Nucleic Acids Research, 2003, Vol. 31, No. 1 **219–223**
DOI: 10.1093/nar/gkg014

SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data

Maximilian Diehn[1,*], Gavin Sherlock[2], Gail Binkley[2], Heng Jin[2], John C. Matese[2], Tina Hernandez-Boussard[2], Christian A. Rees[2], J. Michael Cherry[2], David Botstein[2], Patrick O. Brown[1,3] and Ash A. Alizadeh[1]

[1]Department of Biochemistry, [2]Department of Genetics and [3]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305, USA

Received August 14, 2002; Accepted August 23, 2002

ABSTRACT

The explosion in the number of functional genomic datasets generated with tools such as DNA microarrays has created a critical need for resources that facilitate the interpretation of large-scale biological data. SOURCE is a web-based database that brings together information from a broad range of resources, and provides it in manner particularly useful for genome-scale analyses. SOURCE's GeneReports include aliases, chromosomal location, functional descriptions, GeneOntology annotations, gene expression data, and links to external databases. We curate published microarray gene expression datasets and allow users to rapidly identify sets of co-regulated genes across a variety of tissues and a large number of conditions using a simple and intuitive interface. SOURCE provides content both in gene and cDNA clone-centric pages, and thus simplifies analysis of datasets generated using cDNA microarrays. SOURCE is continuously updated and contains the most recent and accurate information available for human, mouse, and rat genes. By allowing dynamic linking to individual gene or clone reports, SOURCE facilitates browsing of large genomic datasets. Finally, SOURCEs batch interface allows rapid extraction of data for thousands of genes or clones at once and thus facilitates statistical analyses such as assessing the enrichment of functional attributes within clusters of genes. SOURCE is available at http://source.stanford.edu.

INTRODUCTION

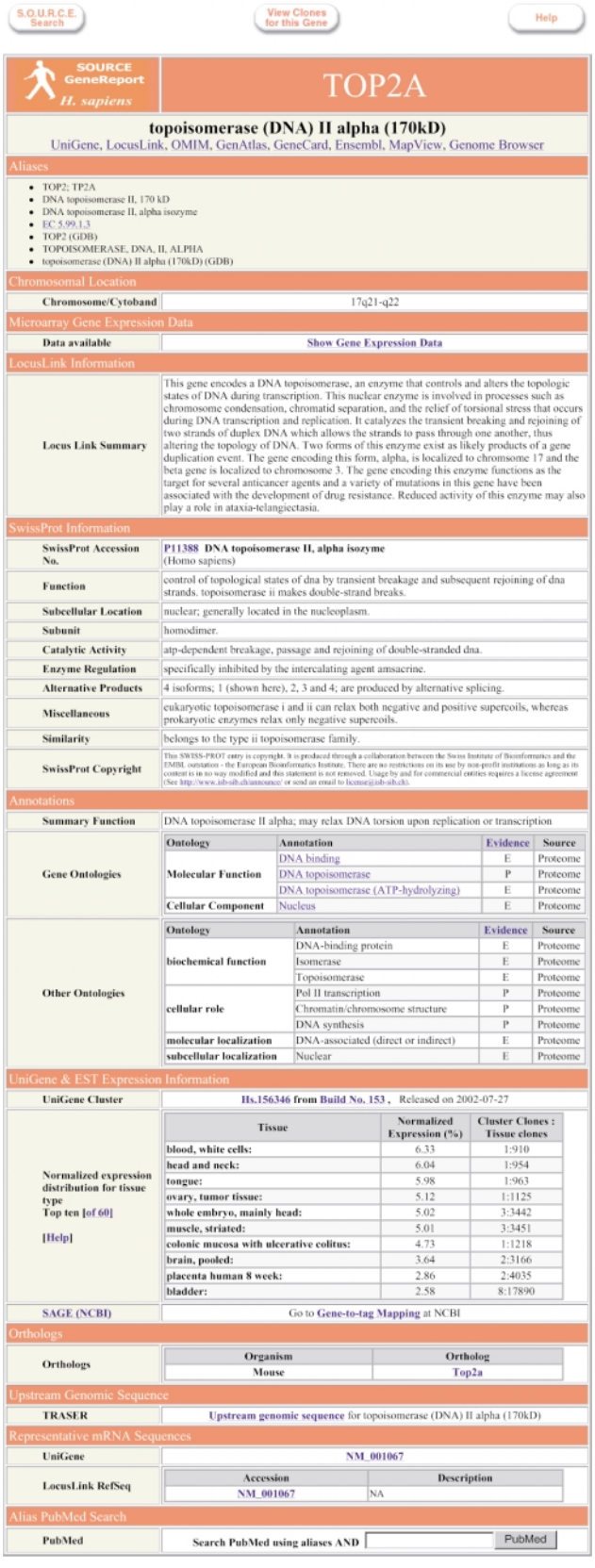
The recent emergence of high throughput structural and functional genomic technologies has led to the rapid growth of genome-scale datasets. The analysis of such datasets largely depends on rapid access to previously described features of the genes being studied. Today, diverse publicly available resources exist that catalog various attributes of genes, ranging from their mapped coordinates within the genome to the enzymatic function of the proteins they encode. These include Online Mendelian Inheritance in Man (OMIM) (1), SwissProt (2), LocusLink (3), UniGene (3), GenBank (4), PubMed (3), as well as many others. Although these resources are highly informative individually, the collection of available content would have more utility if provided in a unified and centralized context and indexed in a robust manner.

