

# 1274 Full-Open Reading Frames of Transcripts Expressed in the Developing Mouse Nervous System

Maria F. Bonaldo,<sup>1</sup> Thomas B. Bair,<sup>2</sup> Todd E. Scheetz,<sup>2,3</sup> Einat Snir,<sup>1</sup> Ike Akabogu,<sup>1</sup> Jennifer L. Bair,<sup>1</sup> Brian Berger,<sup>1</sup> Keith Crouch,<sup>1</sup> Aja Davis,<sup>1</sup> Mari E. Eyestone,<sup>1</sup> Catherine Keppel,<sup>1</sup> Tamara A. Kucaba,<sup>1</sup> Mark Lebeck,<sup>1</sup> Jenny L. Lin,<sup>4</sup> Anna I.R. de Melo,<sup>1</sup> Joshua Rehmann,<sup>1</sup> Rebecca S. Reiter,<sup>4</sup> Kelly Schaefer,<sup>1</sup> Christina Smith,<sup>1</sup> Dylan Tack,<sup>5</sup> Kurtis Trout,<sup>1</sup> Val C. Sheffield,<sup>1,6</sup> Jim J-C. Lin,<sup>4</sup> Thomas L. Casavant,<sup>2,3,5,7</sup> and Marcelo B. Soares<sup>1,8,9,10,11</sup>

Departments of <sup>1</sup>Pediatrics, <sup>2</sup>Center for Bioinformatics and Computational Biology, <sup>3</sup>Ophthalmology and Visual Sciences, <sup>4</sup>Biological Sciences, <sup>5</sup>Electrical and Computer Engineering, <sup>6</sup>Howard Hughes Medical Institute, <sup>7</sup>Biomedical Engineering, <sup>8</sup>Biochemistry, <sup>9</sup>Physiology and Biophysics, <sup>10</sup>Orthopaedics, The University of Iowa, Iowa City, Iowa 52242, USA

As part of the trans-National Institutes of Health (NIH) Mouse Brain Molecular Anatomy Project (BMAP), and in close coordination with the NIH Mammalian Gene Collection Program (MGC), we initiated a large-scale project to clone, identify, and sequence the complete open reading frame (ORF) of transcripts expressed in the developing mouse nervous system. Here we report the analysis of the ORF sequence of 1274 cDNAs, obtained from 47 full-length-enriched cDNA libraries, constructed by using a novel approach, herein described. cDNA libraries were derived from size-fractionated cytoplasmic mRNA isolated from brain and eye tissues obtained at several embryonic stages and postnatal days. Altogether, including the full-ORF MGC sequences derived from these libraries by the MGC sequencing team, NIH\_BMAP full-ORF sequences correspond to ~20% of all transcripts currently represented in mouse MGC. We show that NIH\_BMAP clones comprise 68% of mouse MGC cDNAs  $\geq 5$  kb, and 54% of those  $\geq 4$  kb, as of March 15, 2004. Importantly, we identified transcripts, among the 1274 full-ORF sequences, that are exclusively or predominantly expressed in brain and eye tissues, many of which encode yet uncharacterized proteins.

[Supplemental material is available online at [www.genome.org](http://www.genome.org).]

The Brain Molecular Anatomy Project (BMAP) was initiated in 1998 by the National Institute of Mental Health (NIMH) and the National Institute of Neurological Disorders and Stroke (NINDS) as an interdisciplinary project to establish state-of-the-art technologies and informatics systems to decipher the molecular anatomy of the mammalian brain. One of the aims in the first phase of this project was the discovery of most transcripts expressed in the mouse brain, and the development of a comprehensive nonredundant arrayed collection of BMAP cDNAs and expressed sequence tags (ESTs). It was anticipated that such resources would greatly facilitate large-scale parallel analyses of gene expression studies aimed at localizing the site of expression of all BMAP transcripts in the brain. Toward this goal, we contributed ESTs that defined 28,000 of NCB1's mouse UniGene clusters, from ~80,000 ESTs generated from a comprehensive collection of BMAP cDNAs representing 12 microdissected regions of the adult C57BL/6 mouse brain, spinal cord, and retina (T. Scheetz, M. Bonaldo, B. Berger, K. Crouch, N. Wu, J. Kasperski, M. Eyestone, J. Rehmann, C. Smith, T. Kucaba, et al., in prep).

In 2001, as part of a broader trans-National Institutes of Health (NIH) effort and in close coordination with the NIH-Mammalian Gene Collection (MGC) Program (Strausberg et al. 2002), we initiated the second phase of the BMAP with the objective of identifying and determining the complete and accurate

protein coding sequence of a large number of transcripts expressed in the developing mouse nervous system. Forty-seven (NIH\_BMAP) cDNA libraries were generated from size-fractionated cytoplasmic mRNA obtained from brain and eye tissue at multiple stages of embryonic development and at postnatal days 1, 5, and 15, using a novel approach that we developed for construction of full-length-enriched cDNA libraries. The 175,990 5' ESTs, comprising 14,973 distinct clusters, were derived from these libraries based on alignment to the mouse genome. Of these, 7774 clones were tentatively identified as full-ORF-containing cDNAs, including 4223 transcripts novel to MGC. Further analysis of these 4223 clones resulted in the selection of 2084 cDNAs for full-insert sequence production, of which 1863 have been completed. Final analysis of these sequences led to the identification of 1274 NIH\_BMAP full-ORFs, all of which have been submitted to National Center for Biotechnology Information (NCBI)/MGC.

Here we report the complete and accurate sequence of 1274 NIH\_BMAP full-ORF-containing cDNAs. Thus, added to the 1019 full-ORF MGC sequences derived from NIH\_BMAP cDNA libraries by the MGC sequencing team, NIH\_BMAP full-ORF sequences correspond to ~20% of all transcripts currently represented in mouse MGC (total of 10,295 nonredundant and 12,974 redundant mouse MGC sequences as of March 15, 2004; <http://mgc.nci.nih.gov/>). Most significantly, we show that NIH\_BMAP clones constitute 68% of the mouse MGC clones  $\geq 5$  kb, and 54% of those  $\geq 4$  kb. In addition, we describe a new approach that we developed for construction of full-length-enriched cDNA libraries and successfully used for construction of 47 NIH\_BMAP cDNA

**<sup>11</sup>Corresponding author.**

**E-MAIL** [bento-soares@uiowa.edu](mailto:bento-soares@uiowa.edu); **FAX** (319) 335-9565.

Article and publication are at <http://www.genome.org/cgi/doi/10.1101/gr.2601304>.

libraries. Furthermore, we present the results of an analysis that we performed with the sequences derived from the 1274 NIH\_BMAP cDNAs, to identify those in UniGene clusters with highest relative representations of ESTs derived from brain and eye tissues, respectively. This analysis reveals a number of transcripts that are predominantly expressed in the brain, and several others with distinctive expression in the eye, many of which encode uncharacterized proteins.

## RESULTS AND DISCUSSION

### A New Approach for Construction of Full-length-Enriched cDNA Libraries

A number of methods have been developed for construction of cDNA libraries enriched for full-length cDNAs, each presenting its own advantages and disadvantages (Maruyama and Sugano 1994; Suzuki et al. 1997; Carninci and Hayashizaki 1999; Carninci et al. 2000, 2001; Piao et al. 2001; Shibata et al. 2001; Suzuki and Sugano 2001, 2003). Enrichments achieved with these methods vary over a wide range, at least in part due to confounding factors pertaining to RNA integrity, nuclear RNA contamination, impeding RNA secondary structures, and characteristics and quality of critical components of the system, such as, but not limited to, the cloning vector, and the enzymes required for cDNA synthesis and cloning.

Despite all difficulties and inherent limitations, a great number of full-length-enriched human and mouse cDNA libraries have been produced, and Carninci's "CAP trapper" (Carninci and Hayashizaki 1999; Carninci et al. 2000, 2001, 2002; Shibata et al. 2001; Hirozane-Kishikawa et al. 2003), and Sugano's "Oligo-capping" (Maruyama and Sugano 1994; Suzuki et al. 1997; Suzuki and Sugano 2001, 2003) methods have proven invaluable. As a result, significant progress has been made toward the complete sequence characterization of both the human and the mouse transcriptomes (Okazaki et al. 2002; Strausberg et al. 2002; Carninci et al. 2003; Ota et al. 2004).

An important development in this arena was the establishment of the MGC Program, a trans-NIH initiative to generate a publicly available resource of accurately sequenced full-length ORF clones for all human, mouse, and rat genes (<http://mgc.nci.nih.gov/>). It is noteworthy that MGC's objective is not to obtain strictly full-length cDNAs, that is, complete copies of the mRNAs from the 5' CAP to the 3' poly(A) addition site, but rather full-ORF-containing cDNAs. At present, an important limitation of the MGC program is that it seeks to obtain only one representative full-ORF sequence from each transcription unit, despite the fact that multiple transcripts might be derived from any given transcription unit by virtue of utilization of more than one promoter, alternative splicing, and/or differential polyadenylation.

