doi:10.1006/geno.2002.6698, available online at http://www.idealibrary.com on IDEAL

Profiling Gene Expression Using Onto-Express

Purvesh Khatri,[1] Sorin Draghici,[1,2] G. Charles Ostermeier,[3] and Stephen A. Krawetz[2,3,*]

[1]*Department of Computer Science,* [2]*Institute for Scientific Computing, and* [3]*Department of Obstetrics and Gynecology, Molecular Medicine and Genetics, Wayne State University School of Medicine, Detroit, MI 48201, USA*

**To whom correspondence and reprint requests should be addressed. Fax: (313) 577-8554. E-mail: steve@compbio.med.wayne.edu.*

Gene expression profiles obtained through microarray or data mining analyses often exist as vast data strings. To interpret the biology of these genetic profiles, investigators must analyze this data in the context of other information such as the biological, biochemical, or molecular function of the translated proteins. This is particularly challenging for a human analyst because large quantities of less than relevant data often bury such information. To address this need we implemented an automated routine, called Onto-Express (http://vortex.cs.wayne.edu:8080), to systematically translate genetic fingerprints into functional profiles. Using strings of accession or cluster identification numbers, Onto-Express searches the public databases and returns tables that correlate expression profiles with the cytogenetic locations, biochemical and molecular functions, biological processes, cellular components, and cellular roles of the translated proteins. The profiles created by Onto-Express fundamentally increase the value of gene expression analyses by facilitating the translation of quantitative value sets to records that contain biological implications.

INTRODUCTION

The development of microarray technologies permits researchers to monitor mRNA transcript levels for thousands of genes in a single experiment [1–5]. The sheer power and potential for discovery using this whole transcriptome analysis [6,7] move well beyond traditional strategies that rely on the examination of a single gene or the assessment of a group of genes one at a time. The resulting patterns, that is expression profiles, depict subsets of transcripts that qualitatively reflect gene activity at a given instant or in response to a particular treatment or condition [8,9]. These profiles represent a genetic "fingerprint" that characterizes the cell or tissue being studied (G.C.O. *et al.*, manuscript submitted) [10] and provide a foundation from which to begin an investigation of the underlying biology. Biological functionality must be associated with each element of the genetic fingerprint to appreciate the intricacies of how specific tissues or cells function in a normal state or a disease state, or in response to a given treatment. A description of functionality should include the structural, regulatory, or enzymatic roles of the corresponding proteins; the biological "objectives" to which the proteins contribute; the places within a cell where the proteins are active; and the major biological processes in which these proteins play a role as well as their biochemical activities.

Databases are being constructed to store expression data [11] and algorithms are being implemented to support access to this information [12,13]. Even though these efforts are underway, the interoperability required to convert such data into functional biological profiles is lacking. This is a daunting task when one considers that genetic fingerprints can contain tens of thousands of records (G.C.O. *et al.*, manuscript submitted) [5]. To this end, we developed the Java-based program called Onto-Express (http://vortex.cs.wayne.edu:8080). This algorithm uses UniGene accession or cluster identification numbers generated from microarray investigations to search the public databases to return the functional profiles for each genetic fingerprint. Here we demonstrate the utility of Onto-Express.

RESULTS AND DISCUSSION

Microarray-based genomic surveys and other high-throughput strategies are becoming increasingly important in biology. These methods typically produce data sets containing massive lists of expressed or repressed genes. This information provides only a starting point from which to begin the investigation of specific biological processes. To begin understanding how cells function within a tissue in a normal or diseased state or respond to specific stimuli, these data must be decoded. To facilitate interpretation and stratify the data, information concerning the biological, biochemical, or molecular functions of the translated proteins must be provided.

The stratification of *in silico* and biological data can be an undertaking of epic proportions when investigators use the typical biological paradigm of examining one element at a time. To overcome this impasse, we created Onto-Express.

