Assembly, Verification, and Initial Annotation of the NIA Mouse 7.4K cDNA Clone Set

Vincent VanBuren,[1] Yulan Piao,[1] Dawood B. Dudekula,[1] Yong Qian,[1] Mark G. Carter,[1] Patrick R. Martin,[1] Carole A. Stagg,[1] Uwem C. Bassey,[1] Kazuhiro Aiba,[1] Toshio Hamatani,[1] George J. Kargul,[1] Amber G. Luo,[1] Janet Kelso,[2] Winston Hide,[2] and Minoru S.H. Ko[1,3]

[1]*Developmental Genomics and Aging Section, Laboratory of Genetics, National Institute on Aging (NIA), National Institutes of Health, Baltimore, Maryland, 20892, USA;* [2]*South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa*

A set of 7407 cDNA clones (NIA mouse 7.4K) was assembled from >20 cDNA libraries constructed mainly from early mouse embryos, including several stem cell libraries. The clone set was assembled from embryonic and newborn organ libraries consisting of ~120,000 cDNA clones, which were initially re-arrayed into a set of ~11,000 unique cDNA clones. A set of tubes was constructed from the racks in this set to prevent contamination and potential mishandling errors in all further re-arrays. Sequences from this set (11K) were analyzed further for quality and clone identity, and high-quality clones with verified identity were re-arrayed into the final set (7.4K). The set is freely available, and a corresponding database was built to provide comprehensive annotation for those clones with known identity or homology, and has been made available through an extensive Web site that includes many link-outs to external databases and analysis servers.

[The sequence data from this study have been submitted to GenBank under accession nos. BQ550036–BQ563104.]

Increasing interest in stem cells and early embryos requires a high-quality cDNA clone set for cDNA microarray experiments as well as downstream molecular biological studies of gene action. The application of DNA microarrays is strengthened when high-quality cDNA clones corresponding to array elements are readily available for further investigations. Genes that are uniquely expressed in these early embryonic cell types, however, are usually underrepresented in available cDNA clone sets (Ko 2001). To address this issue, we assembled and released the NIA mouse 15K cDNA clone set previously (Tanaka et al. 2000; Kargul et al. 2001), which was mainly derived from preimplantation and early postimplantation mouse embryos. The clone set has been used in over 200 cDNA microarray facilities worldwide. To complement this clone set and increase the coverage of genes, we have assembled and characterized an additional set of 7407 cDNA clones (NIA mouse 7.4K) from mouse cDNA libraries derived from various stem cell lines, preimplantation embryos, and newborn organs (Fig. 1A). The combined sets therefore represent an estimated 19,000 genes (see below), including a large number of genes uniquely transcribed in stem cells, and embryonic and newborn organ cell types.

[3]Corresponding author.
E-MAIL kom@grc.nia.nih.gov; FAX (410) 558-8331.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.633802. Article published online before print in November 2002.

RESULTS AND DISCUSSION

Assembly and Handling of NIA Mouse 7.4K

Approximately 11,000 cDNA clones (11K) were chosen from a collection of mouse early embryonic, stem cell, and newborn organ libraries as nonredundant representatives of these libraries. To prevent contamination and ensure accurate identities in a final assembly of cDNA clones, clones were handled in individual tubes after the first assembly of a quality-control 11K set. Tubes containing high-quality, sequence-verified clones (see Methods) were then moved to registered positions to form the NIA mouse 7.4K cohorts.

Verified Identity and Annotation for NIA Mouse 7.4K

Careful curation and analysis were applied (see Methods) in selecting 7407 cDNA clones from among 20 cDNA libraries constituting a total of ~120,000 cDNA clones. The resultant clone set has 100% verified sequence identity (clones matching the parental cDNA clones). Sequences are provided along with annotation on our Web site (see URL A in Methods section). The average high-quality (phred score ≥20) read length for representative sequences in the clone set is 470 bp. Other investigators can assess sequence quality throughout the NIA mouse 7.4K clone set from a graphic display of sequence quality reports arranged according to the plate location of clones (Fig. 2; see URL A in Methods).