A significant development in deciphering the transcriptome was the creation of the Mouse BMAP, a trans-NIH initiative aimed at understanding gene expression and function in the nervous system (<http://trans.nih.gov/bmap/index.htm>). Among its objectives is the identification and sequencing of most transcripts expressed in the mouse brain. As part of this effort, and in coordination with the NIH-MGC Program, we began a project aimed at cloning, identifying, and determining the sequence of a large number of full-ORF-containing cDNAs, representing transcripts expressed in the developing mouse nervous system. This project provided us the opportunity to use and rigorously test an approach that we developed for construction of full-length-enriched cDNA libraries, which attempts to overcome a problem commonly observed in full-length cDNA libraries, that is, overrepresentation of full-length cDNAs derived from smaller tran-

scripts and lack or disproportionate representation of full-length cDNAs derived from longer transcripts.

The approach we developed for construction of full-length-enriched cDNA libraries involves four principal steps: (1) size-fractionation and purification of high-quality cytoplasmic poly(A)<sup>+</sup> mRNAs; (2) synthesis of oligo-dT-primed first-strand cDNA from each mRNA size-fraction, individually, using RNaseH<sup>-</sup> reverse transcriptase under optimized conditions to yield full-length cDNAs with short 5' dT-tails; (3) size-selection and purification of double-stranded cDNAs according to the size range of the mRNAs in the size-fraction from which they originated; and (4) separate cloning and limited amplification of cDNAs in different size ranges, using a plasmid vector designed to facilitate transposon-mediated sequencing.

Ultimately, the purpose of the two most distinctive attributes of this approach—that is, (1) the serial and corresponding size fractionation of template (cytoplasmic mRNA) and product (double-stranded cDNAs), and (2) the separate cloning and (limited) amplification of cDNAs in different size ranges—is to maximize representation of transcripts, irrespective of length and abundance, in the final cDNA libraries. Because mRNA complexity is lower in a size-fractionated than in unfractionated RNA, there is greater likelihood for representation of rare transcripts in a library that contains cDNAs in the corresponding size-range than in a cDNA library derived from unfractionated mRNA. This difference is even further increased by separately cloning, electroporating, and propagating in bacteria (for limited amplification) cDNAs and clones, respectively, in different size-ranges. As a result, competition for cloning and amplification among cDNAs that differ significantly in length is eliminated, thus minimizing biases in representation of transcripts in the final library that might otherwise arise due to differences in transcript length.

It should be noted that the size-fractionation approach also presents certain disadvantages. The primary drawback is the fact that representation, in the final libraries, is limited to transcripts within the range encompassed by the mRNA size-fractions used as template for cDNA synthesis. To date, we have successfully derived full-length-enriched cDNA libraries from size-fractionated mRNA up to 7.0 kb in length. It has been our experience that libraries derived from size fractions containing transcripts in the 7.0- to 9.0-kb range are more likely to contain cDNAs derived from contaminating unprocessed nuclear transcripts, which compromises representation of bona fide full-length cDNAs.

We used this approach to construct 47 full-length-enriched cDNA libraries from size-fractionated cytoplasmic mRNA obtained from brain and eye tissue at multiple embryonic stages (upper heads at 9.5 to 10.5 dpc; brain and eyes at 12.5, 13.5, 14.5, 15.5, 16.5, 17.5, and 18.5 dpc), and from postnatal days 1, 5, and 15, of the C57BL/6 mouse strain. A complete list of the libraries constructed, with information on tissue and mRNA size fraction of origin, number of primary recombinants obtained, and total number of 5' ESTs generated, is presented in Table 1.

To assess the effectiveness of the procedures that we used for size fractionation of cytoplasmic mRNA and for size-selection of double-stranded cDNA, and thus demonstrate the quality of the full-length-enriched cDNA libraries generated with our approach, we first verified that cDNA size-ranges in these libraries do indeed correspond to those of the mRNA size fractions from which they originated. Plasmid DNA preparations from 16 libraries were linearized with a homing endonuclease (PI-SceI) and electrophoresed on an agarose gel along with a DNA size ladder. As shown in Figure 1, in all 16 libraries, cDNA sizes varied within the range of the mRNA size fraction used as template for their syntheses. It should be emphasized that shown in Figure 1 are

**Table 1.** A List of the 47 Full-length-Enriched NIH\_BMAP cDNA Libraries With Associated Information Regarding Tissue and mRNA Size Fraction of Origin, Number of Primary Recombinants Obtained, and Total Number of 5'ESTs Generated

Library name	Description	Size (kb)	Number of Recombinants	Number of 5'ESTs
NIH_BMAP_EF0	Brain (18.5 dpc)	0.5–1.0	3,000,000	1,189
NIH_BMAP_EM0	Brain (18.5 dpc)	1.0–2.0	480,000	3,502
NIH_BMAP_EG0p	Brain (18.5 dpc)	2.0–3.0	1,300,000	3,125
NIH_BMAP_EH0p	Brain (18.5 dpc)	3.0–4.0	380,000	4,020
NIH_BMAP_EQ0	Brain (18.5 dpc)	4.0–5.0	210,000	1,640
NIH_BMAP_FP0	Brain (15.5 dpc)	0.5–1.0	995,000	1,887
NIH_BMAP_ER0	Brain (15.5 dpc)	1.0–2.0	6,400,000	3,378
NIH_BMAP_EV0	Brain (15.5 dpc)	2.0–3.0	3,500,000	2,867
NIH_BMAP_EW0	Brain (15.5 dpc)	3.0–4.0	2,200,000	3,202
NIH_BMAP_EX0	Brain (15.5 dpc)	4.0–5.0	160,000	3,709
NIH_BMAP_EY0	Brain (15.5 dpc)	5.0–7.0	78,000	650
NIH_BMAP_FA0	Brain (12.5 dpc)	0.5–1.0	2,400,000	2,775
NIH_BMAP_FB0	Brain (12.5 dpc)	1.0–2.0	14,700,000	1,061
NIH_BMAP_FD0	Brain (12.5 dpc)	2.0–3.0	2,300,000	6,320
NIH_BMAP_FC0	Brain (12.5 dpc)	3.0–4.0	1,250,000	3,774
NIH_BMAP_FI0	Brain (12.5 dpc)	4.0–5.0	140,000	6,452
NIH_BMAP_FO0	Brain (12.5 dpc)	5.0–7.0	170,000	5,737
NIH_BMAP_FV0	Brain pool (13.5, 14.5, 16.5, 17.5 dpc)	0.5–1.0	1,100,000	1,981
NIH_BMAP_FX0	Brain pool (13.5, 14.5, 16.5, 17.5 dpc)	1.0–2.0	1,700,000	2,320
NIH_BMAP_FR0	Brain pool (13.5, 14.5, 16.5, 17.5 dpc)	2.0–3.0	6,600,000	4,369
NIH_BMAP_FW0	Brain pool (13.5, 14.5, 16.5, 17.5 dpc)	3.0–4.0	400,000	5,632
NIH_BMAP_FY0	Brain pool (13.5, 14.5, 16.5, 17.5 dpc)	4.0–5.0	173,000	18,192
NIH_BMAP_GI0	Brain pool (13.5, 14.5, 16.5, 17.5 dpc)	5.0–7.0	179,000	6,756
NIH_BMAP_GK0	Brain, Newborn pool (1, 5, 15 days)	0.5–1.0	23,000,000	1,868
NIH_BMAP_GL0	Brain, Newborn pool (1, 5, 15 days)	1.0–2.0	36,000,000	1,907
NIH_BMAP_GM0	Brain, Newborn pool (1, 5, 15 days)	2.0–3.0	723,000	1,735
NIH_BMAP_GH0	Brain, Newborn pool (1, 5, 15 days)	4.0–5.0	125,000	8,297
NIH_BMAP_GV0	Brain, Newborn pool (1, 5, 15 days)	5.0–7.0	10,000	6,844
NIH_BMAP_HC0	Eye pool (12.5, 13.5, 14.5 dpc)	0.5–1.0	1,455,000	2,489
NIH_BMAP_HA0	Eye pool (12.5, 13.5, 14.5 dpc)	1.0–2.0	1,055,000	1,745
NIH_BMAP_GZ0	Eye pool (12.5, 13.5, 14.5 dpc)	2.0–3.0	2,590,000	4,475
NIH_BMAP_HD0	Eye pool (12.5, 13.5, 14.5 dpc)	3.0–4.0	740,000	4,189
NIH_BMAP_HB0	Eye pool (12.5, 13.5, 14.5 dpc)	4.0–5.0	260,000	8,917
NIH_BMAP_HE0	Eye pool (12.5, 13.5, 14.5 dpc)	5.0–7.0	16,000	2,557
NIH_BMAP_FZ0	Eye pool (15.5, 16.5, 17.5, 18.5 dpc)	0.5–1.0	1,000,000	1,541
NIH_BMAP_HP0	Eye pool (15.5, 16.5, 17.5, 18.5 dpc)	1.0–2.0	1,030,000	1,035
NIH_BMAP_GW0	Eye pool (15.5, 16.5, 17.5, 18.5 dpc)	2.0–3.0	320,000	4,028
NIH_BMAP_HK0	Upper Head (9.5–10.5 dpc)	0.5–1.0	697,000	1,949
NIH_BMAP_HL0	Upper Head (9.5–10.5 dpc)	1.0–2.0	554,000	2,039
NIH_BMAP_HJ0	Upper Head (9.5–10.5 dpc)	2.0–3.0	1,118,000	4,358
NIH_BMAP_HQ0	Upper Head (9.5–10.5 dpc)	3.0–4.0	220,000	4,624
NIH_BMAP_HN0	Upper Head (9.5–10.5 dpc)	4.0–5.0	120,000	9,654
NIH_BMAP_HO0	Upper Head (9.5–10.5 dpc)	5.0–7.0	49,000	5,484
NIH_BMAP_HS0	Upper Head (9.5–10.5 dpc)	5.0–7.0	88,000	1,717
NIH_BMAP_HV0	Eye, Newborn pool (1, 5, 15 days)	0.5–1.0	635,000	0
NIH_BMAP_HW0	Eye, Newborn pool (1, 5, 15 days)	1.0–2.0	581,000	0
NIH_BMAP_HU0	Eye, Newborn pool (1, 5, 15 days)	2.0–3.0	568,000	0