0888-7543/01 $35.00

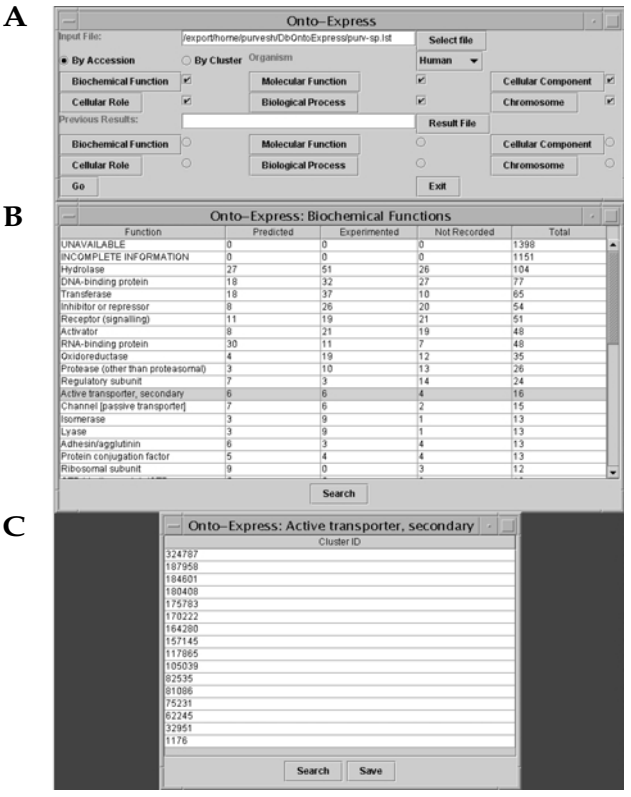

FIG. 1. Program interface windows. Onto-Express uses a graphical user interface. (A) Program window. To initiate the program, the user enters the raw data file path in the appropriate text field. To indicate the format and species contained in the raw data file, the user selects the corresponding radio button (by accession code or by cluster identification code) and organism, for example, human. The checkboxes indicating the desired information are selected and the "Go" button is activated to start data retrieval. (B) Result window. After data retrieval, Onto-Express returns tables that contain the desired information, such as the biochemical functions of the translated proteins. The user can select a function of interest, such as active transporters, secondary, and then click the search button. This will open a window showing the cluster identification numbers that correspond to the function of interest as shown in (C). (C) Search window. The user can highlight one of the cluster identification numbers and click search, which will open the corresponding NCBI web page. Alternatively, the user can click the "save" button and store the generated list in a user-specified file.

This program systematically associates gene expression profiles with the chromosome and cytogenetic gene locations, as well as the biochemical and molecular functions, biological processes, cellular components, and cellular roles of the translated proteins.

Onto-Express starts by reading the input file that exclusively contains single line entries of either accession or cluster identification numbers. The database is built from the LL_tmpl file, which is updated weekly (and can be obtained from ftp://ftp.ncbi.nih.gov/refseq/LocusLink). After creating the database, Onto-Express queries the NCBI map viewer and retrieves the number of genes on each chromosome. If the input file contains accession numbers, Onto-Express retrieves the cluster identification number for each accession number and then builds a list of cluster identification numbers. This eliminates redundancy, as multiple accession numbers are often linked to a single cluster identification number. After building the list of cluster identification numbers, Onto-Express retrieves the locus identification number for each cluster. Onto-Express then queries the database using each locus identification number to collect as much information as possible. For each locus this includes biological process, biochemical function, molecular function, cellular role, cellular component, and chromosomal location. The program builds the corresponding list in memory. If the chromosomal location of a cluster is not available in the database created, Onto-Express queries the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene/). The software could query other databases such as Ensembl (http://us.ensembl.org/) to retrieve additional information or attempt to complete the analysis.

In those cases when an accession number corresponding to a cluster identification number or locus identification number is not available, or for a cluster identification number for which the corresponding locus identification number is not available, it is classified as "Unavailable." If partial information is available for an accession number or for a cluster identification number, the software classifies the observation as "Incomplete."

Onto-Express stores the information collected as simple text files. If the input file contains accession numbers, the software prefixes the output filenames with "accn." Similarly, if the input file contains cluster identification numbers, Onto-Express prefixes the output filenames with "clst." The first line of the output file states which data are stored followed by a semicolon-delimited listing of the acquired data. Once all of the information is collected, Onto-Express displays the information in tabular form as the user requests.

