FINAL REPORT FOR GRANT DE-FG02-98ER62537

TECHNOLOGY DEVELOPMENT FOR GENE DISCOVERY AND FULL-LENGTH SEQUENCING

PERIOD COVERED: JANUARY 1, 2000 TO DECEMBER 31, 2003

Submitted to:

Department of Energy
Dr. Marvin Stodolsky (Marvin.Stodolsky@science.doe.gov)
OBER – Office of Biological and Environmental Research
Biology Division and Genome Task Group
(301) 903-4475

Submitted by:

Marcelo Bento Soares, Ph.D.
Professor of Pediatrics, Biochemistry, Physiology and Biophysics, Orthopaedics
The University of Iowa
Roy J. and Lucille A. Carver College of Medicine
375 Newton Road - 4184 Medical Education Biomedical Research Facility
Iowa City, IA 52242
phone: (319) 335-8250; fax: (319) 335-9565
e-mail: bento-soares@uiowa.edu

July 18, 2004

SUMMARY

In previous years, with support from the U.S. Department of Energy, we developed methods for construction of normalized and subtracted cDNA libraries, and constructed hundreds of high-quality libraries for production of Expressed Sequence Tags (ESTs) (Blackshear et al. 2001; Bonaldo et al. 1996; Dimopoulos et al. 2000; Franco et al. 1995; Grimmond et al. 2000; Ho et al. 1997; Lehane et al. 2003; Lennon et al. 1996; Netz et al. 2001; Reddy et al. 2002; Silveira de Oliveira et al. 2000; Soares et al. 1994; Urmenyi et al. 1999; Whitfield et al. 2002). Our clones were made widely available to the scientific community through the IMAGE Consortium (Lennon et al. 1996), and millions of ESTs were produced from our libraries either by collaborators (Adams et al. 1993; Berry et al. 1995; Blackshear et al. 2001; Franco et al. 1995; Hillier et al. 1996; Lehane et al. 2003; Marra et al. 1999; Verjovski-Almeida et al. 2003; Whitfield et al. 2002) or by our own sequencing laboratory at the University of Iowa (Dimopoulos et al. 2000; Gavin et al. 2002; Hackett et al. 2004; Laffin et al. 2004; Morcuende et al. 2002; Netz et al. 2001; Scheetz et al. 2004a; Scheetz et al. 2004b; Tuggle et al. 2003).

During this grant period, we focused on the development of a method for preferential cloning of tissue-specific and/or rare transcripts, its utilization to expedite EST-based gene discovery for the NIH Mouse Brain Molecular Anatomy Project, further development and optimization of a method for construction of full-length-enriched cDNA libraries, and modification of a plasmid vector to maximize efficiency of full-length cDNA sequencing by the transposon-mediated approach. Following is an outline of our main accomplishments during this grant period.

ACCOMPLISHMENTS

- Development of a method for preferential cloning of tissue-specific and/or rare mRNAs.
- Successful completion of a feasibility study aimed at assessing the efficacy of the method for preferential cloning of tissue-specific and/or rare mRNAs.
- Identification of hippocampus-specific transcripts of low prevalence or not represented in cDNA libraries derived from whole mouse brain using the method developed for cloning of rare mRNAs.
- Further development and optimization of a method for construction of full-length-enriched cDNA libraries.

- Further development and optimization of a vector designed to facilitate transposon-mediated full-insert cDNA sequencing.

DESCRIPTION OF EXPERIMENTAL APPROACHES AND RESULTS

A. METHOD FOR PREFERENTIAL CLONING OF TISSUE-SPECIFIC AND/OR RARE MRNAS

This method was developed to maximize the likelihood of representation of rare mRNAs in a cDNA library. A number of factors contribute to making cloning of rare mRNAs challenging, cloning efficiency being one of them. Typically, cloning efficiencies are not high enough to warrant representation of rare mRNAs (1-5 copies per cell) in cDNA libraries. Reassociation-kinetics analysis indicates that the mRNAs of a typical somatic cell are distributed in three frequency classes: (I) super prevalent [consisting of about 10-15 mRNAs which altogether represent 10-20% of the total mRNA mass), (II) intermediate [1-2,000 mRNAs; 40-45%] and (III) complex [15-20,000 mRNAs; 40-45%] (Bishop et al. 1974). The probability that a given mRNA will be represented in a cDNA library can be expressed by the equation $P(x) = 1-(1-f)^n$, where f=frequency and n=number of recombinant clones. Accordingly, the probability that an mRNA that is present at 1 copy per cell (1 in 500,000 total RNA molecules) will not be represented in a cDNA library of 1 million recombinants is 14%.

