A Guide to Finding Mammalian Gene Collection (MGC) Clones and Evaluating Their Sequence

Part A. MGC Clone Search. A variety of ways exist to determine whether MGC cDNA clones are available for human, mouse, and rat genes and transcripts of interest. Here we describe three approaches. We illustrate by searching for MGC cDNA clones for protein-coding transcripts from the human gene SERPINA1, encoding α-1-anti-trypsin protein.

(As per HUGO-defined convention, all letters of human gene names are capitalized, but only the first letter of mouse and rat gene names is capitalized. Entries into the search engines described below are case insensitive.)

Approach 1. The MGC homepage (Figure 1) provides several search tools. You can search for individual full-protein-coding (full-cds) human, mouse, rat, or bovine clones from this page using gene names or key words. Entering SERPINA1 into the Enter Gene Symbol box (Figure 1, arrow 1) opens a page that shows two MGC clones are available, BC011991 and BC015642 (Figure 2), together with names of the libraries from which they were isolated and links to associated vector and source tissue information.

Figure 1 MGC Website Home Page

[Diagram of MGC Website Home Page]

---

1 All newly isolated MGC, XGC, and ZGC clones are assigned a “BC” accession when their sequence is submitted to GenBank, but only a subset of these candidate clones has a full-CDS. Once a candidate clone is confirmed by full-length sequencing to have a full-CDS (without changes altering the phase of reading frame, the position of the translational start ATG codon, or the position of the stop codon) it is then assigned the Keyword “MGC” in a new GenBank record.
The MGC homepage also provides lists of MGC cDNA clones, as well as lists of MGC cDNA libraries (Figure 1, arrow 2). More detailed lists of MGC clones are provided at the MGC ftp site, described in Part C, below.

(The left side of the MGC Homepage provides links to the XGC and ZGC pages, with similar search functions for Xenopus tropicalis and Xenopus laevis clone or Danio rerio (zebrafish) clones.)

Clicking on the link for BC011991 leads to the GenBank record for this clone in Entrez Nucleotide (Figure 3a). This page provides many details about MGC cDNA clones, including information on how the clone was obtained, the cloning vector, tissue source, nucleotide sequence, and expected translated amino acid sequence. Near the top of the page, the Definition line (Figure 3a, arrow 1) notes that this clone includes a complete protein-coding sequence (cds).

(If this MGC clone had been prepared with a synthetic DNA insert, the Definition line also would indicate whether the natural protein-coding sequence was cloned with or without a stop codon, as illustrated by BC167860 and BC140303.)

The existence of RefSeq alternative splicing isoforms for this gene is noted on the right side of this page (Figure 3a, arrow 2), with links to information on the alternative splice isoforms. Further down on the right side, All links from this record shows links to other resources. These include Order cDNA Clone (Figure 3a, arrow 3), which connects to a clone order page with links to several commercial distributors of these two clones.

Other links under All links from this record connect to data and analysis tools, such as dbSNP, GEO, Map Viewer, PubMed, UniGene, UniSTS, OMIM, and LinkOut. LinkOut connects to gene-related research materials from commercial vendors, including antibodies, peptides, and siRNA reagents.
Figure 3a  GenBank Record for BC011991

Homo sapiens serpin peptidase inhibitor, clade A (alpha-1 antitrypsin, antitrypsin), member 1, mRNA (cDNA clone MGC:9222 IMAGE:3859644), complete cds

Figure 3b  GenBank Record for BC011991 (continued)

Homo sapiens serpin peptidase inhibitor, clade A (alpha-1 antitrypsin, antitrypsin), member 1, mRNA (cDNA clone MGC:9222 IMAGE:3859644), complete cds
Approach 2. A second way to find an MGC clone, such as for the human gene SERPINA1, is to start at the NCBI homepage. First, choose Gene in the database Search dropdown menu (Figure 4, arrow 1), then type the gene symbol and the organism (serpina1 [SYMBOL] AND human [ORGANISM]) or enter the gene id for human SERPINA1 (5265) into the box (Figure 4, arrow 2) and click on Search. This will take you to the Entrez Gene page for human SERPINA1. On the right side of this page, under LINKS (Figure 5, arrow 1), Order cDNA Clone is displayed; this link leads to the same clone ordering information described above. In addition, the Gene page provides a wealth of other valuable information about the SerpinA1 gene.

Approach 3. You can create custom lists of MGC clones by performing searches from the NCBI Homepage using different combinations of search terms. For example, to search for all MGC full-cds clones, entering “MGC [KEYWORD] AND human [ORGANISM]” (Figure 6, arrow) will produce a list all ~30,000 human clones in the MGC. Adding other qualifying terms further restricts the search: “MGC [KEYWORD] AND human [ORGANISM] AND "peptidase inhibitor" yields a list of human MGC clones for peptidase inhibitors. Suggestions for searching the NCBI databases are given at Entrez Help.

