle) for 15 sec, 68°C for 1 min/ kb; then 20 cycles of 95°C for 15 sec, 55°C for 15 sec, 68°C for 1 min/kb; followed by an extension at 68°C for 5 min.

Isolation and Cloning of Full-ORF Amplicons

Ten microliters of 25 µL PCR product for each sample was loaded on a 1% agarose gel. The gel was stained with SYBR Green (Mandel) and visualized using a Typhoon 9400 Variable Mode Imager. Visible bands of expected size were excised from the agarose gel,

and purified using the MinElute Gel Extraction Kit (QIAGEN) following the manufacturer's protocol. Purified amplicons were cloned into the pCR4-TOPO vector using the TOPO TA Cloning Kit for Sequencing (Invitrogen) following the manufacturer's protocol. Plasmid vectors containing PCR amplicon inserts were electroporated into DH10B bacterial cells, and recombinant clones were selected on agar plates containing 50 µg/mL ampicillin and 10 µg/mL kanamycin. Glycerol stocks were prepared from 24 individual clone isolates per amplified gene and stored in 384-well plates.

Restriction Fragment Analysis of cDNA Clones

Plasmid DNA was prepared from overnight cultures of clones using an alkaline lysis method as described (Marra et al. 1997) in 96-well plates. Plasmids were digested with EcoRI and analyzed by separation on a 1.2% agarose gel (Marra et al. 1997; Schein et al. 2004). Restriction fragments were identified and sized using both a modified version of automated BandLeader software (Fuhrmann et al. 2003; P. Saeedi, unpubl.), and manual review using Image software (http://www.sanger.ac.uk/Software/Image). The resulting restriction fragments were compared with restriction fragments predicted computationally from the sequence (i.e., in silico digests).

Probability of Finding an Error-Free Clone

The estimated error rate of SuperScript II Reverse Transcriptase is 1/15,000 (this information was provided by the manufacturer, Invitrogen). Advantage-HF 2 high-fidelity DNA polymerase, used for PCR, has an estimated error rate of 1/50,000 (this information was provided by the manufacturer, Clontech). Based on these error rates, the probability of finding at least one acceptable clone among all sequenced clones for a transcript was calculated as follows:

$$P = 1 - (1 - (14999/15000)^{l} * (49999/50000)^{l*c})^{n},$$

where **p** is the probability of finding at least one acceptable clone among **n** sequenced clones, **n** is the number of clones sequenced, **l** is the ORF length in base pairs, and **c** is the PCR cycle number.

Sequencing of cDNA Inserts

Clone inserts were sequenced on an ABI PRISM 3730 XL DNA Analyzer using BigDye primer cycle sequencing reagents (ABI). First, end reads were obtained using M13 -21 and M13 40 reverse primers. Sequence reads were processed and their quality was assessed using *Phred*, and assembled using *Phrap* (Ewing and Green 1998; Ewing et al. 1998). If the end reads were of high quality (defined by Phred quality scores, Supplemental Table S2) and matched the correct gene, the need for additional sequencing reactions using gene-specific sequencing primers was assessed. When needed, gene-specific sequencing primers were designed by manual selection from primer sequences generated by Consed (Gordon et al. 1998), and full-length sequences were obtained via primer walking.

Sequence Analysis

Sequences were analyzed using Consed. In addition, a number of bioinformatics methods have been developed at the Genome Sciences Centre to automate parts of the bioinformatics pipeline (Butterfield et al. 2002). If the end reads were of sufficient quality to be assembled into contigs by the *Phrap* algorithm, they were aligned with the RefSeq sequence of the corresponding target gene using BLASTN (Altschul et al. 1990), with the following parameters: expectation value (E) = 10.0; word size = 11; pairwise alignment used; query sequences filtered by DUST; penalty for nucleotide mismatch = −3; reward for nucleotide match = 1; threshold for extending hits = 0. If the expected RefSeq sequence was the top hit, the need for further sequencing using gene-specific primers was assessed. For ORFs up to 1 kb, end reads with an average phred20 length of 600 were often sufficient to cover the entire clone insert. If the entire ORF was not covered, gene-specific sequencing primers were designed and used to close gaps.

The full-length sequence was aligned against the RefSeq sequence of the targeted gene using CLUSTAL W (Thompson et al. 1994), and high-quality sequence discrepancies were assessed. Average Phred scores of contigs are shown in Supplemental Table S2. If the clone ORF sequence was identical to RefSeq, an acceptable clone was found. Changes that did not result in amino acid differences did not require validation and were acceptable according to the standards established in the MGC pilot project for clone recovery. Validation of protein coding changes was attempted as follows. The sequence was aligned using BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat; Kent 2002) with human genomic sequence and dbEST (http://www.ncbi.nlm.nih.gov/dbEST), and potential polymorphisms were assessed using dbSNP (http://www.ncbi.nlm.nih.gov/SNP) via Ensembl (http://www.ensembl.org; Hubbard 2002; Hubbard et al. 2002; Birney et al. 2004; Kasprzyk et al. 2004). If the clone sequence contained only changes that were validated in these other data sources, the clone was declared acceptable.