To minimize this problem we devised a strategy based on the idea that cloning of rare mRNAs could be facilitated by depletion of all previously identified mRNAs from the mRNA population prior to cDNA synthesis. We reasoned that if the representation of rare mRNAs in a total cellular RNA preparation could be increased by depletion of all previously identified mRNAs, their cloning would become less challenging. This can be accomplished in a two-step process involving hybridization of total cellular poly(A)+ mRNA with a driver comprising a comprehensive non-redundant collection of cDNAs representing thousands of transcripts previously identified in that mRNA population, followed by selection and isolation of the mRNAs that remain single strand. The latter can then be amplified according to established methodologies and the amplified RNA utilized for synthesis and cloning of cDNAs, thus generating a library enriched for rare mRNAs. Selection and isolation of unhybridized mRNAs can be accomplished in different ways, e.g. RNase H digestion of the RNAs in RNA:DNA heteroduplexes followed by isolation of the intact poly(A)+ single-stranded mRNA using streptavidin-coated oligo-[dT]-beads, or alternatively upon binding of the heteroduplexes to streptavidin-coated magnetic beads if a biotinylated driver is utilized.

To document the feasibility of this approach, we constructed a mouse hippocampus cDNA library enriched for rare mRNAs according to the procedure outlined below.

- A driver comprising 39 000 cDNAs that we identified for the NIH Mouse Brain Molecular Anatomy Project was generated by PCR amplification and hybridized with poly(A)+ RNA from mouse hippocampus.

- The mRNAs in heteroduplexes were destroyed with RNase H and the mRNA that remained intact was poly(A)+ selected and used as template for synthesis of double-strand cDNA according to standard procedures (Bonaldo et al. 1996), except that first-strand cDNA synthesis was primed with an oligonucleotide that contained a T7 RNA Polymerase promoter [5' T7 promoter – Not I – dT18].

- The resulting double-stranded cDNAs were ligated to an adaptor containing the promoter for the T3 RNA polymerase and utilized as template for synthesis of cRNA with T7 RNA Polymerase.

- The cDNA template was digested with RNase-free DNase and first-strand cDNA was synthesized from the cRNA using as primer an oligonucleotide complementary to the adapter sequence present at the 3' end of the cRNAs.

- The cRNA template was hydrolyzed and second strand cDNA was synthesized upon priming of the first-strand cDNA with an oligonucleotide complementary to the T7 – Not I sequence present at the 3' end of the first-strand cDNA.

- The resulting double-stranded cDNA was size-selected by gel filtration over a long (64-cm) and narrow (0.2cm diameter) Bio-gel A-50m (Bio-Rad, 100-200 mesh) column, ligated to EcoRI adapters, digested with NotI, directionally cloned into the *Not*I and *Eco*RI sites of the pT3T7-Pac vector and electroporated into DH10B *E.coli* bacteria.

B. CHARACTERIZATION OF A MOUSE HIPPOCAMPUS CDNA LIBRARY ENRICHED FOR TISSUE-SPECIFIC AND/OR RARE MRNAS

The mouse hippocampus cDNA library enriched for rare mRNAs that we constructed to test the feasibility of our method, named NIH_BMAP_MHI2_S1, was characterized by sequencing and by microarray hybridization.