linearized plasmid DNA preparations of entire libraries, which include the 1691-bp pYX-Asc I cloning vector.

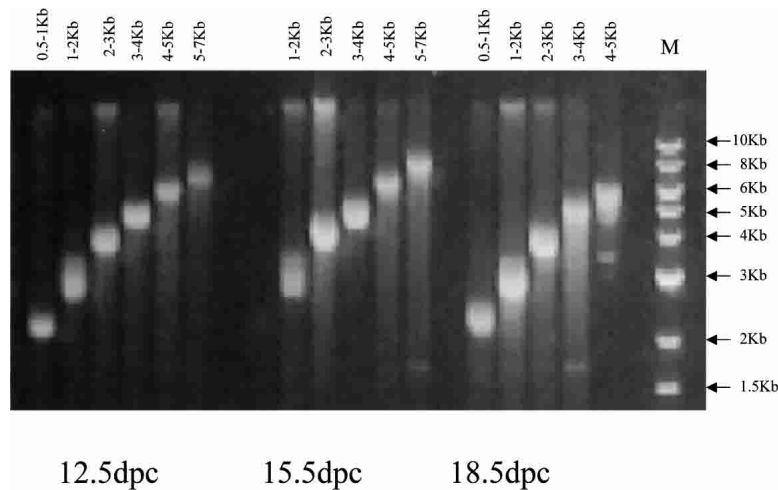
Next, we investigated the correspondence between actual and expected lengths of 1274 cDNAs, by comparing the lengths of their complete sequences with the size ranges expected for the cDNAs in the library they originated. As shown in Table 2, this analysis revealed an overall correspondence that ranged from 50% to 100%. It is noteworthy, however, that this analysis also revealed that 70% to 90% of the sequences with lengths falling below the size range expected for clones in the library were shorter by no >15% of the minimum length in the respective range, thus documenting the effectiveness of this approach to generate libraries enriched for full-length cDNAs. A complete list of all sequences and corresponding libraries is available as Supplemental data.

Here we report the sequence characterization of 1274 full-ORF-containing NIH\_BMAP cDNAs that we identified in these

libraries, all of which have been submitted to NCBI/MGC. It should be emphasized that all NIH\_BMAP libraries were also contributed to the MGC program, while being characterized and sequenced in our laboratory. As a result, an additional 1019 full-ORF sequences were derived from these libraries by the MGC sequencing team.

### Selection of Full-ORF-Containing cDNAs for Full-Insert Sequencing

The first step in the clone selection pipeline involved arraying of the cDNA libraries and production of 5' ESTs. The number of 5' ESTs derived from each library is listed in Table 1. In total, 175,990 5' ESTs were generated and subjected to informatics analyses (sequence homology-based and ab-initio methods) for selection of full-ORF clone candidates. Alignment of the 175,990 5' ESTs to the mouse genomic sequence using a dedicated BLAT



**Figure 1** Distribution of clone sizes in the size-fractionated cDNA libraries to assess the efficacy of the method used for size-fractionation of cytoplasmic mRNA and size selection of double-stranded cDNA. Linearized plasmid DNA from each of 16 size-fractionated libraries representing size fractions from 0.5 to 1 kb to 5 to 7 kb were electrophoresed and compared with a 1-kb standard ladder (New England BioLabs). Note that the clone sizes include the 1.7-kb pYX-Asc I vector.

server (Kent 2002; Karolchik et al. 2003) enabled the identification of 14,973 EST clusters. The 5' most EST from each cluster was then identified and subjected to further analyses. These included (1) BLAST searches against RefSeq, Riken, SWISS-PROT, and MGC databases; (2) genomic context examination, based on data available in the UCSC database, to determine whether a 5' EST overlaps with the start codon of a known gene or mRNA and whether it is the most 5' of all ESTs mapping to that genomic location; and (3) *ab initio* tools for classification of 5' ESTs according to the likelihood that they represent full-ORF-containing cDNAs. The latter are based on recognition of distinctive sequence features, such as the presence of a Kozak sequence motif and the occurrence and localization of start and stop codons, and decision tree optimization based on historic true/false-positive rates (<http://genome.uiowa.edu/techreports.html>). Selected clones were BLASTed against NCBI's "nr" database and GenBank records of each significant hit ( $e$  value  $< 10^{-8}$ ) were examined to seek homologous genes that had an annotated CDS for additional evidence that might further support or contradict the prediction.

A total of 7774 cDNA clones were selected according to these criteria as putative full-ORF-containing cDNAs. Of those, 3551 corresponded to transcription units already represented in the mouse MGC database and hence were not selected for full-insert sequencing, but still remain of interest due to their potential of representing alternatively spliced variants. Of the remaining 4223 clones, 3579 were rearranged and 3' ESTs were generated,

and 644 await processing. Upon visual inspection and analysis by an annotator in the finishing group, 2084 clones were selected for *in vitro* transposition and full-insert sequencing, and the remaining 1495 clones were rejected.

Complete and accurate full-insert sequence has been obtained for 1863 clones, and 221 clones are currently in the finishing pipeline. Final analyses resulted in the classification of 1274 NIH\_BMAP clones as full-ORF-containing cDNAs and in the rejection of 589 finished sequences for one of the following reasons: chimeric (5), frame shift (142), retained intron (113), library artifacts (22), unable to sequence (131), no significant ORF (39), 5' truncation of ORF (54), 3' truncation of ORF (30), a similar clone appeared in MGC while ours was still in the finishing phase (44), and 5' end sequence of the re-arrayed clone did not match that of the original 5' EST (9).

In conclusion, 52% (7774/14,973) of the cDNAs in the nonredundant set of clones that we identified in the NIH\_BMAP

cDNA libraries were ranked as full-ORF-containing candidates based on analyses performed on their 5' ESTs. Of those, 72% to 81% (5635 to 6279/7774; 644 pending and 1495 rejected) remained considered as full-ORF-containing candidates upon further inspection and analysis of their 3' ESTs. Finally, 68% (1274/1863) of the clones selected for full-insert sequencing, ultimately met all criteria required for classification as full-ORF-containing cDNAs.