Accordingly, we have developed a publicly available, web-based resource called SOURCE (http://source.stanford.edu). Unifying data from a broad collection of resources, SOURCE is a database providing dynamic content including genomic map position, biological role, and gene expression data. Currently, this content is available for three organisms (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*), with a number of others slated for addition in the near future. We have designed SOURCE particularly for the analysis of microarray gene expression datasets and have thus emphasized the types of information that are most useful in analyzing and interpreting genome scale gene expression experiments.

DATABASE ORGANIZATION

SOURCE is structured as a set of relationships between two entities: GeneReports and CloneReports. As the name implies, a GeneReport page captures the collection of features attributable to a given gene and its products, where a gene is defined by a unique UniGene cluster. SOURCE contains GeneReports for both characterized and uncharacterized genes. GeneReports for named genes are titled with Human Gene Nomenclature Committee (5) approved conventions for naming genes, as represented within LocusLink, while GeneReports for uncharacterized genes are listed by their UniGene titles. Wherever available, each GeneReport will contain all or a subset of the following categories of data (Fig. 1):

*To whom correspondence should be addressed. Tel: +1 650 498 5998; Fax: +1 650 724 7554; Email: diehn@genome.stanford.edu
Correspondence may also be addressed to Ash A. Alizadeh. Tel: +1 650 498 5998; Fax: +1 650 724 7554; Email: arasha@genome.stanford.edu
The authors wish it to be known that, in their opinion, M.D. and A.A.A. should be regarded as joint First Authors



S.O.U.R.C.E. Search | View Clones for this Gene | Help

SOURCE GeneReport *H. sapiens*

TOP2A

topoisomerase (DNA) II alpha (170kD)

UniGene, LocusLink, OMIM, GenAtlas, GeneCard, Ensembl, MapView, Genome Browser

Aliases

- TOP2; TP2A
- DNA topoisomerase II, 170 kD
- DNA topoisomerase II, alpha isozyme
- EC 5.99.1.3
- TOP2 (GDB)
- TOPOISOMERASE, DNA, II, ALPHA
- topoisomerase (DNA) II alpha (170kD) (GDB)