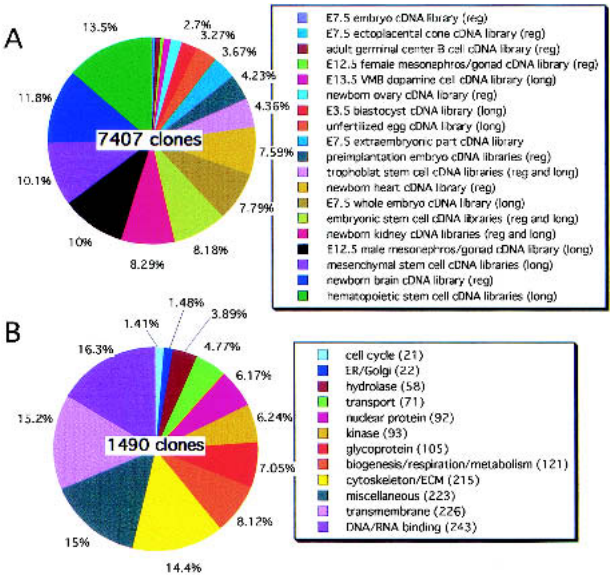

Figure 1 Summary descriptions of the NIA mouse 7.4K clone set. (*A*) Library Composition of NIA mouse 7.4K. (*B*) Functional classification of 1490 clones with mouse gene identity in RefSeq (BLAST ≥400) or homology in NCBI nr (BLAST ≥100).

Annotation for the NIA mouse 7.4K clone set is available over the Internet through a Web-based interface (URL B) that provides links to other databases and analysis servers (Fig. 3), and as a downloadable tab-delimited file (URL A). Of the 7407 cDNA clones, 3174 (40%) known genes (as outlined in Methods) were identified by comparison of cDNA clone sequences to NCBI RefSeq and NCBI nr using BLAST. An alternative, less stringent approach to determine gene identity/homology (see Methods) found NCBI RefSeq identities for 926 clone sequences and NCBI nr homologies for another 3509 clones, therefore providing tentative identity/homology annotation for ~60% (4435) of the clone set. The available annotation for the set (URL A and B) used the more stringent criteria for 3174 known genes. SWISS-PROT keywords and Gene Ontology (GO) annotation (Ashburner et al. 2000), as well as links to TIGR and NCBI Unigene gene expression indices, are all provided on our Web site (see URL B). Nucleotide sequences may also be searched against the NIA cDNA clone database using BLAST (see URL C).

To assess agreement of sequences from the NIA mouse 7.4K with genomic sequence, as well as to establish the chromosomal location of transcripts in the set, cDNA clone sequences were compared with genomic sequence using BLAST. Most clones aligned with the UCSC mouse draft genome (http://genome.cse.ucsc.edu/), with 6467 clones (~87% of set) having at least one (but often more than one) spliced segment (likely exon) aligning to the genome with an E-value $\leq 1e-100$.

For a first-pass assessment of the functional composition of the clone set, the 7.4K cDNA clones were clustered by SWISS-PROT keywords (Bairoch and Apweiler 2000). Of 3445 clone sequences submitted previously to GenBank, ~43% (1490) were found to have known functional identity (SWISS-PROT keywords relevant to function). At this time (7/11/02), GO annotation was available for only 723 clones in the NIA mouse 7.4K, so clustering by SWISS-PROT keywords was chosen as a more comprehensive classification strategy. For clones with identities or homologies (4435), keywords were collected by automated annotation of those clones with previously assigned GenBank accession numbers (3445), using the Web-based EST Annotation Machine (http://bio.ifom-firc.it/EST_MACHINE/index.html). The 1490 clones found to have keywords related to function were clustered using the Web-based Keyword Clustering Machine (http://bio.ifom-firc.it/KW_CLUST/index.html). The clustered groups were then examined manually to assess major functional categories (Fig. 1B).

The NIA mouse 7.4K cDNA clone set, ~60% of them novel (see Methods), is nonredundant both within the new set and with the verified sequences from the NIA mouse 15K clone set (Tanaka et al. 2000; Kargul et al. 2001). Given that ~12K cDNA clones of NIA mouse 15K are unique, these two sets together represent ~19K unique cDNA clones. NIA mouse 7.4K is composed of both 2652 cDNA clones with an average insert size of ~1.5 kb (Tanaka et al. 2000) and 4755 clones enriched for long inserts, with an average insert size ~2.5–3.0 kb (Piao et al. 2001). The clone sets are open to the research community without restriction, and are therefore expected to help stimulate the exploration of mammalian embryology, aiding in the construction of DNA microarrays and the biological analysis of clones of interest identified in microarray experiments (see URL A for clone set availability).