Sequence characterization of a mouse hippocampus cDNA library enriched for rare mRNAs

A total of 8,835 high-quality 3' ESTs were generated from this library and a number of analyses were performed to determine whether our method was successful in enriching for rare mRNAs. First, to assess whether the subtractive hybridization step of the procedure was effective, we determined the representation of the driver population in this set of 8,835 ESTs. The driver population consisted of 39,000 cDNAs comprising 5,835 clusters. Of the 8,835 ESTs derived from the NIH_BMAP_NHI2_S1 library, 2,730 ESTs clustered in with a driver cluster; the remainder formed 3,918 new clusters. We therefore concluded that the subtraction was successful in reducing representation of the driver population. Next, to determine whether the method was successful in generating a library enriched for novel (often rare) mRNAs, we compared EST discovery rates achieved with this library with those of three other libraries:

(a) NIH_BMAP_MHI: a non-normalized mouse hippocampus library derived from the same mRNA as that utilized for construction of the NIH_BMAP_MHI2_S1 library;
(b) NIH_BMAP_MHI_N: a normalized mouse hippocampus cDNA library that was derived from NIH_BMAP_MHI;
(c) NIH_BMAP_M_S1: a subtracted library derived from a pool of normalized libraries, including NIH_BMAP_MHI_N, from ten regions of the mouse brain (cerebellum, brain stems, olfactory bulbs, hypothalamus, cortex, amygdala, basal ganglia, pineal gland, striatum, hippocampus); the driver used in this subtraction consisted of a pool of 20,000 BMAP cDNAs obtained from non-normalized and normalized libraries of these ten regions of the mouse brain.

More specifically, in this analysis we determined the number of EST clusters that were contributed exclusively by each of these libraries, as an indicator of their respective novelty rates, representation of rare mRNAs, and overall contribution to the gene discovery goals of the mouse

Brain Molecular Anatomy Project. The numbers presented below correspond to EST clusters exclusively contributed by each library:

(a) NIH_BMAP_MHI: 108 (from a total of 4,724 ESTs in 2,766 clusters)
(b) NIH_BMAP_MHI_N: 106 (from a total of 635 ESTs in 596 clusters)
(c) NIH_BMAP_M_S1: 1,459, of which 81 were derived from hippocampus (from a total of 3722 ESTs in 3,213 clusters)
(d) NIH_BMAP_MHI2_S1: 2,119 (from a total of 8835 ESTs in 4,715 clusters)

These analyses showed unequivocally that the NIH_BMAP_MHI2_S1 library from mouse hippocampus is well enriched for rare mRNAs, thus demonstrating the effectiveness of our method. Furthermore, it is noteworthy that this library enabled discovery of a large number of novel EST clusters representing mRNAs from hippocampus, thus contributing rather significantly to the gene discovery goals of the NIH Mouse Brain Molecular Anatomy Project.

cDNA microarray hybridization analysis of a mouse hippocampus cDNA library enriched for rare mRNAs

We performed a series of cDNA microarray hybridization experiments to demonstrate that the mouse hippocampus cDNA library constructed according to this novel procedure was indeed enriched for cDNAs derived from rare transcripts. We fabricated a glass slide cDNA microarray specifically for these experiments, with PCR products/probes obtained from 557 cDNA clones derived from three mouse hippocampus libraries: a non-normalized (NIH_BMAP_MHI), a normalized (NIH_BMAP_MHI_N), and this subtracted library enriched for rare mRNAs (NIH_BMAP_MHI2_S1). This probe set was printed four times in each slide. The Cy3- and Cy5-labeled cDNA targets used in these hybridizations were synthesized from RNA isolated from hippocampus and from the remainder of the mouse brain (i.e. whole brain minus hippocampus). A total of twelve hybridizations were performed (dye-swapping): six with Cy3- labeled cDNA derived from RNA obtained from hippocampus + Cy5-labeled cDNA derived from RNA obtained from the remaining of the mouse brain; six with Cy5-labed cDNA derived from RNA obtained from hippocampus + Cy3-labed cDNA derived RNA obtained from the remaining of the mouse brain. For each cDNA probe, we computed the weighted average from the four replicas present on each slide as the intensity level for each channel; the weights are reciprocal to the variances associated with the probes.