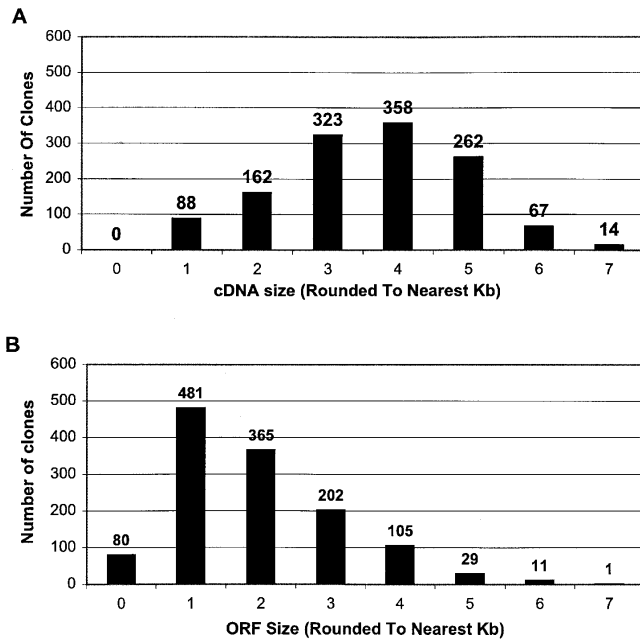
### Distribution of cDNA Size and ORF Length in NIH\_BMAP Clones: Comparison to the Remainder Mouse MGC cDNAs

The distribution of cDNA sizes and ORF lengths in the 1274 NIH\_BMAP cDNA sequences reported in this manuscript are presented in Figure 2, A and B, respectively. In addition, a complete list of the 1274 clones with their respective sequence and ORF lengths is available as Supplemental data to this manuscript. As shown in Figure 2A, although cDNA sizes range from 507 to 7146 bp, most sequences are in the 3.0- to 5.0-kb range (74%; 943/1274), with a peak of ~4 kb (28%; 358/1274). On the other hand, as shown in Figure 2B, although ORF lengths range from 204 to 6849 bp, the majority fall within 1.0 to 3.0 kb (82%; 1048/1274), with a peak at ~1 kb (38%; 481/1274).

A comparison of the size distribution of the NIH\_BMAP cDNAs that are in the mouse MGC database with that of all other mouse MGC cDNAs revealed a striking difference, in that ~90%

**Table 2.** Correspondence Between Actual and Expected Lengths of 1274 cDNAs

cDNA sequence length	Percentage of sequence lengths that fall within the size range expected for the library	Percent of sequence lengths that fall below the size range expected for the library	Percent of sequence lengths shorter by no more than 15% of the expected size range for the library
0.5–1.0 kb	100% (22/22)	0	0
1.0–2.0 kb	79% (53/67)	21% (14/67)	80% (11/14; $\geq 0.85$ kb)
2.0–3.0 kb	76% (116/152)	24% (36/152)	83% (30/36; $\geq 1.7$ kb)
3.0–4.0 kb	50% (139/279)	50% (140/279)	90% (126/140; $\geq 2.5$ kb)
4.0–5.0 kb	70% (363/521)	30% (158/521)	68% (108/158; $\geq 3.4$ kb)
5.0–7.0 kb	61% (142/233)	39% (91/233)	73% (66/91; $\geq 4.25$ kb)



**Figure 2** Size distribution of the 1274 full-ORF clones. For the 1274 full-ORF clones derived from NIH-BMAP libraries and sequenced locally: the distribution of insert sizes (A) and the distribution of ORF sizes (B).

of the cDNAs in the latter group fall in the 1.0- to 3.0-kb range, with a peak ~2 kb (Fig. 3A). This is in sharp contrast with the size distribution of the NIH\_BMAP cDNAs, with ~75% of the sequences falling in the 3.0- to 5.0-kb range. This difference is particularly notable in the subset of MGC sequences derived from longer cDNAs (Fig. 3B), with NIH\_BMAP clones composing 68% (493/727) of all mouse MGC clones  $\geq 5$  kb and 54% of those  $\geq 4$  kb (1060/1961). It is noteworthy that all NIH\_BMAP cDNA sequences, comprising 20% of the MGC database, were included in this analysis irrespective of whether they were generated by our group at the University of Iowa or by laboratory members of the MGC sequencing team.

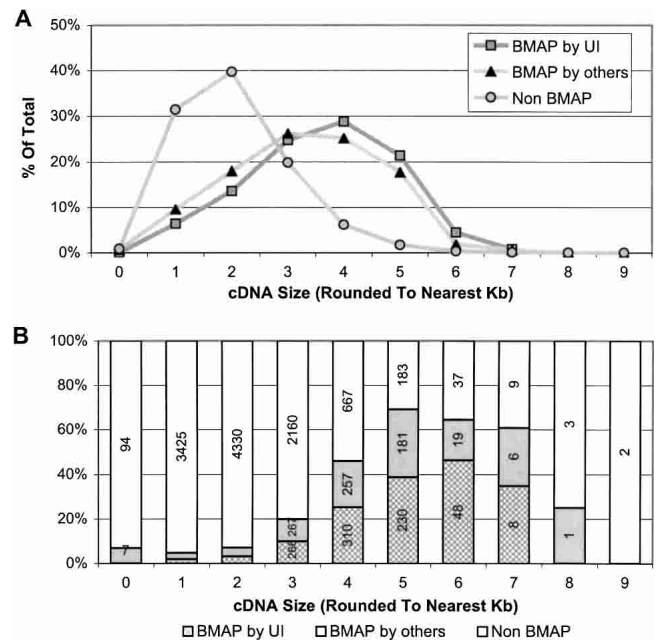
### NIH\_BMAP cDNAs Predominantly Expressed in Brain and Eye Tissues

We have analyzed sequences derived from the 1274 full-ORF-containing NIH\_BMAP cDNAs reported in this manuscript to identify those corresponding to transcripts most distinctively expressed in brain and eye tissues, respectively. We used NCBI's mouse UniGene database (build no. 135) to identify UniGene clusters containing sequences derived from the 1274 NIH\_BMAP cDNAs. We then obtained information on the tissue of origin of every EST constituent of the 1274 clusters, using a locally curated translation file that specifies the tissue source(s) for each library identification in the mouse UniGene, and determined the relative representations of ESTs derived from brain and eye cDNA libraries, respectively, in each cluster (shown as percentages in Table 3A,B). Next, we selected the NIH\_BMAP transcripts in clusters with highest relative representations of brain and eye ESTs. Last, we determined, within each group, the subset of ESTs that originated from embryonic brain and embryonic eyes, respectively. This was calculated for each cluster, based on the ratio of the number of embryonic brain- (or embryonic eye-) ESTs per total number of brain (adult + embryonic) or eye (adult + embryonic) ESTs in the cluster (also shown as percentages in Table 3A,B).

It should be emphasized that only ESTs obtained from libraries derived exclusively from brain or eye tissues were counted as evidence for expression in the brain or in the eyes, respectively. Thus, this analysis provides a conservative estimate of transcript expression in these tissues.

A total of 3,089,497 ESTs in the mouse UniGene were used for these analyses, comprising 531,536 brain ESTs (17.2%), of which 216,120 are embryonic (7.0%), and 132,625 eye ESTs (4.3%), of which 27,721 are embryonic (0.9%).

Fifty-two NIH\_BMAP transcripts expressed exclusively or predominantly in the brain, arbitrarily defined as those in UniGene clusters with >75% of ESTs derived from brain tissue, were identified in this analysis (Table 3A; Fig. 4). Included in this set are several transcripts known to be distinctively expressed in the brain, such as calneuron 1 (Wu et al. 2001); seizure-related gene 6 (Shimizu-Nishikawa et al. 1995); kinesin family member 1A (Okada et al. 1995); cerebellin 3 (Pang et al. 2000); T-box brain gene 1 (Bulfone et al. 1995); adenylate cyclase activating polypeptide 1 receptor 1 (Sheward et al. 1998); leucine-rich repeat LGI family, member 1 (Gu et al. 2002; Kalachikov et al. 2002); sulfotransferase family 4A, member 1 (Sakachikawa et al. 2002); protocadherin 8 (Strehl et al. 1998); adenosine deaminase, RNA-specific, B2 (Mittaz et al. 1997); potassium inwardly-rectifying channel, subfamily J, member 9 (Lesage et al. 1994); and glutamate receptor, ionotropic AMPA1 GluR1 (Puckett et al. 1991). Interestingly, however, this analysis revealed several transcripts that are still uncharacterized and that seem to be mainly, but not highly, expressed in the brain (e.g., Mm.296323, Mm.337426, Mm.334249, Mm.334408). Yet a third class included uncharacterized transcripts apparently differentially and highly expressed



**Figure 3** BMAP clones/sequences in the mouse MGC collection as of March 17, 2004. Three categories of clones are shown: 1077 BMAP clones sequenced by UI, 1019 BMAP clones sequenced by others, and non-BMAP clones (which represent all the clones in the MGC collection that are not BMAP clones). (A) Size distribution as a percentage of total clones within each category. (B) Percentage of clones contributed by each category, by clone size. The number of BMAP clones in MGC sequenced by UI from clone sizes 1 and 2 are 69 and 146, respectively. The number of BMAP clones in MGC sequenced by others for clone sizes 1 and 2 are 98 and 183, respectively.