METHODS

Assembly and Verification

All cDNA clones were constructed with the pSPORT1 vector (Invitrogen) as described previously (Piao et al. 2001). Briefly, double-stranded cDNAs were synthesized with an oligo-dT primer (Invitrogen: 5′-pGACTAGTTCTAGATCGCGAGCGG CCGCCCTTTTTTTTTTTTTT-3′) from extracted RNA, treated with T4 DNA polymerase, and purified by ethanol-precipitation. The cDNAs were ligated to Lone-linker LL-Sal4, purified by phenol/chloroform, and separated from free linkers by Centricon 100. The cDNAs were then digested with *Sal*I and *Not*I enzymes, and cloned into the *Sal*I/*Not*I site of pSPORT1 plasmid vector. The DH10B *Escherichia coli* host was transformed with the ligation mixture by the standard chemical method (Piao et al. 2001) (see URL A for more details pertaining the construction of specific libraries).

About 120,000 3′-ESTs, ~70,000 of which were newly collected from more than 20 new cDNA libraries, were clustered into ~23,000 unique genes using StackPack (Electric Genetics), which uses a series of tools including d2-cluster (clustering), Phrap (assembly/alignment), and Craw (alignment analysis) (Christoffels et al. 2001). After removing genes that were already represented in the NIA mouse 15K cDNA clone set, ~11,000 unique cDNA clones were re-arrayed into 96-well microtiter plates. To avoid well-to-well contamination, all the clones were then subjected to one-round of single colony isolation and stored in individually labeled tubes. Clone identity was verified by comparing sequence reads from both 5′- and 3′-ends before and after single colony isolation using BLAST (Altschul et al. 1997), and clones without reads matching the appropriate parental sequences were removed from the final set.

Verified clones (9218) were analyzed further using their 5′- and 3′-ESTs to remove redundancies and mitochondrial sequences, and to limit low-complexity sequences and repeat sequences. First, we searched for sequence similarity between sequences within the present clone set and with NIA mouse 15K, and removed redundancies where similarity with a BLAST score ≥300 was discovered. Second, clones with a BLAST score ≥200 against the NCBI nonredundant gene col-

Sequence Information of H4022C04-5

[NIA Mouse Genomics Home Page] [NIA Mouse cDNA Project Home Page] [NIA 15k Mouse cDNA Clone Set Page]

NCBI 'nt' Database Top Hit	Mus musculus zinc finger protein 113 (Zfp113), mRNA	*BLAST SCORE*	1572.00
Mouse Genome Top Hit	chr5	*BLAST SCORE*	1604.04
Mouse 'RefSeq' Database Top Hit	Mus musculus zinc finger protein 113 (Zfp113), mRNA	*BLAST SCORE*	1526

NAMES DATA Chromatogram View of the Sequence

CloneName	Sequence Name	SANBI Cluster	Library Description	Reference
H4022C04	H4022C04 -5		NIA Mouse 7.4k Clone Set	Unpublished

SEARCH SEQUENCE *LENGTH* 817 bp *SELFSCORE* 1571 *% of Low Complexity* 0.00

The trace file for this sequence was read by PHRED and the bases were called with an average 99% accuracy. We attempt to trim the vector, Not1 site and PolyT from this sequence. All data processing e.g. BLASTs (hence the name SEARCHSEQ) and RepeatMasking is done using this sequence.

TATCCGCTGTGTGGGTCTGTAACAGGGTCCAACAGGACCTGTTTCGTGTCCGGCCCCCTAAGTCTTACGTCCCTGTCACCAGGTCCTTCAAAACCTGCTG
CTTGCTGAGTCTACAGCTAGCCGACCAGTGACCAGAGCATTTCCCCCTCGCCCATGGAAACCCAGGCAGACCATGCATCTCAGGCACCTCTGCCCCTGTT
AGAGAGCGCTCTTTCTTCTTCAAAAGTCCCTTCCTTTCCTGACAAGGACAGTCTGGGGGACAAGATGTTGGCTGCTGCGCTTCTAAAGGCCAAGTCTCAG
GAGTTGGCGACCTTTGAGGATGTAGCTGTGTACTTCATCTGGAGAGAGTGGAAGCGTCTGGAGCCTGCACAAAGGGACCTCTACAGAGATGTGATGCTGG
AGAACTATGGGAACGTGTTCTCACTGGATGGCGATTCCAAAACTGGAAATGATCGAGTAATCTCTGAAGGTATGGGGTCATGTGAGATGATACTGGGGCG
ATTCCAAAAGGATGTTTCTCAGGGTCTTAAGTTTGAAGAAGCCTACGAACAAGAAGTCAGTCTGCAAAGGCAGCTAGGGAATTCCTCTGTTGGAAGACTG
AATAGGAAAATACAAGAGTTTCACCAAGTTAGAGTTGAGGAAAAGCTCACCCATATTGGAGAAAGAAATAAGAAGTACAGTGAGTTTGGAAACAGCTTCA
CTGTGAATTCCAAACTTATTTCACATCAAAGACTTCAGATGGGAGACAAACCCCATAAGTGTGATGAATGTAGCAAGAGTTTTAATCGTACTTCAGACCT
TATTCAACATCAGAGAA