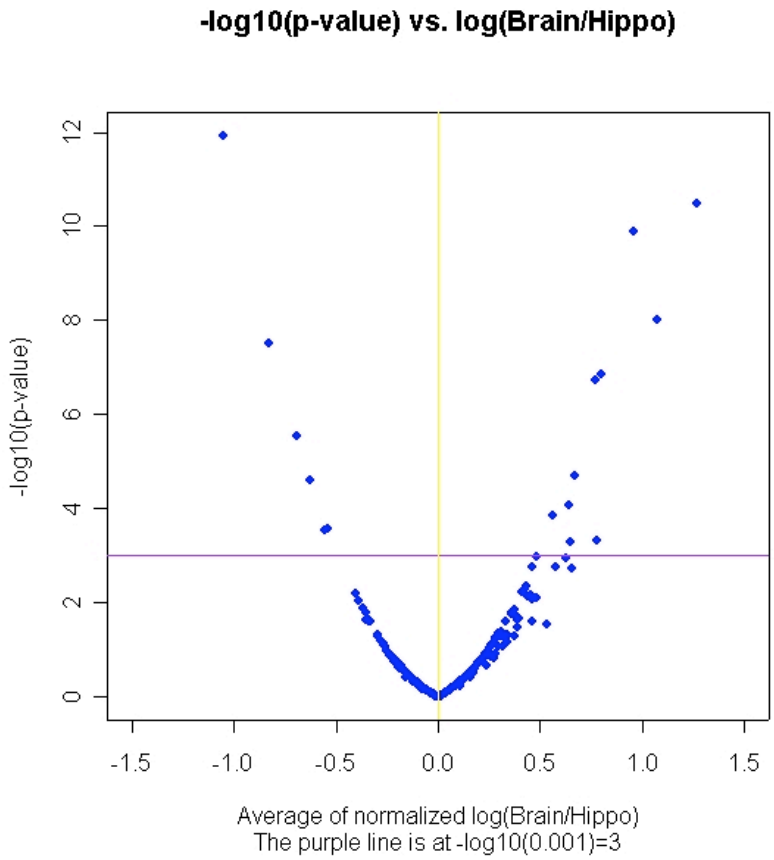

Figure 1: P-values: Volcano plot.

The p-value is calculated by conducting t-test for the intensity of each gene after normalization (taking SPLM correction for weighted average) from 12 slides. The p-values are between 0 and 1. The p-values were transformed into –log scale, so that smaller (more significant) p-values will now have more positive values in –log scale. P-value of 0.05 is approximately equal to 3 in –log scale (purple line in the graph). The graph is the plot of –log(p-value) against the difference in normalized intensity of whole brain without hippocampus versus hippocampus. In this volcano plot, the area encompassed from -1.5 to 0 on the X axis contains the probes corresponding to transcripts that are more highly expressed in hippocampus than in the remaining of the brain; those most significantly differentially expressed are in the upper left quadrant (above the purple line). In turn, the area from 0 to +1.5 on the X axis scale, comprises the probes representing transcripts expressed at lower levels in the hippocampus than in the remaining of the brain, with those most differentially expressed

falling in the upper right quadrant. The p-values and t-statistics values of probes falling in the two upper quadrants are presented on the table below.