Figure 1 shows the user interface windows of Onto-Express. The user first enters the input filename that contains the list of either accession or cluster identification numbers. Alternatively, one can browse through directories by pressing the "Select File" button to locate the input file. The user then informs Onto-Express of the input file format by selecting either the "By Accession" or "By Cluster" radio button. Next, the user selects the organism studied from the drop-down list. Subsequently, the user chooses the information they would like to retrieve by selecting or deselecting the check boxes. By default, the software collects all the information. To start data collection, the user activates the "Go" button. The "Go" button then remains unavailable until Onto-Express finishes collecting the data. Version 1.0 processes one input file at a time. While Onto-Express is collecting the information, the upper six buttons, labeled "Biochemical Function," "Molecular Function," "Cellular Component," "Cellular Role," "Biological Process," and "Chromosome," are unavailable (that is, grayed out). They are available only after Onto-Express has collected all of the information. When available, the user can click any of the buttons to display the

doi:10.1006/geno.2002.6698, available online at http://www.idealibrary.com on IDEAL

A

Onto–Express: Biochemical Functions

Function	Predicted	Experimented	Not Recorded	Total
UNAVAILABLE	0	0	0	1398
INCOMPLETE INFORMATION	0	0	0	1151
Hydrolase	27	51	26	104
DNA-binding protein	18	32	27	77
Transferase	18	37	10	65
Inhibitor or repressor	8	26	20	54
Receptor (signalling)	11	19	21	51
Activator	8	21	19	48
RNA-binding protein	30	11	7	48
Oxidoreductase	4	19	12	35
Protease (other than proteasomal)	3	10	13	26
Regulatory subunit	7	3	14	24
Active transporter, secondary	6	6	4	16
Channel [passive transporter]	7	6	2	15
Isomerase	3	9	1	13
Lyase	3	9	1	13
Adhesin/agglutinin	6	3	4	13
Protein conjugation factor	5	4	4	13
Ribosomal subunit	9	0	3	12

Search

B

Onto–Express: Biological Processes

Function	Predicted	Experimented	Not Recorded	Total
UNAVAILABLE	0	0	0	1398
INCOMPLETE INFORMATION	0	0	0	979
signal transduction	46	28	14	88
cell proliferation	24	13	1	38
oncogenesis	28	6	2	36
transcription regulation from Pol II promoter	7	24	3	34
transcription from Pol II promoter	11	14	4	29
developmental processes	21	3	4	28
cell adhesion	14	6	6	26
small molecule transport	15	9	1	25
cell cycle control	13	5	6	24
cell-cell signalling	9	8	6	23
neurogenesis	14	3	3	20
protein modification	10	7	1	18
immune response	9	4	5	18
proteolysis and peptidolysis	8	7	3	18
protein phosphorylation	7	10	0	17
protein complex assembly	6	7	3	16
energy pathways	10	1	5	16

Search

C

Onto–Express: Cellular Components

Function	Predicted	Experimented	Not Recorded	Total
UNAVAILABLE	0	0	0	1398
INCOMPLETE INFORMATION	0	0	0	1166
integral plasma membrane protein	75	16	12	103
nucleus	37	54	9	100
cytoplasm	11	40	14	65
membrane fraction	24	18	19	61
plasma membrane	9	28	12	49
integral membrane protein	30	3	4	37
soluble fraction	5	13	6	24
mitochondrion	7	9	8	24
extracellular space	9	7	8	24
endoplasmic reticulum	6	10	2	18
peripheral plasma membrane protein	7	3	3	13
cytosol	1	8	2	11
cytoskeleton	5	1	5	11
extracellular	4	4	2	10
lysosome	1	2	7	10
Golgi apparatus	0	9	1	10
actin cytoskeleton	6	2	2	10