ACKNOWLEDGMENTS

We thank D.S. Gerhard, E.A. Feingold, F.S. Collins, R.L. Strausberg, and members of the Mammalian Gene Collection Consortium for helpful discussions and support for this project; J.E. Schein, J.M. Hirst, and the following groups at the British Columbia Cancer Agency Genome Sciences Centre for helpful discussions and assistance: Gene Expression Laboratory, Mapping Group, Sequencing Group, and Bioinformatics Group. This project, as part of the trans-NIH initiative MGC effort, was funded with federal funds from the National Cancer Institute, NIH, #N01-C0-12400. We also thank the British Columbia Cancer Foundation for their support. M.A.M. is a Scholar of the Michael Smith Foundation for Health Research.

The content of this publication does not necessarily reflect the views or policies of the U. S. Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Barnes, W.M. 1994. PCR amplification of up to 35-kb DNA with high fidelity and high yield from λ bacteriophage templates. *Proc. Natl. Acad. Sci.* **91:** 2216–2220.

Birney, E., Andrews, D., Bevan, P., Caccamo, M., Cameron, G., Chen, Y., Clarke, L., Coates, G., Cox, T., Cuff, J., et al. 2004. Ensembl 2004. *Nucleic Acids Res.* **32:** D468–470.

Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. 1993. dbEST—Database for "expressed sequence tags". *Nat. Genet.* **4:** 332–333.

Butterfield, Y.S., Marra, M.A., Asano, J.K., Chan, S.Y., Guin, R., Krzywinski, M.I., Lee, S.S., MacDonald, K.W., Mathewson, C.A., Olson, T.E., et al. 2002. An efficient strategy for large-scale high-throughput transposon-mediated sequencing of cDNA clones. *Nucleic Acids Res.* **30:** 2460–2468.

Cheng, S., Fockler, C., Barnes, W.M., and Higuchi, R. 1994. Effective amplification of long targets from cloned inserts and human genomic DNA. *Proc. Natl. Acad. Sci.* **91:** 5695–5699.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. *Genome Res.* **8:** 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Fuhrmann, D.R., Krzywinski, M.I., Chiu, R., Saeedi, P., Schein, J.E., Bosdet, I.E., Chinwalla, A., Hillier, L.W., Waterston, R.H., McPherson, J.D., et al. 2003. Software for automated analysis of DNA fingerprinting gels. *Genome Res.* **13:** 940–953.

Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8:** 195–202.

Hubbard, T. 2002. Biological information: Making it accessible and integrated (and trying to make sense of it). *Bioinformatics* (Suppl.) **18:** S140.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30:** 38–41.

Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., and Birney, E. 2004.

EnsMart: A generic system for fast and flexible access to biological data. *Genome Res.* **14:** 160–169.
Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. *Genome Res.* **12:** 656–664.
Marra, M.A., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D., and Waterston, R.H. 1997. High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7:** 1072–1084.
Rozen, S. and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132:** 365–386.
Schein, J., Kucaba, T., Sekhon, M., Smailus, D., Waterston, R., and Marra, M. 2004. High-throughput BAC fingerprinting. In *Methods in molecular biology* (eds. S. Zhao and M. Stodolsky), Vol. 255, Bacterial artificial chromosomes, pp. 143–156. Humana Press, Totowa, NJ.
Schuler, G.D. 1997. Sequence mapping by electronic PCR. *Genome Res.* **7:** 541–550.
Schuler, G.D. 1998. Electronic PCR: Bridging the gap between genome mapping and genome sequencing. *Trends Biotechnol.* **16:** 456–459.
Shevchenko, Y., Bouffard, G.G., Butterfield, Y.S., Blakesley, R.W., Hartley, J.L., Young, A.C., Marra, M.A., Jones, S.J., Touchman, J.W., and Green, E.D. 2002. Systematic sequencing of cDNA clones using the transposon Tn5. *Nucleic Acids Res.* **30:** 2469–2477.
Strausberg, R.L., Feingold, E.A., Klausner, R.D., and Collins, F.S. 1999. The mammalian gene collection. *Science* **286:** 455–457.
Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., Altschul, S.F., et al. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci.* **99:** 16899–16903.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

WEB SITE REFERENCES

http://genome.ucsc.edu/cgi-bin/hgBlat; Human BLAT Search.
http://mgc.nci.nih.gov; Mammalian Gene Collection.
http://www.broad.mit.edu/cgi-bin/primer/primer3_www.cgi; Primer3.
http://www.ensembl.org; Ensembl.
http://www.ncbi.nlm.nih.gov/dbEST; Expressed Sequence Tags database.
http://www.ncbi.nlm.nih.gov/RefSeq; NCBI Reference Sequences.
http://www.ncbi.nlm.nih.gov/SNP; dbSNP home page.
http://www.ncbi.nlm.nih.gov/UniGene; UniGene.
http://www.sanger.ac.uk/Software/Image; Image—The fingerprint image analysis system.