Related resources:
“Tips for Finding cDNA Clones” is an NCBI page with details on locating MGC clones. The UCSC Genome Bioinformatics website offers a training page with many helpful tutorials and guides to using the UCSC genome browser, including a tutorial, “Fishing for Genes in the UCSC Browser,” with advice on finding information on MGC clones.

Part B. Evaluating MGC Clone Sequence Integrity. The annotation of MGC clone sequences in GenBank records is based on comparing the cDNA clone sequence to the genome sequences. For example, the GenBank record for the SERPINA1 MGC clone BC015642, under FEATURES, notes in the misc_difference category (Figure 3b, arrow) that this clone contains a T>C change at position 737, with valine encoded by the genome and alanine by the cDNA; it also states that “the chimpanzee genome agrees with the cDNA sequence, suggesting that this difference is unlikely to be due to an artifact.”
Figure 5  Gene Page for SerpinA1

Figure 6  A Custom Search for MGC Clones
Annotations in the MGC clone GenBank records were frozen on March 23, 2009 at the conclusion of the MGC project. As a result, subsequent revisions of the human, mouse, and rat genomes that result in changes to annotations of the protein-coding sequence (CDS) of a gene will not be reflected in the GenBank records for these MGC clones. New annotations might, for example, reposition the ATG translation start codon so it now lies upstream (5') to the start ATG noted in the GenBank record.

Prior to ordering an MGC clone, therefore, users should compare the MGC cDNA sequence against the most recent genome build, to determine whether more recent data have altered the cds annotation of the MGC clone. A variety of tools are available to perform such analysis, four of which are described here.

Tool 1. A detailed comparison of MGC clone sequences to the current reference genome and RefSeq transcripts is available from Evidence Viewer. The Entrez Gene page for SERPINA1, for example, provides links (Figure 5, arrow 2) to Evidence Viewer, which displays an extensive list of SERPINA1 RefSeq transcripts, MGC clones, and other clones, including alternative isoforms, together with details on nucleotide mismatches and indels associated with each.

Tool 2. A second way to display differences between an MGC clone sequence and the current version of the genome is to do a BLAST search of the MGC clone against the databases which harbor genome and RefSeq sequences. Although BLAST can be accessed from the NCBI homepage (Figure 4, arrow 3), a convenient BLAST link is also provided on the upper right side of GenBank record (Figure 3a, arrow 4). Clicking on the BLAST link for BC015642.2 takes the user to the BLAST page, where one can choose to compare this clone sequence against the human genome and transcript databases. Doing so reveals a single nt difference in the alignment between the MGC clone and its RefSeq transcript homolog, NM_001127707.1, at position 737. Thus the annotated difference at position 737, noted on the MGC clone page is current.

Tool 3. SPLIGN provides a third way to align a cDNA clone sequence to the current version of a reference genome, with exon-by-exon graphics and sequence displays. If you Enter BC015642 in the cDNA box and select homo sapiens in the pull-down menu under Genome, SPLIGN generates a display of the exon structure of this transcript, showing a single mismatch in exon 3 (Figure 7, arrow 1). Clicking on exon 3 reveals its sequence, with the T>C change at position 737 highlighted in red (Figure 6, arrow 2).

Tool 4. The UCSC Genome Browser provides detailed graphical views and alignment statistics of MGC clone sequences against their respective genomes. Entering the gene name SERPINA1 or the clone accession number BC015642 into the position or search term box of the UCSC Genome Browser Gateway page (Figure 8, arrow) and clicking submit leads to a schematic view of clones for SERPINA1 aligned against the human genome, including all relevant MGC clones (if you had first activated the MGC Genes track). A red line in exon 3 of the schematic line-figure of BC015642 signals a sequence difference from the reference genome.

Clicking on the line-figure of BC015642 connects you to its clone GenBank record, with summary statistics on alignments of the clone sequence to the Genome Reference Consortium (GRC) human genomic sequence and to multiple RefSeq mRNAs. This clone page shows that the exon structure of BC015642 shares 100% identity with multiple RefSeq RNAs and that its sequence is 99.93% identical to the reference human genome sequence on chromosome 14. Clicking on "BC015642:1-1371," under "mRNA (alignment details)," displays a single CDS nt difference at position 737.