Table 1: Top 16 genes from SPLM analysis

Gene ID	p-value	t-stat	t-nume	t-deno
UI-M-BZ1-bkx-h-02-0-UI.s1	0e+00	-7.1099	-1.0537	0.1482
UI-M-AQ0-aaj-g-08-0-UI.s1	0e+00	6.6316	1.2665	0.1910
UI-M-CD0-ayj-a-03-0-UI.s1	0e+00	6.4323	0.9625	0.1496
UI-M-AQ0-aae-h-02-0-UI.s1	0e+00	5.7338	1.0741	0.1873
UI-M-BZ1-bdp-f-01-0-UI.s1	0e+00	-5.5406	-0.8265	0.1492
UI-M-AQ0-aad-d-04-0-UI.s1	0e+00	5.2682	0.7999	0.1518
UI-M-BZ1-bfr-g-04-0-UI.s1	0e+00	5.2171	0.7735	0.1483
UI-M-BZ1-bfr-e-24-0-UI.s1	0e+00	-4.6855	-0.6939	0.1481
UI-M-BH3-awc-g-02-0-UI.s4	0e+00	4.2744	0.6685	0.1564
UI-M-BZ1-blq-d-04-0-UI.s1	0e+00	-4.2107	-0.6271	0.1489
UI-M-AQ0-aah-e-06-0-UI.s1	1e-04	3.9358	0.6428	0.1633
UI-M-BZ1-blk-g-12-0-UI.s1	1e-04	3.8087	0.5656	0.1485
UI-M-BZ1-bll-f-11-0-UI.s1	3e-04	-3.6384	-0.5387	0.1481
UI-M-BZ1-bfs-c-06-0-UI.s1	3e-04	-3.6275	-0.5575	0.1537
UI-M-AQ0-aaj-b-11-0-UI.s1	5e-04	3.5063	0.7772	0.2216
UI-M-AH0-acs-c-03-0-UI.s1	5e-04	3.4743	0.6459	0.1859

Transcripts that are expressed at a significantly higher level in hippocampus than in the remaining of the brain have negative t-statistics values while those expressed at significantly higher level in the remaining of the brain have positive t-statistics values. As evidenced in this table, 6 out of the 8 probes derived from the NIH_BMAP_MHI2_S1 library (clone names starting with UI-M-

BZ1) represent transcripts expressed at a significantly higher level in hippocampus than in the remaining of the brain (negative t-statistics values), as expected if our method was successful in enriching for hippocampus-specific and/or rare mRNAs. The remainder probes, prevalent in brain regions other than the hippocampus (positive t-statistics values) were derived from the other libraries (clone names starting with UI-M-AH0, UI-M-AQ0, UI-M-BH3, UI-M-CD0).

In conclusion, both sequence and microarray analyses demonstrated the effectiveness of the method that we developed under DOE support to generate cDNA libraries enriched for tissue-specific and/or rare mRNAs.

C. Optimization of a method for construction of full-length-enriched cDNA libraries

The approach we developed for construction of full-length-enriched cDNA libraries involves four principal steps: (a) size-fractionation and purification of high-quality cytoplasmic poly(A)+ mRNAs, (b) synthesis of oligo-dT-primed first-strand cDNA from each mRNA size-fraction, individually, utilizing RNaseH$^-$ Reverse Transcriptase under optimized conditions to yield full-length cDNAs with short 5' dT-tails, (c) 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 (d) separate cloning and limited amplification of cDNAs in different size ranges, utilizing a plasmid vector designed to facilitate transposon-mediated sequencing.

Ultimately, the purpose of the two most distinctive attributes of this approach, i.e. (*i*) the serial and corresponding size fractionation of template (cytoplasmic mRNA) and product (double-stranded cDNAs), and (*ii*) 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. Since mRNA complexity is lower in a size-fraction 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.

During this grant period, we conducted a number of experiments aimed at optimizing several steps in this procedure, and in particular the method utilized for size fractionation of the mRNA. Different gel conditions were tested and experiments were performed to assess effectiveness of size selection and intactness of the mRNA after fractionation. Our efforts were very successful in optimizing and streamlining these procedures, which we utilized for the construction of approximately fifty full-length-enriched cDNA libraries from the developing mouse nervous system for the NIH Mouse Brain Molecular Anatomy Project (Bonaldo et al., 2004, Genome Research, in press). In addition, given the ultimate goal of utilizing the full-length-enriched libraries generated with this method for production of complete and accurate sequence of full-ORF-containing cDNAs, we modified and thoroughly tested a plasmid vector specifically designed to facilitate transposon-mediated sequencing. The resulting vector, pYX-AscI, is a 1,691 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 non-essential 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) 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 non-viable clone. Additional information on the pYX-AscI vector, including its complete sequence, can be obtained at http://image.llnl.gov/image/html/vectors.shtml. To date, we have generated complete and accurate sequence from over 1,500 full-ORF-containing cDNAs that we identified in the full-length-enriched libraries that we constructed for the NIH Mouse Brain Molecular Anatomy Project, all of which have been contributed to the NIH Mammalian Gene Collection program. A manuscript describing the construction of the full-length-enriched libraries and sequencing of complete open reading frames of approximately 1,500 transcripts expressed in the developing mouse nervous system will be published later this year in a special issue of Genome Research on the ORFeome (Bonaldo et al: "1,274 full-open reading frames of transcripts expressed in the developing mouse nervous system").