Search

D

Onto–Express: Cellular Roles

Function	Predicted	Experimented	Not Recorded	Total
UNAVAILABLE	0	0	0	1398
INCOMPLETE INFORMATION	0	0	0	1394
Pol II transcription	20	30	33	83
Cell structure	11	6	25	42
RNA processing/modification	18	8	3	29
Energy generation	8	4	14	26
Protein synthesis	12	5	6	23
Protein degradation	4	2	17	23
Other metabolism	6	9	5	20
DNA repair	10	5	3	18
Differentiation	7	5	6	18
Cell stress	8	1	3	12
Chromatin/chromosome structure	8	2	2	12
Amino-acid metabolism	2	7	3	12
DNA synthesis	5	2	3	10
RNA splicing	5	5	0	10
Protein translocation	4	5	1	10
Nuclear-cytoplasmic transport	0	1	2	3
Pol I transcription	0	1	1	2

Search

E

Onto–Express: Molecular Functions

Function	Predicted	Experimented	Not Recorded	Total
UNAVAILABLE	0	0	0	1398
INCOMPLETE INFORMATION	0	0	0	970
transcription factor	21	18	9	48
DNA binding	14	19	4	37
protein binding	6	28	2	36
RNA binding	22	5	4	31
tumor suppressor	23	5	1	29
transcription co-activator	7	17	2	26
transcription co-repressor	2	18	0	20
RNA polymerase II transcription factor	4	13	3	20
receptor	7	9	3	19
adenosinetriphosphatase	10	6	1	17
protein serine/threonine kinase	7	10	0	17
signal transduction	10	4	2	16
cell adhesion	5	5	6	16
transcription activating factor	1	14	0	15
cytoskeletal structural protein	6	2	7	15
protein kinase	6	7	1	14
electron transporter	5	7	1	13

Search

F

Onto–Express: Chromosome

Chromos...	p-Arm	q-Arm	unknown-Arm	Total	NCBI Total	% Located
1	82	45	157	284	1804	0.15742794
2	44	45	115	204	1104	0.18478261
3	51	26	71	148	956	0.15481171
4	15	23	68	106	658	0.16109422
5	16	47	89	152	844	0.1800948
6	54	32	77	163	978	0.16666667
7	19	34	74	127	815	0.15582822
8	30	15	62	107	591	0.18104906
9	18	32	61	111	693	0.16017316
10	20	30	83	133	659	0.20182094
11	36	49	82	167	988	0.16902834
12	26	39	88	153	944	0.16207626
13	0	30	35	65	314	0.20700637
14	0	42	36	78	554	0.14079422
15	0	30	42	72	565	0.12743363
16	35	30	46	111	751	0.14780293
17	32	56	54	142	1011	0.140455
18	9	15	25	49	293	0.1672355
19	42	54	53	149	1179	0.12637828

Search

FIG. 2. Result windows. Onto-Express returns six different categories of information. The following panels represent these groupings. (A) Biochemical function. These data depict the principle structural, regulatory, or enzymatic function of the proteins. (B) Biological process. This panel describes the biological "objective" to which the proteins contribute. (C) Cellular component. The information in this panel illustrates the place in a cell where the protein is active. (D) Cellular role. The major biological process involving the protein is demonstrated in this panel. (E) Molecular function. The biochemical activity of the proteins is summarized by these data. The first column in each panel lists specific instances. The second column represents the number of instances that are predicted by similarity or by analogy to other proteins, whereas the third column represents the number of times the specific instances were supported by experimental evidence. The fourth column represents how many times the instance was identified but the type of supporting evidence was not recorded. The fifth column shows the sum of the second, third, and fourth columns. (F) Chromosome. This panel represents data concerning the chromosomal location of the genes assessed. The first column reports the specific chromosome number. The second and third columns represent the number of genes located on p and q arms, respectively. The fourth column explains the number of genes for which the exact location was not reported. The fifth column is the sum of the second, third, and fourth columns. The sixth column represents the total number of genes known to be located on the respective chromosomes and used to standardize the total of genes identified, as shown in the last column.