**Table 3A.** Top 52 NIH\_BMAP Transcripts Most Distinctively Expressed in the Brain

UniGene cluster	No. of brain ESTs in cluster (adult and embryonic)	Percentage of brain ESTs in cluster (brain specificity)	No. of embryonic brain ESTs in cluster	Percentage of embryonic brain ESTs (embryonic specificity)	UniGene description
Mm.296323	5	100	2	40	Similar to zinc finger protein 334
Mm.337426	5	100	2	40	Transcribed sequence with weak similarity to protein pir:l68600 ( <i>H. sapiens</i> ) l68600 dipeptidyl aminopeptidase like protein—human
Mm.334249	4	100	4	100	Transcribed sequence with weak similarity to protein ref:NP_079090.1 ( <i>H. sapiens</i> ) hypothetical protein FLJ23109; likely ortholog of mouse E-cadherin binding protein E7 [ <i>H. sapiens</i> ]
Mm.334408	2	100	2	100	Transcribed sequences
Mm.353170	2	100	1	50	Transcribed sequence with strong similarity to protein ref:NP_116220.1 ( <i>H. sapiens</i> ) CAP-binding protein complex interacting protein 2 [ <i>H. sapiens</i> ]
Mm.223128	1	100	0	0	olfactory receptor MOR202-8
Mm.249245	1	100	0	0	<i>Mus musculus</i> 13d embryo head cDNA, RIKEN full-length enriched library, clone:3110006O06 product:unclassifiable, full insert sequence
Mm.97163	42	95.45	12	28.57	cerebellin 3 precursor protein
Mm.44245	90	90.91	53	58.89	adenylate cyclase activating polypeptide 1 receptor 1
Mm.248796	287	90.25	28	9.76	sulfotransferase family 4A, member 1
Mm.326240	37	90.24	5	13.51	copine IV
Mm.320206	51	89.47	6	11.76	calneuron 1
Mm.270641	169	88.95	35	20.71	hippocalcin-like 4
Mm.4920	139	87.97	42	30.22	glutamate receptor, ionotropic, AMPA1 ( $\alpha$ 1)
Mm.242088	36	87.8	12	33.33	adenosine deaminase, RNA-specific, B2
Mm.279818	107	87.7	11	10.28	RIKEN cDNA 3110035E14 gene
Mm.44413	92	87.62	16	17.39	RIKEN cDNA 1500003N10 gene
Mm.276408	183	87.14	113	61.75	kinesin family member 1A
Mm.268079	106	86.89	60	56.6	chondroitin sulfate proteoglycan 3
Mm.298251	25	86.21	6	24	leucine-rich repeat LGI family, member 1
Mm.35474	177	85.92	33	18.64	stathmin-like 4
Mm.333266	6	85.71	4	66.67	Transcribed sequence with weak similarity to protein sp:O60721 ( <i>H. sapiens</i> ) NKX1_HUMAN Sodium/potassium/calcium exchanger 1 precursor (Na <sup>+</sup> /K <sup>+</sup> /Ca <sup>2+</sup> -exchange protein 1) (Retinal rod Na-Ca <sup>2+</sup> K exchanger)
Mm.261168	101	85.59	11	10.89	potassium inwardly-rectifying channel, subfamily J, member 9
Mm.42823	65	84.42	36	55.38	RIKEN cDNA 6530413N01 gene
Mm.103811	43	84.31	19	44.19	protocadherin 8
Mm.39752	137	84.05	19	13.87	RIKEN cDNA 2900041A09 gene
Mm.40741	42	84	17	40.48	RIKEN cDNA 9630019K15 gene
Mm.229330	141	83.93	100	70.92	seizure-related gene 6
Mm.121274	78	83.87	12	15.38	RIKEN cDNA 9630048M01 gene
Mm.44473	10	83.33	8	80	Transcribed sequences
Mm.196944	78	82.98	38	48.72	RIKEN cDNA 6330404E20 gene
Mm.240965	71	82.56	24	33.8	expressed sequence N28178
Mm.246605	102	82.26	21	20.59	RIKEN cDNA 4930471K13 gene
Mm.308525	55	82.09	22	40	T-box brain gene 1
Mm.252890	18	81.82	7	38.89	hypothetical protein 5330438O12
Mm.27005	120	81.63	22	18.33	visinin-like 1
Mm.41708	80	81.63	20	25	leucine-rich repeat LGI family, member 3
Mm.103502	51	80.95	17	33.33	RIKEN cDNA 5730507A09 gene

(continued)

**Table 3A.** *Continued*

UniGene cluster	No. of brain ESTs in cluster (adult and embryonic)	Percentage of brain ESTs in cluster (brain specificity)	No. of embryonic brain ESTs in cluster	Percentage of embryonic brain ESTs (embryonic specificity)	UniGene description
Mm.35088	51	79.69	35	68.63	cholinergic receptor, nicotinic beta polypeptide 2 (neuronal)
Mm.349133	11	78.57	7	63.64	Transcribed sequence with weak similarity to protein sp:Q9Y2J0 ( <i>H. sapiens</i> ) RP3A_HUMAN Rabphilin-3A
Mm.334237	46	77.97	5	10.87	<i>Mus musculus</i> cDNA clone IMAGE:6593004, with apparent retained intron
Mm.332656	7	77.78	3	42.86	Transcribed sequence with moderate similarity to protein pir:T00390 ( <i>H. sapiens</i> ) T00390 KIAA0614 protein—human (fragment)
Mm.113877	30	76.92	3	10	glycine receptor, $\alpha 2$ subunit
Mm.256342	208	76.75	115	55.29	kinesin family member 5C
Mm.289702	273	76.69	60	21.98	synaptotagmin 1
Mm.32191	209	76.56	68	32.54	gamma-aminobutyric acid (GABA-B) receptor, 1
Mm.241147	32	76.19	6	18.75	solute carrier family 8 (sodium/calcium exchanger), member 2
Mm.19047	60	75	36	60	RIKEN cDNA 9330157P13 gene
Mm.331076	15	75	4	26.67	Similar to neuronal transmembrane protein Slitrk3
Mm.149738	6	75	1	16.67	Similar to hypothetical protein
Mm.323867	3	75	0	0	Similar to CAP-binding protein complex interacting protein 1
Mm.329662	3	75	2	66.67	Similar to beta-galactose-3-O-sulfo-transferase 3; Gal3ST-3

in the brain (e.g., Mm.279818, Mm.44413, Mm.246605). In addition, we identified transcripts that seem to be expressed at relatively low levels, but specifically, in embryonic brain (e.g., Mm.334249, Mm.334408).

In contrast, only four of the 1274 NIH\_BMAP transcripts were found in UniGene clusters containing  $\geq 50\%$  ESTs derived from eye tissue (Table 3B), of which three appear to be moderately or highly expressed in both embryonic and adult eyes (crystallin  $\beta$  A4, crystallin  $\beta$  A2, and dopachrome tautomerase). The fourth (Mm.246812), however, is a rare and yet uncharacterized embryonic eye transcript (BC067074) formerly, and only once, identified in a kidney cDNA library from a 14-month-old male mouse (BC042787). BLAT analysis indicates that the eye and kidney transcripts use different alternatively spliced 3' terminal exons. This analysis also revealed a number of NIH\_BMAP transcripts that are highly, yet not predominantly, expressed in the eye, some of which are known (e.g., Mm.1008, Mm.1860) and others are yet uncharacterized (e.g., Mm.55143). Several transcripts were identified in the 25% to 50% range of eye-specific expression, among which are a number of yet-uncharacterized transcripts expressed over a wide range in the eye.

In conclusion, based on the preliminary characterization of the 1274 NIH\_BMAP full-ORF-containing cDNA sequences reported in this manuscript, we anticipate that these NIH\_BMAP full-length-enriched cDNA libraries will prove invaluable not only for identification of additional full-ORF sequences derived from transcripts expressed in the developing mouse nervous system but also as a resource for identification of brain-specific transcripts, resulting from brain-specific splicing and/or polyadenylation, utilization of brain-specific promoters, as well as brain-specific antisense transcripts.

## METHODS

### Tissue Dissection

Time-pregnant C57BL/6 mice were purchased from either Charles River (Wilmington, MA) or Harlan (Indianapolis, IN). For embryonic days 9.5 to 10.5, the head was cut just anterior to the developing mandible and through the middle of the hindbrain. The developing eye was included in this cut. For embryonic days 12.5, 13.5, 14.5, 15.5, 16.5, 17.5, and 18.5 and postnatal days 1, 5, and 15, eyes and brains were dissected separately. The freshly dissected tissues were collected in DEPC-treated phosphate buffered saline and used immediately for the isolation of cytoplasmic RNA.