DATA DISSEMINATION

Data produced during this funding period were presented at the "TRANSCRIPTOME" Conference series and at the "Beyond Identification of Transcribed Sequences" Workshops. All cDNA libraries and sequences generated in this project have been made publicly available. A manuscript describing the new method developed during this grant period for preferential cloning of tissue-specific and/or rare mRNAs is in preparation (Malchenko, Bonaldo and Soares).

LITERATURE CITED

Adams, M.D., M.B. Soares, A.R. Kerlavage, C. Fields, and J.C. Venter. 1993. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat Genet* **4:** 373-380.

Berry, R., T.J. Stevens, N.A. Walter, A.S. Wilcox, T. Rubano, J.A. Hopkins, J. Weber, R. Goold, M.B. Soares, and J.M. Sikela. 1995. Gene-based sequence-tagged-sites (STSs) as the basis for a human gene map. *Nat Genet* **10:** 415-423.

Bishop, J.O., J.G. Morton, M. Rosbash, and M. Richardson. 1974. Three abundance classes in HeLa cell messenger RNA. *Nature* **250:** 199-204.

Blackshear, P.J., W.S. Lai, J.M. Thorn, E.A. Kennington, N.G. Staffa, D.T. Moore, G.G. Bouffard, S.M. Beckstrom-Sternberg, J.W. Touchman, M.F. Bonaldo, and M.B. Soares. 2001. The NIEHS Xenopus maternal EST project: interim analysis of the first 13,879 ESTs from unfertilized eggs. *Gene* **267:** 71-87.

Bonaldo, M., G. Lennon, and M. Soares. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* **6:** 791-806.

Bonaldo, M.F., Bair, T.B., Scheetz, T.E., Snir, E., Akabogu, I., Bair, J.L., Berger, B., Crouch, K., Davis, A., Eyestone, M., Keppel, C., Kucaba, T., Lebeck, M., Lin, J.L., Melo, A.I.R., Rehmann, J., Reiter, R.S., Schaefer, K., Smith, C., Tack, D., Trout, K., Sheffield, V.C., Lin, J.J-C., Casavant, T.C. and Soares, M.B. (2004). 1,274 full-open reading frames of transcripts expressed in the developing mouse nervous system. *Genome Research*, in press.

Dimopoulos, G., T.L. Casavant, S. Chang, T. Scheetz, C. Roberts, M. Donohue, J. Schultz, V. Benes, P. Bork, W. Ansorge, M.B. Soares, and F.C. Kafatos. 2000. Anopheles gambiae pilot gene discovery project: identification of mosquito innate immunity genes from expressed sequence tags generated from immune-competent cell lines. *Proc Natl Acad Sci U S A* **97:** 6619-6624.

Franco, G.R., M.D. Adams, M.B. Soares, A.J. Simpson, J.C. Venter, and S.D. Pena. 1995. Identification of new Schistosoma mansoni genes by the EST strategy using a directional cDNA library. *Gene* **152:** 141-147.

Gavin, A.J., T.E. Scheetz, C.A. Roberts, B. O'Leary, T.A. Braun, V.C. Sheffield, M.B. Soares, J.P. Robinson, and T.L. Casavant. 2002. Pooled library tissue tags for EST-based gene discovery. *Bioinformatics* **18:** 1162-1166.

Grimmond, S., N. Van Hateren, P. Siggers, R. Arkell, R. Larder, M.B. Soares, M. de Fatima Bonaldo, L. Smith, Z. Tymowska-Lalanne, C. Wells, and A. Greenfield. 2000. Sexually dimorphic expression of protease nexin-1 and vanin-1 in the developing mouse gonad prior to overt differentiation suggests a role in mammalian sexual development. *Hum Mol Genet* **9:** 1553-1560.

