Completing the Mammalian Gene Collection: rgeted Large Scale Generation and Analysis of Full-ORF Human cDNA Clones

Á, Butterfield YSN, Coughlin SM, Zeng T, Griffith M, Griffith OL, Petrescu AS, Smailus DE, Khattra J, McDonald HL, McKay SJ, Jones SJM, Holt RA, Marra MA

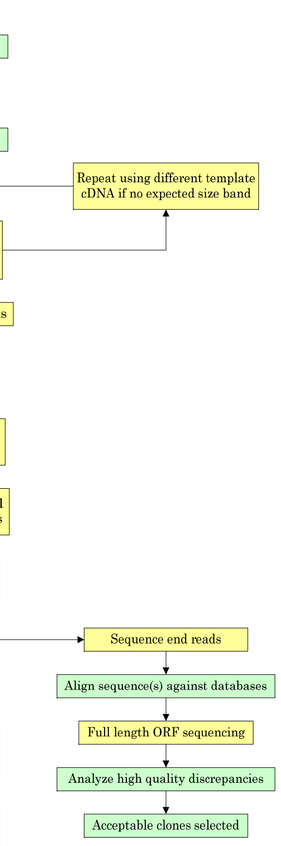
Canada's Michael Smith Genome Sciences Centre
www.bcgsc.ca

n (MGC) consortium (http://mgc.nci.nih.gov) is to collection of all human genes. To date, over 15,200 n reading frames (ORF) have been obtained via lowever, there are still thousands of genes missing ing is no longer effective in finding new clones. As ve have developed a scalable method for the targeted quences. First, cDNA is synthesized from RNA ved by PCR amplification of transcripts using gene mplification step is performed on groups of 96 genes, are cloned into a sequencing vector, and full length or a number of subclones from each gene. Sequences ap and analyzed using Consed. In addition, a number by our group to automate various parts of the h the Reference Sequence (RefSeq) database; then ed using genomic sequence, dbEST and Ensembl eq or containing only validated changes will become containing novel splice variants or polymorphisms lts will make an important contribution to the human esource for researchers further characterizing these e applied to other species as well.

ocess

rgeted 384 human transcripts (four sets of 96, Table Groups of 96 genes were selected based on similarity expression as determined from UniGene (NCBI). from various human tissues (Figure 2). Gene specific d method developed at the Genome Sciences Centre, y PCR in 96-well plates using a high fidelity DNA gel electrophoresis. To date, expected size amplicons 1). Gel fragments of expected size products were asmid vectors were electroporated into bacterial cells, lates containing appropriate antibiotics. Plasmid 2 clones per gene using an alkaline lysis method. nts were separated on an agarose gel and compared t patterns. Twelve clones per gene were entered into obtained using M13-21 and M13 reverse primers were f the end reads were of high quality (defined by Phred he need for additional sequencing reactions using gene ull length sequences were compared with the RefSeq repancies was attempted.

ery process.



2. Clone Acquisition Process (cont'd)

Figure 2. Agarose gel electrophoresis of double stranded cDNA. The sources of RNA are shown on top of the gel. cDNA was synthesized from 1 µg high quality mRNA per sample using the SuperScript Choice System. 1 µl of the resulting 20 µl cDNA per sample was loaded in the five sample wells of a 1% agarose gel. The gel was stained with Sybr Green and visualized using a Typhoon 9400 Variable Mode Imager.

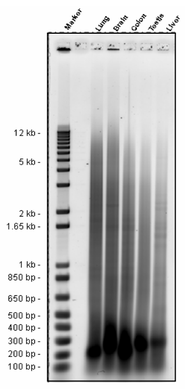
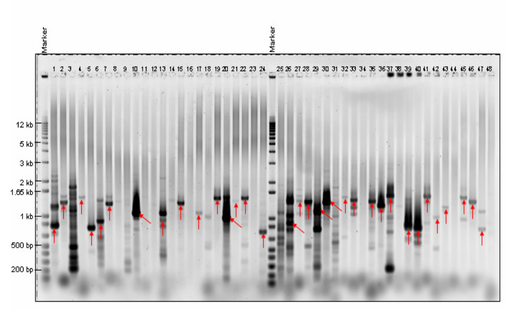
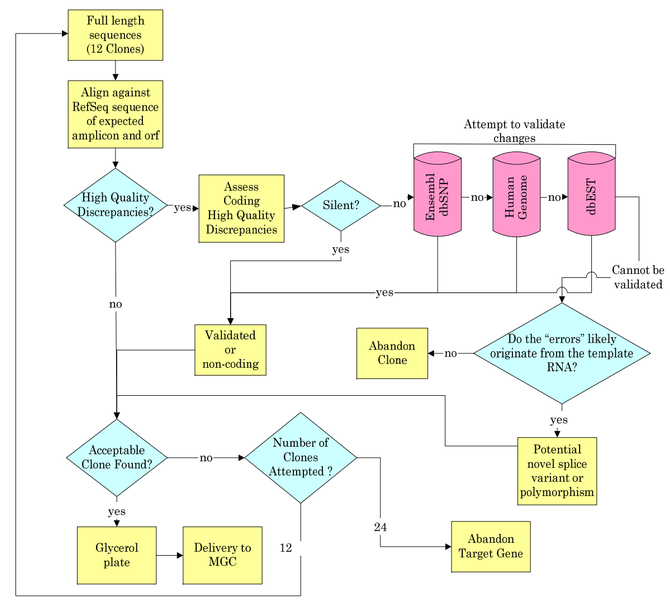


Figure 3. Electrophoretic analysis of PCR amplified ORFs. PCR amplification was performed using lung cDNA template and gene specific primers for 96 target genes. The results of 48 of these are shown below. Ten microliters of a 25 µl reaction for each sample was loaded on a 1% agarose gel. The gel was stained with Sybr Green and visualized using a Typhoon 9400 Variable Mode Imager. Expected size amplicons of target genes are marked with arrows.