### Isolation of Cytoplasmic mRNA

Cytoplasmic RNA was isolated essentially as described before (Favaloro et al. 1980). The tissue was homogenized in lysis buffer (140 mM NaCl, 1.5 mM MgCl<sub>2</sub>, 10 mM Tris-HCl pH 8.6, 0.5 NP-40, and 10 mM vanadyl-ribonucleoside complexes) by using a tissue grinder (Kontes) with a loose pestle. Five milliliters of lysed tissue was then transferred to a 13 mL Sarstedt tube with 5 mL sucrose-containing lysis buffer (0.7 M sucrose) with 1% NP-40 and centrifuged in a Sorvall HB-6 rotor at 14,000g for 20 min at 4°C. The upper layer was carefully transferred to a Sarstedt tube containing one volume of 2 $\times$  proteinase K buffer (20 mM Tris-HCl at pH 7.5, 10 mM EDTA, 1% SDS). Proteinase K was added to a final concentration of 200  $\mu$ g/mL and incubated at 37°C for 30 min. The RNA was extracted with one volume of phenol-chloroform, centrifuged in a Sorvall SS34 rotor at 12,000g for 20 min at 4°C and then precipitated with 2.5 volumes of ethanol and 0.1 volume of sodium acetate (pH 5.2). RNA samples were digested with 10 U RNase-free DNase (Roche) in 1 mM EDTA, 50 mM MgCl<sub>2</sub>, and 250 mM Tris HCl (pH 7.5) buffer for 30 min at

**Table 3B.** Top 50 NIH\_BMAP Transcripts Most Distinctively Expressed in the Eye

UniGene cluster	No. of eye ESTs in cluster (adult and embryonic)	Percentage of eye ESTs in cluster (eye specificity)	No. of embryonic eye ESTs in cluster	Percentage of embryonic eye ESTs (embryonic eye specificity)	Description
Mm.40324	84	86.6	59	70.24	Crystalline, $\beta$ A4
Mm.86656	48	85.71	25	52.08	Crystalline, $\beta$ A2
Mm.19987	54	56.25	22	40.74	dopachrome tautomerase
Mm.246812	1	50	1	100	Mus musculus, clone IMAGE:4221513, mRNA
Mm.214385	31	45.59	9	29.03	mab-21-like 2 ( <i>C. elegans</i> )
Mm.278887	3	42.86	0	0	Mus musculus cDNA clone IMAGE:6406267, partial cds
Mm.287100	21	42	4	19.05	nuclear receptor subfamily 2, group E, member 1
Mm.348020	8	40	0	0	RIKEN cDNA 5830461H18 gene
Mm.221027	2	40	0	0	one cut domain, family member 3
Mm.166744	1	33.33	0	0	RIKEN cDNA 2900056M07 gene
MM.55143	98	31.21	33	33.67	Dickkopf homolog 3 ( <i>Xenopus laevis</i> )
Mm.303897	4	30.77	0	0	Similar to nucleolar protein 4; nucleolar localized protein
Mm.103747	3	30	0	0	neuropeptide FF-amide peptide precursor
Mm.214758	8	29.63	1	12.5	solute carrier family 8 (sodium/calcium exchanger), member 3
Mm.291498	7	26.92	4	57.14	Similar to keratin 6 irs3
Mm.329322	17	26.56	1	5.88	formin-family protein FHOS2
Mm.310822	6	25	1	16.67	A kinase (PKA) anchor protein 6
Mm.275387	21	24.71	0	0	RIKEN cDNA 1810041L15 gene
Mm.154658	15	24.59	2	13.33	cDNA sequence BC043114
Mm.273082	20	24.39	0	0	synaptic vesicle glycoprotein 2b
Mm.99790	13	24.07	0	0	hypothetical protein LOC231503
Mm.1860	62	23.22	7	11.29	phosphatidylinositol membrane-associated
Mm.272870	25	22.73	3	12	RIKEN cDNA 4933432H23
Mm.1008	80	22.66	0	0	prostaglandin D2 synthase
Mm.44736	30	22.39	0	0	dynamin
Mm.252987	29	21.97	1	3.45	solute carrier family 12, member 5
Mm.194148	10	21.28	1	10	rhomboid, veinlet-like 4 ( <i>Drosophila</i> )
Mm.275499	7	21.21	0	0	RIKEN cDNA 2600011E07 gene
Mm.27503	22	20.95	5	22.73	RIKEN cDNA 1810009A15 gene
Mm.101022	10	20.83	1	10	RIKEN cDNA A730032D07 gene
Mm.151931	5	20.83	0	0	RIKEN cDNA 6720466O15 gene
Mm.250981	27	20.61	8	29.63	Eph receptor B2
Mm.268383	29	20.14	8	27.59	ribosomal protein S6 kinase, polypeptide 2
Mm.28424	14	19.18	4	28.57	SRY-box containing gene 12
Mm.22398	17	19.1	2	11.76	jagged 1
Mm.297419	4	19.05	0	0	Similar to KIAA 1987 protein
Mm.41108	20	18.69	5	25	cDNA sequence AF322649
Mm.142655	27	18.62	9	33.33	jumonji, AT rich interactive domain 1C (Rbp2-like)
Mm.272210	5	18.52	1	20	ciliary neurotrophic factor receptor
Mm.207715	18	18.37	4	22.22	RIKEN cDNA 5730410E15 gene
Mm.268532	15	18.07	0	0	Unknown (protein for IMAGE:5694541)
Mm.270469	38	18.01	6	15.79	DNA segment, Chr 5, Brigham & Women's Genetics 0860 expressed
Mm.99905	6	17.65	0	0	<i>Mus musculus</i> adult male corpus striatum cDNA, RIKEN full-length enriched library, clone:C030002C11 product: unknown EST, full insert sequence
Mm.121514	12	17.39	2	16.67	RIKEN cDNA 1500032A09 gene
Mm.103714	4	17.39	2	50	serine threonine kinase 32
Mm.146779	11	16.92	0	0	RIKEN cDNA 1500001A10
Mm.297250	14	16.87	1	7.14	Similar to hypothetical protein FLJ14547

(continued)

**Table 3B.** *Continued*

UniGene cluster	No. of eye ESTs in cluster (adult and embryonic)	Percentage of eye ESTs in cluster (eye specificity)	No. of embryonic eye ESTs in cluster	Percentage of embryonic eye ESTs (embryonic eye specificity)	Description
Mm.234832	10	16.67	0	0	Transcribed sequence with strong similarity to protein ref:NP_055148.1 ( <i>H. sapiens</i> ) immunoglobulin superfamily, member 4 [ <i>H. sapiens</i> ]
Mm.126793	3	16.67	0	0	Transcribed sequence with weak similarity to protein pir:A41234 ( <i>H. sapiens</i> ) A41234 melanocyte-specific protein Pmel-17 precursor—human ankyrin 2, brain
Mm.220242	30	16.48	6	20	

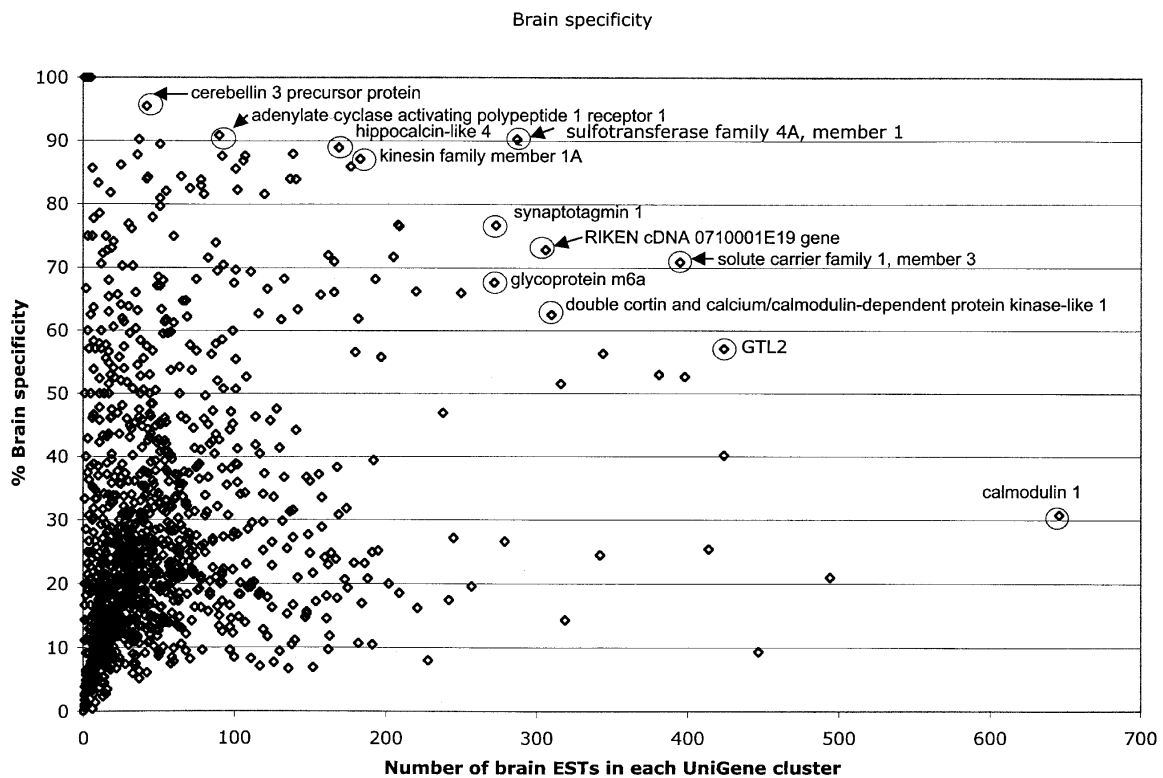
37°C, and poly(A)<sup>+</sup> mRNA was purified using Oligotex mRNA kit (Qiagen) or Dynalbeads mRNA direct kit (DynaL).