The web pages noted above provide many other kinds of information and useful links, besides the ones described. For example, at the upper right of an MGC clone GenBank
record, such as for BC015642, under **Customize View** (Figure 3a, arrow 5) a user can activate the **Features added by NCBI** (followed by **Update View**) to display the 37 SNPs listed in dbSNP that overlap with this cDNA sequence.
Figure 7  Splign Alignment of BC015642 vs. Reference Human Genome

Figure 8  USCS Genome Browser Gateway Page
Part C. Information on MGC Clones Posted at the NCBI-MGC ftp site:
(A Glossary of Abbreviations is provided below)

Supplementary Table A: Lists targets and results of MGC PCR-Rescue (RT-PCR cloning) and DNA synthesis: Rows: Assigned transcripts. Columns: Organism, rescue_uid (PCR rescue ID), laboratory, geneid, target nuc_acc.ver, target nuc_gi, target prot_gi, target cds_length, mgc_acc, mgc_acc_other_gene, non_mgc_acc.

Supplementary Table B: Lists clones produced by the MGC Project which are Not Full-CDS: Rows: information on each transcript accession. Columns: organism, taxid, MGCid, imageid, nuc_acc, nuc_acc.ver, nuc_gi, nuc_len, prot_acc, prot_acc.ver, prot_gi, prot_length, geneid, gene symbol, gene name, gene type.

Supplementary Table C: Total MGC Full-CDS Collection: Rows: information on each transcript accession. Columns: organism, taxid (species code), MGCid, imageid, nuc_acc, nuc_acc.ver, nuc_gi, nuc_len, prot_acc, prot_acc.ver, prot_gi, prot_len, geneid, gene symbol, gene name, gene type.

Supplementary Table D: List of protein-coding genes missing from the MGC full-cds collection: Rows: Genes that have RefSeq transcript with prefix NM_ or XM_, ordered by organism, and geneid. Columns: organism, geneid, gene_symbol, gene_name, longest_CDS_length_of_NM_RefSeq, NM_RefSeq_with_longest_CDS, list of RefSeq_accession.ver

Supplementary Table E: List of OMIM Disease Genes Represented in MGC: Rows: information on each transcript accession. Columns: OMIM_id; OMIM_DESCR; human_GeneID; human_MGC_accession.ver; HomoloGene_id; mouse_GeneID; mouse_MGC_accession.ver; rat_GeneID; rat_MGC_accession.ver

Glossary of Abbreviations in Supplementary Tables (above):
rescue_uid (PCR rescue ID)
geneid (gene ID)
target nuc_acc.ver (RefSeq transcript accession, with version number after decimal point)
nuc_gi (gi number that corresponds to the RefSeq transcript)
nuc_len (length of entire cloned cDNA, including linker sequences and, in some cases, 5’ and/or 3’ UTR sequences)
target prot_gi (protein gi number corresponding to RefSeq transcript)
mgc_acc (accession of full-cds clone for intended RefSeq transcript)
mgc_acc_other_gene (accession of full-cds clone from gene other than the target gene)
non_mgc_acc (accession of a partial-cds or no-cds clone)

Individual MGC, XGC, & ZGC Project Data
All MGC, Xenopus Gene Collection, and Zebrafish Gene Collection clones and their related sequences are included and grouped by organism, with two files for each: “MGC full-CDS” clones “MGC Other” clones. “MGC Other” clones include partial CDS clones; short-CDS clones (CDS < 50% of the longest RefSeq transcript CDS); and clones for transcripts lacking annotated CDS, including possible pseudogenes. Files include information about the clone and sequence in tab-delimited format; fasta nucleotide sequences; GenBank flatfile for nucleotide record; protein fasta sequences; and GenBank flatfile for protein record.
**Part D. Gene Expression Data Related to MGC Clones.** Gene expression profiles for MGC clones can be accessed directly through the GenBank clone page links to UniGene, located on the lower right side of the page (under **All links from this record**). For BC015642, this link leads to a UniGene page that displays both GEO and EST profiles of SERPINA1 gene expression.

mRNA, EST, SAGE, and microarray expression profiles for MGC genes are also displayed on the **UCSC Genome Browser**, when these tracks are activated.

**AceView** provides gene expression information, including tissues of origin for cDNA clones and average expression levels of transcripts.

**GeneNote** provides gene-specific microarray expression profiles. See, for example, the **results** for SERPINA1.

Other sources of gene expression profiles include: **CGAP Tools; MGC-GXD; GENSAT; BioGPS; and DAVID.**

A rapid way to survey all NCBI databases for information about a gene or a specific cDNA clone (once its accession is identified) is to search “All Databases” for a gene name or clone accession, from the **NCBI home page.** This will reveal on one page whether, for example, expression data are available in GEO, markers exist in UniSTS, or TaqMan probes are listed in Probe.