Hackett, J.D., H.S. Yoon, M.B. Soares, M.F. Bonaldo, T.L. Casavant, T.E. Scheetz, T. Nosenko, and D. Bhattacharya. 2004. Migration of the plastid genome to the nucleus in a peridinin dinoflagellate. *Curr Biol* **14:** 213-218.

Hillier, L.D., G. Lennon, M. Becker, M.F. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, W. Gish, M. Hawkins, M. Hultman, T. Kucaba, M. Lacy, M. Le, N. Le, E. Mardis, B. Moore, M. Morris, J. Parsons, C. Prange, L. Rifkin, T. Rohlfing, K. Schellenberg, M. Marra, and et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res* **6:** 807-828.

Ho, P.L., M.B. Soares, T. Maack, I. Gimenez, G. Puorto, M.F. Furtado, and I. Raw. 1997. Cloning of an unusual natriuretic peptide from the South American coral snake Micrurus corallinus. *Eur J Biochem* **250:** 144-149.

Laffin, J.J., T.E. Scheetz, F. Bonaldo Mde, R.S. Reiter, S. Chang, M. Eyestone, H. Abdulkawy, B. Brown, C. Roberts, D. Tack, T. Kucaba, J.J. Lin, V.C. Sheffield, T.L. Casavant, and M.B. Soares. 2004. A comprehensive nonredundant expressed sequence tag collection for the developing Rattus norvegicus heart. *Physiol Genomics* **17:** 245-252.

Lehane, M.J., S. Aksoy, W. Gibson, A. Kerhornou, M. Berriman, J. Hamilton, M.B. Soares, M.F. Bonaldo, S. Lehane, and N. Hall. 2003. Adult midgut expressed sequence tags from the tsetse fly Glossina morsitans morsitans and expression analysis of putative immune response genes. *Genome Biol* **4:** R63.

Lennon, G., C. Auffray, M. Polymeropoulos, and M.B. Soares. 1996. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33:** 151-152.

Marra, M., L. Hillier, T. Kucaba, M. Allen, R. Barstead, C. Beck, A. Blistain, M. Bonaldo, Y. Bowers, L. Bowles, M. Cardenas, A. Chamberlain, J. Chappell, S. Clifton, A. Favello, S. Geisel, M. Gibbons, N. Harvey, F. Hill, Y. Jackson, S. Kohn, G. Lennon, E. Mardis, J. Martin, R. Waterston, and et al. 1999. An encyclopedia of mouse genes. *Nat Genet* **21:** 191-194.

Morcuende, J.A., X.D. Huang, J. Stevens, T.A. Kucaba, B. Brown, H. Abdulkawy, T.E. Scheetz, S. Malchenko, F. Bonaldo, T.L. Casavant, and B. Soares. 2002. Identification and initial characterization of 6,000 expressed sequenced tags (ESTs) from rat normal-growing cartilage and swarm rat chondrosarcoma cDNA libraries. *Iowa Orthop J* **22:** 28-34.

Netz, D.J., H.G. Sahl, R. Marcelino, J. dos Santos Nascimento, S.S. de Oliveira, M.B. Soares, M. do Carmo de Freire Bastos, and R. Marcolino. 2001. Molecular characterisation of aureocin A70, a multi-peptide bacteriocin isolated from Staphylococcus aureus. *J Mol Biol* **311:** 939-949.

Reddy, A.R., W. Ramakrishna, A.C. Sekhar, N. Ithal, P.R. Babu, M.F. Bonaldo, M.B. Soares, and J.L. Bennetzen. 2002. Novel genes are enriched in normalized cDNA libraries from drought-stressed seedlings of rice (Oryza sativa L. subsp. indica cv. Nagina 22). *Genome* **45:** 204-211.