### Size Fractionation of poly(A)<sup>+</sup> mRNA

Poly(A)<sup>+</sup> mRNA (~5 µg) was ethanol precipitated, resuspended in deionized formamide, and loaded on a 1% low melting temperature agarose gel, next to the 1-kb RNA ladder used as reference for size fractionation of the mRNA. The RNA ladder, but not the poly(A)<sup>+</sup> mRNA, was stained with ethidium bromide and exposed to UV light to guide the otherwise blind size fractionation of the poly(A)<sup>+</sup> mRNA. Gel slices containing poly(A)<sup>+</sup> mRNA fractions of 0.5 to 1.0 kb, 1.0 to 2.0 kb, 2.0 to 3.0 kb, 3.0 to 4.0 kb, 4.0 to 5.0 kb, and 5.0 to 7.0 kb were melted at 65°C and digested with 0.3 U of Agarase (Promega) at 40°C.

### Construction of cDNA Libraries From Size-Fractionated Cytoplasmic mRNA

cDNA libraries were constructed from each individual mRNA size fraction, essentially as we previously described (Bonaldo et al. 1996). Typically, each cDNA library was derived from 0.2 µg size-fractionated poly(A)<sup>+</sup> mRNA. Briefly, first-strand cDNA synthesis was primed with a dT<sub>18</sub> oligonucleotide containing a NotI site, for directional cloning, and a library-specific sequence-tag of 10 nucleotides positioned between the NotI site and the dT<sub>18</sub> (Gavin et al. 2002), under conditions that we have optimized for generation of cDNAs with short dA/dT tails (Soares and Bonaldo 1998), in a reaction with 400 U reverse transcriptase (Superscript II, Life Technologies) and 0.5 mM each dATP, dTTP, dGTP, and methyl-dCTP, for 2 h at 37°C. Double-strand cDNA was synthe-



**Figure 4** Graph of brain-specific versus EST composition. Gene identities were annotated for several of the most prevalent and most brain-specific clusters.

sized by nick translation in a reaction with *Escherichia coli* DNA polymerase (New England Biolabs), *E. coli* DNA Ligase (New England Biolabs), and RNase H (USB), and size fractionated by agarose gel electrophoresis according to the size range of the mRNA size fraction used as template for first-strand cDNA synthesis. Size-selected cDNAs were ligated to EcoRI adaptors, digested with NotI, phosphorylated, and directionally cloned into the pYX-Asc I vector, doubly digested with EcoRI and NotI. The cDNA libraries were electroporated into phage (T1)-resistant-DH10B *E. coli*, and after 1 h at 37°C, 30% of the library was plated onto agar plates containing ampicillin, from which individual colonies were robotically picked and arrayed into 384-well plates. The remaining 70% was grown at 37°C overnight, and plasmid DNA was prepared by using a Qiagen kit.

We used the following library tags: (1) brain libraries: CAGCCACGAC (E18.5 dpc), GTGCGTGGA (E15.5 dpc), TGAGAGAGCC (E12.5 dpc), AGCGAGACAG (pool of E13.5, 14.5, 16.5, and 17.5 dpc), CGAACTGAAT (E9.5 to 10.5 dpc), and CGAACTGAAT (pool of postnatal days 1, 5, and 15); (2) eye libraries: TTATTGAAGT (pool of E12.5, 13.5, and 14.5 dpc), CTGCGCTCCTC (pool of E15.5, 16.5, 17.5, and 18.5 dpc), and AATAATTACG (pool of postnatal days 1, 5, and 15).

pYX-Asc I is a 1691-bp plasmid that we derived from the pYX vector originally constructed and kindly provided by Dr. M.J. Brownstein (NIH). The modifications that we introduced in the pYX plasmid include the addition of an AscI site to the polylinker and the deletion of a region containing nonessential sequence. The latter modification was introduced after our observation of multiple transposon integrations within this region in the *in vitro* transposition reactions performed for transposon-facilitated sequencing. The resulting vector (pYX-Asc I) is thus ideal for transposon-facilitated sequencing because, with the exception of the short polylinker sequence, transposon integration into any sequence in the vector renders it a nonviable clone. Additional information on the pYX-Asc I vector, including its complete sequence, can be obtained at <http://image.llnl.gov/image/html/vectors.shtml>.

### Transposon-Facilitated Sequencing and Finishing

The cDNA clones that were identified as full-ORF-containing candidates not yet represented in MGC were rearranged and sequence-verified (both 5' and 3' end-sequences were obtained). Selected clones were colony-purified and grown individually for 15 h at 37°C, and their cultures were combined into pools. An aliquot of each culture was saved as a glycerol stock. Each pool consisted of seven to 16 clones with a combined size of 40 to 50 kb. Plasmid DNA from each pool was purified by using a Wizard Plus SV miniprep DNA purification system (Promega), and a sample of the purified plasmid was loaded to an agarose gel as a quality-control measure and to determine DNA quantity. Approximately 150 ng of purified DNA was used in a transposition reaction performed with a Template Generation System (Finnzymes) according to the manufacturer's instructions. One quarter of the transposition reaction product was electroporated into DH10B cells, incubated for 1 h at 37°C and plated on agar plates under appropriate antibiotic selection. A Genetix QBot was used to array 384 bacterial colonies obtained from each pool into 384-well microtiter plates. The 384-well plates were incubated overnight at 37°C, and double-stranded plasmid DNA templates were prepared from the resulting glycerol stocks by using a microwave-mediated cell lysis method (Marra et al. 1999).

Two parallel sequencing reactions were performed on each DNA template by using the ABI PRISM dRhodamine terminator cycle sequencing kit and primers that correspond to priming sites on the transposon element. Reaction products were electrophoresed on an ABI PRISM 3700 DNA Analyzer. Once the sequence reads were generated, the phred/phrap/consed package (Ewing et al. 1998; Gordon et al. 1998) was used to assign quality scores, to assemble the sequence reads, and to view and to edit the assemblies. For each pool, the 3' and 5' ESTs of the pooled clones were included in the assembly. Each pool's assembly typically included one contig for each clone. Each pool's assembly was split

into assemblies that corresponded to the individual clones by using *ace\_splitter* (<http://genome.uiowa.edu/pubsoft/software.html>). The assembly of each clone was viewed in Consed to determine the overall error rate and to identify low quality regions and regions covered in only one direction.

All full-insert cDNA sequences were finished according to the following quality criteria: no gaps; no ambiguous bases (Ns); cumulative average phrap score of at least 40 (error rate not to exceed one error in 10,000 bases), and a phrap score of at least 30 for each individual base in the assembly, <5% single stranded coverage, and, for those rare single-stranded covered regions, at least three reads of coverage, or two reads of coverage and a phrap score of at least 40 for each base. Clones that did not meet these standards were finished via directed sequencing by using custom oligonucleotides. The oligos were designed in consed and ordered from Integrated DNA Technologies, Inc. The sequence reads generated with the oligos were added to the assemblies of the reads derived from the respective clones until the quality standards were met. The sequence of all finished clones was translated and aligned to known sequences by using BLAT and BLAST to identify possible problems (frame shifts, retained introns, deletions, substitutions). Clones that appeared to have a full intact ORF were submitted to MGC as full-ORF clones. All other clones were submitted to GenBank as full-insert sequences. It is noteworthy that this sequencing pipeline also tracks the status of each clone, all oligos ordered and final sequences of each clone in an easy to use graphical format.