Chromosomal Location

Chromosome/Cytoband	17q21-q22

Microarray Gene Expression Data

Data available	Show Gene Expression Data

LocusLink Information

Locus Link Summary	This gene encodes a DNA topoisomerase, an enzyme that controls and alters the topologic states of DNA during transcription. This nuclear enzyme is involved in processes such as chromosome condensation, chromatid separation, and the relief of torsional stress that occurs during DNA transcription and replication. It catalyzes the transient breaking and rejoining of two strands of duplex DNA which allows the strands to pass through one another, thus altering the topology of DNA. Two forms of this enzyme exist as likely products of a gene duplication event. The gene encoding this form, alpha, is localized to chromsome 17 and the beta gene is localized to chromosome 3. The gene encoding this enzyme functions as the target for several anticancer agents and a variety of mutations in this gene have been associated with the development of drug resistance. Reduced activity of this enzyme may also play a role in ataxia-telangiectasia.

SwissProt Information

SwissProt Accession No.	P11388 DNA topoisomerase II, alpha isozyme (Homo sapiens)
Function	control of topological states of dna by transient breakage and subsequent rejoining of dna strands. topoisomerase ii makes double-strand breaks.
Subcellular Location	nuclear; generally located in the nucleoplasm.
Subunit	homodimer.
Catalytic Activity	atp-dependent breakage, passage and rejoining of double-stranded dna.
Enzyme Regulation	specifically inhibited by the intercalating agent amsacrine.
Alternative Products	4 isoforms; 1 (shown here), 2, 3 and 4; are produced by alternative splicing.
Miscellaneous	eukaryotic topoisomerase i and ii can relax both negative and positive supercoils, whereas prokaryotic enzymes relax only negative supercoils.
Similarity	belongs to the type ii topoisomerase family.
SwissProt Copyright	This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See http://www.isb-sib.ch/announce/ or send an email to license@isb-sib.ch).

Annotations

Summary Function	DNA topoisomerase II alpha; may relax DNA torsion upon replication or transcription

Gene Ontologies

Ontology	Annotation	Evidence	Source
Molecular Function	DNA binding	E	Proteome
	DNA topoisomerase	P	Proteome
	DNA topoisomerase (ATP-hydrolyzing)	E	Proteome
Cellular Component	Nucleus	E	Proteome

Other Ontologies

Ontology	Annotation	Evidence	Source
biochemical function	DNA-binding protein	E	Proteome
	Isomerase	E	Proteome
	Topoisomerase	E	Proteome
cellular role	Pol II transcription	P	Proteome
	Chromatin/chromosome structure	P	Proteome
	DNA synthesis	P	Proteome
molecular localization	DNA-associated (direct or indirect)	E	Proteome
subcellular localization	Nuclear	E	Proteome

UniGene & EST Expression Information

UniGene Cluster	Hs.156346 from Build No. 153 , Released on 2002-07-27

Normalized expression distribution for tissue type
Top ten [of 60]
[Help]

Tissue	Normalized Expression (%)	Cluster Clones : Tissue clones
blood, white cells:	6.33	1:910
head and neck:	6.04	1:954
tongue:	5.98	1:963
ovary, tumor tissue:	5.12	1:1125
whole embryo, mainly head:	5.02	3:3442
muscle, striated:	5.01	3:3451
colonic mucosa with ulcerative colitus:	4.73	1:1218
brain, pooled:	3.64	2:3166
placenta human 8 week:	2.86	2:4035
bladder:	2.58	8:17890

SAGE (NCBI)	Go to Gene-to-tag Mapping at NCBI

Orthologs

Orthologs	Organism	Ortholog
	Mouse	Top2a

Upstream Genomic Sequence

TRASER	Upstream genomic sequence for topoisomerase (DNA) II alpha (170kD)