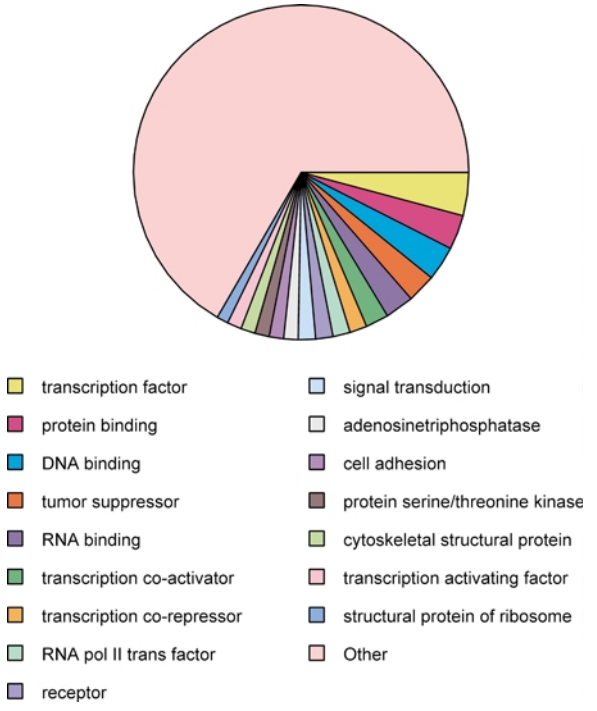

FIG. 3. Ontogeny of spermatozoal RNA. mRNAs were isolated from sperm and used to create labeled cDNAs for microarray analysis. The molecular function of the proteins that represent each expressed sequence tag identified were data mined using Onto-Express. The different sections of the pie chart represent the proportion of proteins identified to have the molecular function that is indicated by color in the legend. "Other" indicates protein groups with fewer than 14 observations.

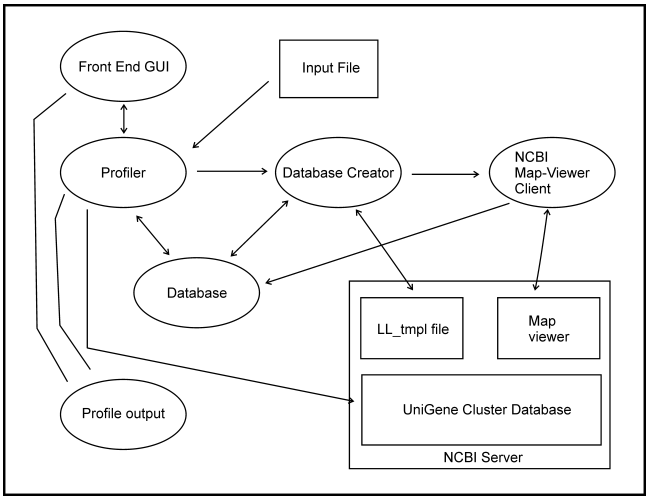

FIG. 4. Software architecture of Onto-Express. Onto-Express uses a GUI interface, a profiler, a database creator, an NCBI map-viewer client, the NCBI server, and database components. The profiler invokes the database creator, which in turn invokes the NCBI map-viewer client. The client queries the NCBI map-viewer and stores the information retrieved in the database. Following the database creator populating the database from the "LL_tmpl" file, the profiler reads the input file. If the input file is comprised of accession numbers, the algorithm retrieves their respective UniGene cluster identification numbers. For each cluster identification number, the algorithm determines the locus identification number to query the database for biochemical function, biological process, cellular component, cellular role, molecular function, and chromosomal location. The profiler stores the information in simple text files, which the front-end GUI displays in tabular format.

respective information in a tabular format as shown in Fig. 1B.

While Onto-Express is collecting information associated with the input file, the user may view any of the previously collected data. The user can view the current results at a later point in time by activating the lower six buttons. The user can either enter the output file name or browse through directories to select the previously generated output file of interest.

Figure 2 shows the results for biochemical function, cellular component, cellular role, molecular function, biological process, and chromosomal location. The information is presented in five columns (Figs. 2A–2E). The first column denotes the function name. The second row of this column, labeled "Unavailable," contains the number of hits for which the corresponding locus identification number was not available, while the third row, labeled "Incomplete Information," contains the number of hits for which partial information was available. The second and third columns are labeled "Predicted" and "Experimented" and show the number of accession or cluster identification numbers having predicted and experimental proof provided for a given function. The fourth column presents how many times the function was identified for which supporting evidence was "Not Recorded." The fifth column, labeled "Total," is the sum of second, third, and fourth columns.