Blast Search Sequence Against NCBI Databbases
Blast Search Sequence Against NCBI Genome Trace Databases
Run NCBI ORF Finder For Search Sequence
Blast Search Sequence Against Ensembl Mouse Genome Database
Blast Search Sequence Against Human Novel Transcripts
Find Initiation Codons in Search Sequence
Search Sequence Blast against Human Genome

Blast Search Sequence Against NCBI Protein Databases
Blast Search Sequence Against NIA Mouse cDNA Database
Blast Search Sequence Against Human Genome Database
Blast Search Sequence Against TIGR Mouse Gene Index
Translate Search Sequence to Protein Sequence
Find EST Frequency
Search Sequence Blat against Mouse Genome

Figure 2 Sample screen shot of online annotations and links for a cDNA clone sequence in NIA mouse 7.4K.

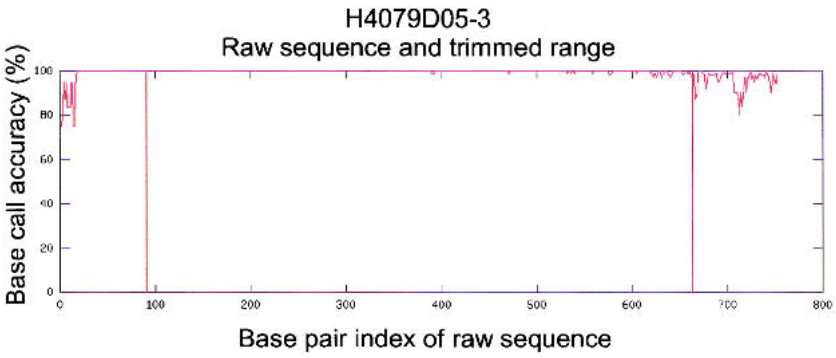

Figure 3 3′ sequence quality for clone H4079D05 (example). Sequence quality for each base is given as a percent accuracy derived from phred scores. Vertical lines mark the ends of quality-trimmed sequence, whereas the final high-quality sequence lies between the lines.

lection (nr) with "mitochond" in the description were removed from the set. Third, RepeatMasker (http://www.geospiza.com/products/tools/repeatmasker.htm) was used to find known repeats in the sequences, and clones with ≥85% repeat sequence or sequences with <50 bp of nonrepeat sequence information were removed. Finally, DUST (Hancock and Armstrong 1994) was used to screen out clones with ≥40% low-complexity sequence. This left 7407 clones in the set. The final 7.4K clone set was re-arrayed from the ~11K sequence verified set by simply rearranging tubes, thereby reducing the potential for contamination during clone selection. Our analysis and clone set assembly strategy is outlined in Figure 4.

Annotation

Two approaches were used to provide annotation for the NIA mouse 7.4K clone set and approximate the number of previously characterized genes within the set. First, to determine the number of known genes in the set, we used BLAST to find masked sequences with significant hits against NCBI RefSeq (Pruitt and Maglott 2001) or NCBI nr. These filtered BLAST hits often returned multiple alignments because of the interruptions in the masked sequences, so significant alignments were realigned using BLAST with unfiltered sequence to get one contiguous alignment. Sequences with likely known identities were assessed by determining those query sequences with at least 80% alignment to a RefSeq or nr target, and within this alignment, those sequences with at least 90% identity with the target sequence. Second, we established RefSeq mouse gene identity for clones with a BLAST alignment score ≥400 against a RefSeq entry. Continuing this second approach, putative homologies were established for clones (those without RefSeq identity) with a BLAST score ≥100 against a NCBI nr entry. The first approach above produced the more stringent result, and was the method chosen for initial annotation (see Results and Discussion).