Representative mRNA Sequences

UniGene	NM_001067

LocusLink RefSeq

Accession	Description
NM_001067	NA

Alias PubMed Search

PubMed — Search PubMed using aliases AND [ ] PubMed

1. Aliases associated with a gene, captured from OMIM, LocusLink, SwissProt, UniGene, and the Mouse Genome Database (6).
2. Gene expression data from curated DNA microarray experiments.
3. Biological roles and summary of functions curated by LocusLink and SwissProt.
4. Ontology annotations, capturing both canonical Gene Ontology annotations (i.e., biological process, molecular function, and cellular component) (7), and alternative ontologies (8).
5. Virtual Tissue Northern Blot, representing the mRNA expression of the gene through relative frequencies of Expressed Sequence Tags (ESTs) from cDNA libraries derived from various tissues.
6. Chromosome localization information, with direct links to NCBIs MapView, Ensembl (9), and UCSC genome browsers (10).
7. Direct link to SOURCE GeneReports for orthologs of mouse and human genes.
8. Direct link to TRASER, an upstream (putative promoter-containing) sequence retrieval tool for predicted human genome mRNAs.
9. Direct links to a host of publicly available resources and SOURCE CloneReports.

In addition to these data, GeneReports also include representative mRNA accessions with direct links to their NCBI GenBank records. Furthermore, each GeneReport page allows formatting of boolean PubMed literature queries using user-defined search terms and all aliases for the given gene. This allows rapid identification of previously published work relevant to each user's interests.

SOURCE CloneReports capture data available for all human, mouse, and rat ESTs within dbEST (11) for which a physical clone has been annotated, regardless of association with a UniGene cluster. Each CloneReport contains a subset of the data from the dbEST record(s) of the cDNA clone, including the putative identity of its EST sequences, as well as links to the corresponding GeneReport and dbEST. When multiple EST sequences are available for a given clone, information for both 5′ and 3′ sequencing reads are displayed. Furthermore, CloneReports contain direct hyperlinks to BLAST searches of databases including the non-redundant nucleotide section of GenBank, dbEST, and high-throughput genome sequences.

Since many of the resources on which SOURCE is based (including UniGene, LocusLink, and SwissProt) are frequently

Figure 1. SOURCE GeneReport for topoisomerase II alpha (TOP2A). This screenshot depicts an example human GeneReport. Data included for this particular gene include a link to SOURCE CloneReports for ESTs mapping to TOP2A, links to outside databases such as LocusLink and the UCSC Genome Browser, aliases, chromosomal location, a link to SOURCE's microarray gene expression data, the LocusLink descriptive summary, SwissProt functional information, GeneOntology annotations, virtual northern EST expression data, a link to the SOURCE GeneReport for the mouse ortholog of TOP2A, a link to TRASER for upstream sequence retrieval, representative GenBank mRNA accession numbers, and a form for formatting boolean PubMed queries using all of TOP2A's aliases.

updated, the SOURCE database is re-loaded on a weekly basis to ensure that it contains the most up-to-date information. An automated process checks for updates of the various outside databases, downloads these files, and populates database tables accordingly. In this fashion, we ensure that the connections between external databases which are made within SOURCE are as accurate as possible. This means that both the mapping of clones to genes and the functional attributes associated with those genes is dynamic and thus current.

Currently, SOURCE employs Oracle Server Enterprise Edition version 8.1.7 and runs on an eight processor Sun E4500 under SunOS 5.8. Most of SOURCE's analysis and display software was written in Perl. The table structure for SOURCE can be found at http://genome-www.stanford.edu/microarray/doc/external2.pdf.