Scheetz, T.E., J.J. Laffin, B. Berger, S. Holte, S.A. Baumes, R. Brown, 2nd, S. Chang, J. Coco, J. Conklin, K. Crouch, M. Donohue, G. Doonan, C. Estes, M. Eyestone, K. Fishler, J. Gardiner, L. Guo, B. Johnson, C. Keppel, R. Kreger, M. Lebeck, R. Marcelino, V. Miljkovich, M. Perdue, L. Qui, J. Rehmann, R.S. Reiter, B. Rhoads, K. Schaefer, C. Smith, I. Sunjevaric, K. Trout, N. Wu, C.L. Birkett, J. Bischof, B. Gackle, A. Gavin, A.J. Grundstad, B. Mokrzycki, C. Moressi, B. O'Leary, K. Pedretti, C. Roberts, N.L. Robinson, M. Smith, D. Tack, N. Trivedi, T. Kucaba, T. Freeman, J.J. Lin, M.F. Bonaldo, T.L. Casavant, V.C. Sheffield, and M.B. Soares. 2004a. High-throughput gene discovery in the rat. *Genome Res* **14:** 733-741.

Scheetz, T.E., J. Zabner, M.J. Welsh, J. Coco, F. Eyestone Mde, M. Bonaldo, T. Kucaba, T.L. Casavant, M.B. Soares, and P.B. McCray, Jr. 2004b. Large-scale gene discovery in human airway epithelia reveals novel transcripts. *Physiol Genomics* **17:** 69-77.

Silveira de Oliveira, J., A. Rossan de Brandao Prieto da Silva, M.B. Soares, M.A. Stephano, W. de Oliveira Dias, I. Raw, and P.L. Ho. 2000. Cloning and characterization of an alpha-neurotoxin-type protein specific for the coral snake Micrurus corallinus. *Biochem Biophys Res Commun* **267:** 887-891.

Soares, M.B., M.F. Bonaldo, P. Jelene, L. Su, L. Lawton, and A. Efstratiadis. 1994. Construction and characterization of a normalized cDNA library. *Proc Natl Acad Sci U S A* **91:** 9228-9232.

Tuggle, C.K., J.A. Green, C. Fitzsimmons, R. Woods, R.S. Prather, S. Malchenko, B.M. Soares, T. Kucaba, K. Crouch, C. Smith, D. Tack, N. Robinson, B. O'Leary, T. Scheetz, T. Casavant, D. Pomp, B.J. Edeal, Y. Zhang, M.F. Rothschild, K. Garwood, and W. Beavis. 2003. EST-based gene discovery in pig: virtual expression patterns and comparative mapping to human. *Mamm Genome* **14:** 565-579.

Urmenyi, T.P., M.F. Bonaldo, M.B. Soares, and E. Rondinelli. 1999. Construction of a normalized cDNA library for the Trypanosoma cruzi genome project. *J Eukaryot Microbiol* **46:** 542-544.

Verjovski-Almeida, S., R. DeMarco, E.A. Martins, P.E. Guimaraes, E.P. Ojopi, A.C. Paquola, J.P. Piazza, M.Y. Nishiyama, Jr., J.P. Kitajima, R.E. Adamson, P.D. Ashton, M.F. Bonaldo, P.S. Coulson, G.P. Dillon, L.P. Farias, S.P. Gregorio, P.L. Ho, R.A. Leite, L.C. Malaquias, R.C. Marques, P.A. Miyasato, A.L. Nascimento, F.P. Ohlweiler, E.M. Reis, M.A. Ribeiro, R.G. Sa, G.C. Stukart, M.B. Soares, C. Gargioni, T. Kawano, V. Rodrigues, A.M. Madeira, R.A. Wilson, C.F. Menck, J.C. Setubal, L.C. Leite, and E. Dias-Neto. 2003. Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni. *Nat Genet* **35:** 148-157.

Whitfield, C.W., M.R. Band, M.F. Bonaldo, C.G. Kumar, L. Liu, J.R. Pardinas, H.M. Robertson, M.B. Soares, and G.E. Robinson. 2002. Annotated Expressed Sequence Tags and cDNA Microarrays for Studies of Brain and Behavior in the Honey Bee. *Genome Res.* **12:** 555-566.