Figure 2F shows the related chromosomal information. The first column is the chromosome number, the second column is the total of accession numbers or cluster identification numbers located on the p-Arm, the third is the total of accession numbers or cluster identification numbers located on the q-Arm, followed by the fourth column indicating the number of ESTs identified on that chromosome for which their exact location is not known. The fifth column is the sum of the second, third, and fourth columns. The sixth column, labeled "NCBI Total," shows the number of genes located on that chromosome. The seventh column, labeled "% located," is the ratio of the fifth and the sixth columns expressed as a percentage.

Application of Onto-Express

It is well documented that mature spermatozoa contain mRNA [14–20]. However, the function or purpose of the mRNAs have yet to be established. To begin unraveling the mystery of why spermatozoa contain mRNA, we characterized the genetic fingerprint of mature spermatozoa generated from a series of microarray experiments (G.C.O. *et al*., manuscript submitted), using Onto-Express. Transcription factors, protein binding factors, and DNA binding proteins are the functional groups containing the largest number of identified proteins (Fig. 3). The

doi:10.1006/geno.2002.6698, available online at http://www.idealibrary.com on IDEAL

identification of transcription factors was somewhat surprising, bearing in mind that spermatozoa are considered transcriptionally quiescent [21-24]. The identification of DNA and protein binding factors is consistent with spermatozoal mRNAs encapsulating spermatogenic gene expression [20,25,26]. These results illustrate that Onto-Express can transform genetic fingerprints having unknown function into profiles that underscore the biology of the system. When interpreting these observations the question often arises: "What are the identities of the various members of each functional group?" Onto-Express provides a graphical search interface to retrieve the accession or cluster identification numbers associated with specific functions (Figs. 1B and 1C). The accession and cluster identification numbers are linked to NCBI (http://www.ncbi.nlm.nih.gov/), allowing the user to effortlessly determine which genes were identified in each functional group. In this manner, the user can associate the biological functions of interest to the primary data obtained from the array. For example, the identification of transcription factors in mature spermatozoa was somewhat surprising. The specific factors (for example, UniGene Cluster Identification Numbers Hs.26102, Hs.108106, Hs.6557, Hs.147049, Hs.182528, Hs.28423, Hs.155402, Hs.79058, Hs.2815) are consistent with their functional relevancy in spermatogenesis [27]. It is clear that this tool will find many uses in our emerging studies of various transcriptomes.

MATERIALS AND METHODS

Figure 4 shows the architecture of Onto-Express. The program has been implemented in Java using an object-oriented approach to facilitate platform-independence and maintenance. The basic components of Onto-Express include the graphical user interface (GUI), a profiler, a database creator, and a local database that is updated from the NCBI server. To generate the gene expression profiles, the GUI invokes the profiler, which in turn invokes the database creator. The database creator invokes the NCBI map-viewer client and, in turn, the client repeatedly queries NCBI map-viewer to collect gene/chromosome information. When the NCBI map-viewer client returns, the database creator reads in the LL_tmpl file and creates a local database. This ensures access to the most current information. After the database creator finishes, the profiler starts reading the input file and repeatedly queries the database to collect the information needed. When the profiler has completed reading the input file and collected all of the required information, it outputs the collected information to text files. The front end GUI then reads the profiler output files and displays the assimilated data as the user requests.

ACKNOWLEDGMENTS

This work was supported in part by NIH grant HD36512 awarded to S.A.K. G.C.O. is supported in part by Wayne State University School of Medicine Dean's Post-Doctoral Fellowship.

RECEIVED FOR PUBLICATION OCTOBER 30;
ACCEPTED DECEMBER 12, 2001.