3. Sequence Analysis/Bioinformatics

Figure 4. Bioinformatics sequence analysis pipeline.



3. Sequence Analysis/Bioinformatics [cont'd]

The summary of the sequence analysis approach is shown in Figure 4. Sequences are processed and assembled using Phred, Phrap and 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. Once obtained, the full length sequence is aligned against the RefSeq sequence of the targeted gene using an automated method utilizing Clustal W, and high quality sequence discrepancies are assessed. If the clone ORF sequence is identical to RefSeq, an acceptable clone has been found. Changes that do not result in amino acid differences do not require validation and are acceptable according to the standards established in the MGC pilot project for clone recovery. Validation of protein coding changes is attempted as follows. The sequence is aligned with human genomic sequence and dbEST using an automated method utilizing BLAT, and potential polymorphisms are assessed, currently manually, using Ensembl dbSNP. If the clone sequence contains only changes that are validated in these other data sources, the clone is declared acceptable and the gene is considered rescued. To date, we have rescued 258 (67%) of the targeted 384 genes (Table 1).

4. Progress Summary

Table 1. Progress summary of clone recovery using RT-PCR. Acceptable clones contain ORFs with identical sequences to RefSeq, or contain only validated changes. The numbers/percentages indicate numbers/percentages of genes.

Genes targeted	1st set		2nd set		3rd set		4th set		Total	
	#	%	#	%	#	%	#	%	#	%
Attempted by RT-PCR	96	*100*	96	*100*	96	*100*	96	*100*	384	*100*
Expected size amplicons	88	*92*	87	*91*	78	*81*	92	*96*	345	*90*
Clones available	88	*92*	86	*90*	78	*81*	91	*95*	343	*89*
Acceptable clones found	75	*78*	67	*70*	56	*58*	60	*63*	258	*67*
No acceptable clones	13	*14*	15	*16*	19	*20*	15	*16*	62	*16*
Clones pending analysis	0	*0*	4	*4*	3	*3*	16	*17*	23	*6*

Among the genes that did not yield acceptable clones, about two thirds matched a RefSeq sequence other than the targeted gene, and the remaining third matched the targeted ORF. Approximately 20% of the genes with clones matching the target failed due to various unvalidated errors. These may have been introduced by the RT-PCR process itself (the reverse transcriptase used has an estimated error rate of 1/15,000 and the error rate of high fidelity DNA polymerase is 1/50,000). For about 80% of these genes; however, we found that at least half of the clones for the targeted gene shared a common unvalidated error. In cases where clones corresponded to the targeted gene, but the majority could not be rescued due to a common "error", it is likely that we isolated biologically valid expressed sequences, and these changes are not artifacts. 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. We propose further discussion regarding clones affected by these shared non validated errors, and would argue that they likely represent novel splice variants or other real polymorphisms and are biologically valid transcripts produced from these genes.

We plan to scale up the RT-PCR based protocol described above for high throughput clone acquisition. Using this protocol, we were able to obtain expected size amplicons for ORFs up to 4 kb. In order to target genes with ORFs longer than 4 kb, we are currently experimenting with long PCR To further increase the rate of recovery, we are addressing various issues that caused different types of failures. The cases where clone insert sequences matched another gene instead of the targeted, and included the correct PCR primers, indicate that other transcripts of similar size to the target gene and containing homologous sequences were cloned. Generally this is more likely to occur for paralogous gene family members. To reduce these kinds of errors in the future, we plan to further refine our automated primer design process using the publicly available tool electronic PCR (ePCR), that 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. There were a number of genes that could not be rescued due to various errors that could not be validated. For these, more clones will be sequenced, or clones containing the least number of errors will be repaired by use of site directed mutagenesis. In genes where the same error failed at least 50% of the clones, we believe that these kinds of "errors" are not likely to be introduced by RT-PCR, but rather originate from the template and are biologically valid expressed sequences.

5. Splice variant discovery

We often observed multiple bands on the PCR gel instead of the expected size full-ORF amplicon alone (Figure 3). To test whether some of these represented additional splice variants of the target gene, we cloned some of these extra bands from the 4th targeted set of 96, and further processed them similarly to the expected size amplicons (Figure 1). An agarose gel containing a subset of these amplicons is shown in Figure 5. Sequence analysis revealed, that in fact, most of these extra bands contained potential alternative splice forms of the targeted gene (Table 2). In some instances, one isolated PCR band yielded more than one splice variant (Table 2, Figure 6). In these cases, the sizes of splice variants were very similar, such as they could not be resolved by agarose gel electrophoresis of PCR products.