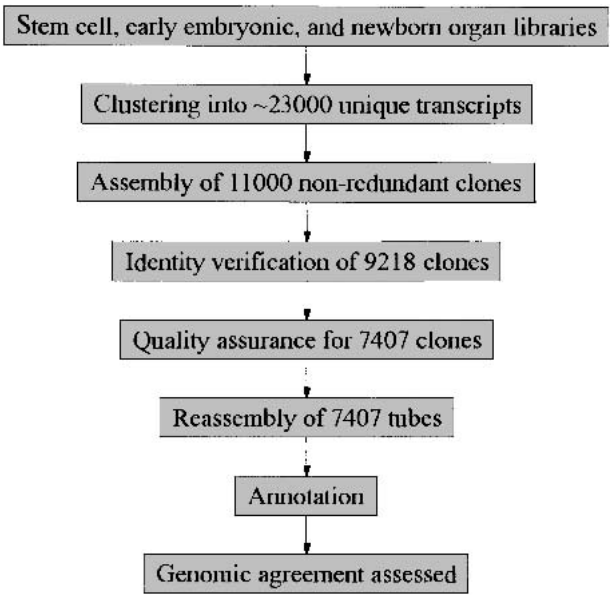

Figure 4 Strategy summary for analysis and assembly of NIA mouse 7.4K.

URLs

(A) Information on clone set generation and composition, as well as information on distribution of the freely available NIA mouse 7.4K cDNA clone set, can be found at: http://lgsun.grc.nia.nih.gov/cDNA/NIA_7_4k.html.

(B) Detailed annotations for each clone can be searched by clone name or keyword at: http://lgsun.grc.nia.nih.gov/cgi-bin/pro1.

(C) BLAST searches against NIA cDNA collections and libraries may be performed at: http://lgsun.grc.nia.nih.gov/cgi-bin/pro8.

ACKNOWLEDGMENTS

We thank T.S. Tanaka, T. Yoshikawa, W.L. Kimber, S.A. Jaradat, R. Matoba, A. Sharov, and R. Nagaraja for valuable help and discussion; M.A. Espiritu, A. Ebrahimi, J.J. Evans, S.J. Olson, M. Roque-Briewer, and N. Caffo at Applied Biosystems for contract-based sequencing; S. Chacko for setting up the mouse genome database on Biowulf; D. Schlessinger for critical review of the manuscript; and D.L. Longo and R.J. Hodes at NIA for encouragement and support. This study used the high-performance computational capabilities of the Biowulf/LoBoS3 cluster at the National Institutes of Health.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25(17):** 3389–3402.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25(1):** 25–29.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28(1):** 45–48.

Christoffels, A., Gelder, A., Greyling, G., Miller, R., Hide, T., and Hide, W. 2001. STACK: Sequence tag alignment and consensus knowledgebase. *Nucleic Acids Res* **29(1):** 234–238.

Hancock, J.M. and Armstrong, J.S. 1994. SIMPLE34: An improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Comput. Appl. Biosci.* **10(1):** 67–70.

Kargul, G.J., Dudekula, D.B., Qian, Y., Lim, M.K., Jaradat, S.A., Tanaka, T.S., Carter, M.G., and Ko, M.S.H. 2001. Verification and initial annotation of the NIA mouse 15K cDNA clone set. *Nat. Genet.* **28(1):** 17–18.

Ko, M.S.H. 2001. Embryogenomics: Developmental biology meets genomics. *Trends Biotechnol.* **19:** 511–518.

Piao, Y., Ko, N.T., Lim, M.K., and Ko, M.S.H. 2001. Construction of long-transcript enriched cDNA libraries from submicrogram amounts of total RNAs by a universal PCR amplification method. *Genome Res.* **11(9):** 1553–1558.

Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29(1):** 137–140.

Tanaka, T.S., Jaradat, S.A., Lim, M.K., Kargul, G.J., Wang, X., Grahovac, M.J., Pantano, S., Sano, Y., Piao, Y., Nagaraja, R., et al. 2000. Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proc. Natl. Acad. Sci.* **97(16):** 9127–9132.

WEB SITE REFERENCES

http://www.geospiza.com/products/tools/repeatmasker.htm; RepeatMasker.

http://bio.ifom-firc.it/EST_MACHINE/index.html; EST annotation machine.

http://bio.ifom-firc.it/KW_CLUST/index.html; keyword clustering machine.

http://genome.cse.ucsc.edu/; UCSC Mouse Genome Project Working Draft, February 2002 assembly.

Received July 16, 2002; accepted in revised form September 11, 2002.