REFERENCES

1. Schena, M., Shalon, D., Davis, R., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:** 467–470.
2. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95:** 14863–14868.
3. Lockhart, D. J., *et al*. (1996). DNA expression monitoring by hybridization of high density oligonucleotide arrays. *Nat. Biotechnol.* **14:** 1675–1680.
4. Shalon, D., Smith, S. J., and Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*. **6:** 639–645.
5. Ramsay, G. (1998). DNA chips: State-of-the-art. *Nat. Biotechnol.* **16:** 40–44.
6. Iyer, V. R., *et al*. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* **283:** 83–87.
7. Perou, C. M., *et al*. (2000). Molecular portraits of human breast tumours. *Nature* **406:** 747–752.
8. Golub, T. R., *et al*. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286:** 531–537.
9. Tamayo, P., *et al*. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96:** 2907–2912.
10. Kudoh, K., *et al*. (2000). Monitoring the expression profiles of doxorubicin-induced and doxorubicin-resistant cancer cells by cDNA microarray. *Cancer Res.* **60:** 4161–4166.
11. Baxevanis, A. D. (2000). The molecular biology database collection: an online compilation of relevant database resources. *Nucleic Acid Res*. **28:** 1–7.
12. Ashburner, M., *et al*. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.
13. Consortium, T. G. (2001). Creating the gene ontology resource: design and implementation. *Genome Res*. **8:** 1425–1433.
14. Concha, I., *et al*. (1993). U1 and U2 snRNA are localized in the sperm nucleus. *Exp. Cell Res*. **204:** 378–338.
15. Kumar, G., Patel, D., and Naz, R. K. (1993). c-MYC mRNA is present in human sperm cells. *Cell. Mol. Biol. Res.* **39:** 111–117.
16. Chiang, M. H., Steuerwald, N., Lambert, H., Main, E. K., and Steinleitner, A. (1994). Detection of human leukocyte antigen class I messenger ribonucleic acid transcripts in human spermatozoa via reverse transcription-polymerase chain reaction. *Fertil. Steril.* **61:** 276–280.
17. Rohwedder, A., Liedigk, O., Schaller, J., Glander, H., and Werchau, H. (1996). Detection of mRNA transcripts of β1 integrins in ejaculated human spermatozoa by nested reverse-transcrition polymerase chain reaction. *Mol. Hum. Reprod.* **2:** 499–505.
18. Miller, D., *et al*. (1999). A complex population of RNAs in human ejaculate spermatozoa: implications for understanding molecular aspects of spermiogenesis. *Gene* **237:** 385–392.
19. Richter, W., Dettmer, D., and Glander, H. J. (1999). Detection of mRNA transcripts of cyclic nucleotide phosphodiesterase subtypes in ejaculated spermatozoa. *Mol. Hum. Reprod.* **5:** 732–736.
20. Wykes, S. M., Miller, D., and Krawetz, S. A. (2000). Mammalian spermatozoal mRNAs: tools for the functional analysis of male gametes. *J. Submicrosc. Cytol. Pathol.* **32:** 77–81.
21. Kierszenbaum, A. l., and Tres, L. L. (1975). Structural and transcriptional features of the mouse spermatid genome. *J. Cell Biol*. **65:** 258–270.
22. Stewart, T. A., Bellve, A. R., and Leder, P. (1984). Transcription and promoter usage of the myc gene in normal somatic and spermatogenic cells. *Science* **226:** 202–210.
23. Clermont, Y., and Leblond, C. (1955). Spermiogenesis of man, monkey, ram and other mammals as shown by the periodic acid-schiff technique. *Am. J. Anat.* **96:** 229–253.
24. Balhorn, R., *et al*. (1999). Protamine mediated condensation of DNA in mammalian sperm. In *The Male Gamete: From Basic Science to Clinical Applications* (C. Gagnon, Ed.), pp. 55–70. Cache River Press, Vienna, Il.
25. Kramer, J., and Krawetz, S. A. (1997). RNA in sperm: implications within the sperm genome. *J. Biol. Chem*. **271:** 11619–11622.
26. Miller, D. (1997). RNA in the ejaculate spermatozoon: a window into molecular events in spermatogenesis and a record of the unusual requirements of haploid gene expression and post-meiotic equilibration. *Mol. Hum. Reprod.* **3:** 669–676.
27. Nelson, J. E., and Krawetz, S. A. (1995). Computer assisted promoter analysis of a human sperm specific nucleoprotein gene cluster. *DNA Seq*. **5:** 329–337.