### ACKNOWLEDGMENTS

We would like to thank Dr. Michael Brownstein (National Institute of Mental Health, NIH) for kindly providing the pYX plasmid. We also thank Dr. Steven O. Moldin (National Institute of Mental Health, NIH), Dr. Hemin Chin (National Eye Institute, NIH), and Dr. Robert Strausberg (while director of the Cancer Genomics Office at the National Cancer Institute, NIH, currently vice president for research at The Institute for Genomic Research [TIGR]) for facilitating the coordination between this component of the NIH-Mouse BMAP and the NIH-MGC Program. Dr. Bair's work was supported by an NRSA post-doctoral fellowship no. 1F32HG002881-01A1. V.C.S. is an investigator of the Howard Hughes Medical Institute. This work was supported by contract no. N01MH12006 to the University of Iowa (M. Bento Soares, principal investigator), entitled "Gene Discovery in the Developing Nervous System."

### REFERENCES

- Bonaldo, M.F., Lennon, G., and Soares, M.B. 1996. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* **6**: 791-806.
- Bulfone, A., Smiga, S.M., Shimamura, K., Peterson, A., Puelles, L., and Rubenstein, J.L. 1995. T-brain-1: A homolog of Brachyury whose expression defines molecularly distinct domains within the cerebral cortex. *Neuron* **15**: 63-78.
- Carninci, P. and Hayashizaki, Y. 1999. High-efficiency full-length cDNA cloning. *Methods Enzymol.* **303**: 19-44.
- Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M., and Hayashizaki, Y. 2000. Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10**: 1617-1630.
- Carninci, P., Shibata, Y., Hayatsu, N., Itoh, M., Shiraki, T., Hirozane, T., Watahiki, A., Shibata, K., Konno, H., Muramatsu, M., et al. 2001. Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel  $\lambda$ -FLC family allows enhanced gene discovery rate and functional analysis. *Genomics* **77**: 79-90.
- Carninci, P., Shiraki, T., Mizuno, Y., Muramatsu, M., and Hayashizaki, Y. 2002. Extra-long first-strand cDNA synthesis. *Biotechniques* **32**: 984-985.
- Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Aizawa, K., Arakawa, T., Ishii, Y., Sasaki, D., et al. 2003. Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* **13**: 1273-1289.
- 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.

- Favaloro, J., Treisman, R., and Kamen, R. 1980. Transcription maps of polyoma virus-specific RNA: Analysis by two-dimensional nuclease S1 gel mapping. *Methods Enzymol.* **65**: 718–749.
- Gavin, A.J., Scheetz, T.E., Roberts, C.A., O'Leary, B., Braun, T.A., Sheffield, V.C., Soares, M.B., Robinson, J.P., and Casavant, T.L. 2002. Pooled library tissue tags for EST-based gene discovery. *Bioinformatics* **18**: 1162–1166.
- Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8**: 195–202.
- Gu, W., Wevers, A., Schroder, H., Grzeschik, K.H., Derst, C., Brodtkorb, E., de Vos, R., and Steinlein, O.K. 2002. The LGI1 gene involved in lateral temporal lobe epilepsy belongs to a new subfamily of leucine-rich repeat proteins. *FEBS Lett.* **519**: 71–76.
- Hirozane-Kishikawa, T., Shiraki, T., Waki, K., Nakamura, M., Arakawa, T., Kawai, J., Fagiolini, M., Hensch, T.K., Hayashizaki, Y., and Carninci, P. 2003. Subtraction of cap-trapped full-length cDNA libraries to select rare transcripts. *Biotechniques* **35**: 510–516, 518.
- Kalachikov, S., Evgrafov, O., Ross, B., Winawer, M., Barker-Cummings, C., Martinelli Boneschi, F., Choi, C., Morozov, P., Das, K., Teplitskaya, E., et al. 2002. Mutations in LGI1 cause autosomal-dominant partial epilepsy with auditory features. *Nat. Genet.* **30**: 335–341.
- Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**: 51–54.
- Kent, W.J. 2002. BLAT: The BLAST-like alignment tool. *Genome Res.* **12**: 656–664.
- Lesage, F., Duprat, F., Fink, M., Guillemare, E., Coppola, T., Lazdunski, M., and Hugnot, J.P. 1994. Cloning provides evidence for a family of inward rectifier and G-protein coupled K<sup>+</sup> channels in the brain. *FEBS Lett.* **353**: 37–42.
- Marra, M.A., Kucaba, T.A., Hillier, L.W., and Waterston, R.H. 1999. High-throughput plasmid DNA purification for 3 cents per sample. *Nucleic Acids Res.* **27**: e37.
- Maruyama, K. and Sugano, S. 1994. Oligo-capping: A simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138**: 171–174.
- Mittaz, L., Antonarakis, S.E., Higuchi, M., and Scott, H.S. 1997. Localization of a novel human RNA-editing deaminase (hRED2 or ADARB2) to chromosome 10p15. *Hum. Genet.* **100**: 398–400.
- Okada, Y., Yamazaki, H., Sekine-Aizawa, Y., and Hirokawa, N. 1995. The neuron-specific kinesin superfamily protein KIF1A is a unique monomeric motor for anterograde axonal transport of synaptic vesicle precursors. *Cell* **81**: 769–780.
- Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaïdo, I., Osato, N., Saito, R., Suzuki, H., et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**: 563–573.
- Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K., et al. 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* **36**: 40–45.
- Pang, Z., Zuo, J., and Morgan, J.I. 2000. Cbln3, a novel member of the precerebellin family that binds specifically to Cbln1. *J. Neurosci.* **20**: 6333–6339.
- Piao, Y., Ko, N.T., Lim, M.K., and Ko, M.S. 2001. Construction of long-transcript enriched cDNA libraries from submicrogram amounts of total RNAs by a universal PCR amplification method. *Genome Res.* **11**: 1553–1558.
- Puckett, C., Gomez, C.M., Korenberg, J.R., Tung, H., Meier, T.J., Chen, X.N., and Hood, L. 1991. Molecular cloning and chromosomal localization of one of the human glutamate receptor genes. *Proc. Natl. Acad. Sci.* **88**: 7557–7561.
- Sakakibara, Y., Suiko, M., Pai, T.G., Nakayama, T., Takami, Y., Katafuchi, J., and Liu, M.C. 2002. Highly conserved mouse and human brain sulfotransferases: Molecular cloning, expression, and functional characterization. *Gene* **285**: 39–47.
- Sheward, W.J., Lutz, E.M., Copp, A.J., and Harmar, A.J. 1998. Expression of PACAP, and PACAP type 1 (PAC1) receptor mRNA during development of the mouse embryo. *Brain Res. Dev. Brain Res.* **109**: 245–253.
- Shibata, Y., Carninci, P., Watahiki, A., Shiraki, T., Konno, H., Muramatsu, M., and Hayashizaki, Y. 2001. Cloning full-length, cap-trapper-selected cDNAs by using the single-strand linker ligation method. *Biotechniques* **30**: 1250–1254.
- Shimizu-Nishikawa, K., Kajiwara, K., Kimura, M., Katsuki, M., and Sugaya, E. 1995. Cloning and expression of SEZ-6, a brain-specific and seizure-related cDNA. *Brain Res. Mol. Brain Res.* **28**: 201–210.
- Soares, M.B. and Bonaldo, M.F. 1998. *Construction and screening of normalized cDNA libraries*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- 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.
- Strehl, S., Glatt, K., Liu, Q.M., Glatt, H., and Lalande, M. 1998. Characterization of two novel protocadherins (PCDH8 and PCDH9) localized on human chromosome 13 and mouse chromosome 14. *Genomics* **53**: 81–89.
- Suzuki, Y. and Sugano, S. 2001. Construction of full-length-enriched cDNA libraries: The oligo-capping method. *Methods Mol. Biol.* **175**: 143–153.
- . 2003. Construction of a full-length enriched and a 5'-end enriched cDNA library using the oligo-capping method. *Methods Mol. Biol.* **221**: 73–91.
- Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A., and Sugano, S. 1997. Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene* **200**: 149–156.
- Wu, Y.Q., Lin, X., Liu, C.M., Jamrich, M., and Shaffer, L.G. 2001. Identification of a human brain-specific gene, calneuron 1, a new member of the calmodulin superfamily. *Mol. Genet. Metab.* **72**: 343–350.

## WEB SITE REFERENCES

- <http://genome.uiowa.edu/techreports.html>; Publication repository for the Coordinated Laboratory for Computational Genomics.
- <http://trans.nih.gov/bmap/index.htm>; The BMAP Web site.
- <http://mgc.nci.nih.gov>; Primary Web site for the Mammalian Genome Project.
- <http://image.llnl.gov/image/html/vectors.shtml>; Information on all vectors used in the IMAGE clone collection.
- <http://genome.uiowa.edu/pubsoft/software.html>; Software distribution site for the Coordinated Laboratory for Computational Genomics.
- <http://www.genome.washington.edu/UWGC/analysis/tools/Phrap.cfm>; 1994—Basic description of the Phrap Program.

Received March 29, 2004; accepted in revised form April 27, 2004.