GENE EXPRESSION DATA

An integral mission of SOURCE is to curate and consolidate gene expression data from microarray experiments in order to allow researchers easy and intuitive access to this rapidly growing body of information. While many authors of microarray datasets have made their raw data available on their own websites, accessing these one at a time is tedious and hinders rapid analysis. This is particularly important for researchers generating their own microarray datasets, for whom the examination of co-regulated genes under diverse conditions is critical to successful analyses. While several efforts exist for centrally archiving raw microarray data [e.g., Gene Expression Omnibus (12)], these databases do not re-analyze published data nor do they provide them in a format that is readily searchable at the single gene level. For SOURCE, only datasets for which raw data have been made publicly available are considered for inclusion and these are then curated and re-analyzed in order to ensure proper data processing and display. Currently, SOURCE contains 10 human and 2 mouse microarray datasets, generated using either cDNA or Affymetrix microarrays, and totaling greater than four million gene expression measurements.

Figure 2A shows the SOURCE display for the gene expression of DNA topoisomerase II alpha (TOP2A) across the cell cycle of HeLa cells (13). The measurements are displayed as a temporally ordered matrix of gene expression data where rows represent genes (i.e., unique cDNA elements) and columns represent experimental samples. Coloured pixels capture the magnitude of the response for any gene. Shades of red and green represent induction and repression, respectively.

An important component of SOURCE's gene expression interface is the ability to list the most highly correlated genes of a given gene through a simple click on that gene's expression 'color bar.' This allows rapid identification of co-regulated groups of genes and facilitates quick access to information that is crucial to the interpretation of new microarray experiments. Figure 2B shows TOP2A's 10 most correlated neighbours in a dataset of normal tissues and cell lines (14). As can be seen, TOP2A expression is highest in transformed cell lines, normal testis and fetal liver. Additionally, many of the neighbours are genes known to be involved in cell proliferation (e.g., CDC2, CCNB2, and MAD2L1), consistent with TOP2A's role in cell cycle progression.

SOURCE also displays *in silico* generated expression information calculated from EST abundance data. In the absence of useful systematic genome-scale expression data, the EST data provide an accessible source of information that identifies at least some of the sites where a gene is expressed. For example, SNAP25, a synaptic vesicle associated protein specific to neurons (15), is highly overrepresented in EST libraries stemming from central nervous system samples compared to all other tissues (Fig. 2C). Such information is often useful when examining microarray expression data of cellular mixtures, as is the case with tissue and tumor samples.

DATA ACCESS

SOURCE allows users to query individual genes as well as retrieve selected attributes for many genes in batch. When searching for individual genes, users can query the database via a gene's name (whether the official HGNC name or a historical alias), the LocusLink identifier, the current UniGene cluster identifier, the GenBank accession of a sequence associated with the gene through UniGene, or a cDNA clone identifier. The flexibility of this search interface is important, since users may have access to only a few of these attributes for the genes they are studying. In order to increase the likelihood of successful gene name searches, we have assembled the largest collection of gene aliases available on the web by combining synonym data from a large number of sources.

The capacity to access gene-level data through searches using clone identifiers is particularly practical for users of DNA microarrays, as most spotted array platforms employ cDNA clones, each of which may be represented by multiple ESTs. In this fashion, SOURCE can reveal potentially chimeric cDNA clones, which are associated with ESTs that map to multiple UniGene clusters or genes. Currently, no other publicly available database offers this search functionality for accessing both gene- and clone-level data.

SOURCE allows for dynamic linking to both GeneReports and CloneReports. This feature is particularly useful when browsing large data sets. For example, when visualizing datasets with TreeView (16), linking of the gene or clone names to SOURCE allows users to find detailed information about each gene or clone with just a click. Similarly, external websites, such as supplements to published functional genomic datasets (e.g., see http://genome-www.stanford.edu/hostresponse/) are made much more generally useful by linking of each gene or clone name to SOURCE.

One of the most important and unique features of SOURCE is the ability to simultaneously extract data for thousands of genes in batch, thus eliminating the need for laborious cross-referencing of data from external databases. This is particularly useful for functional genomic studies, where it is necessary to continually update information on the genes and clones being examined. For instance, researchers interested in the mapped position or subcellular localization of a list of genes can extract these attributes with ease, and perform statistical analyses such as assessing the enrichment of certain functional attributes

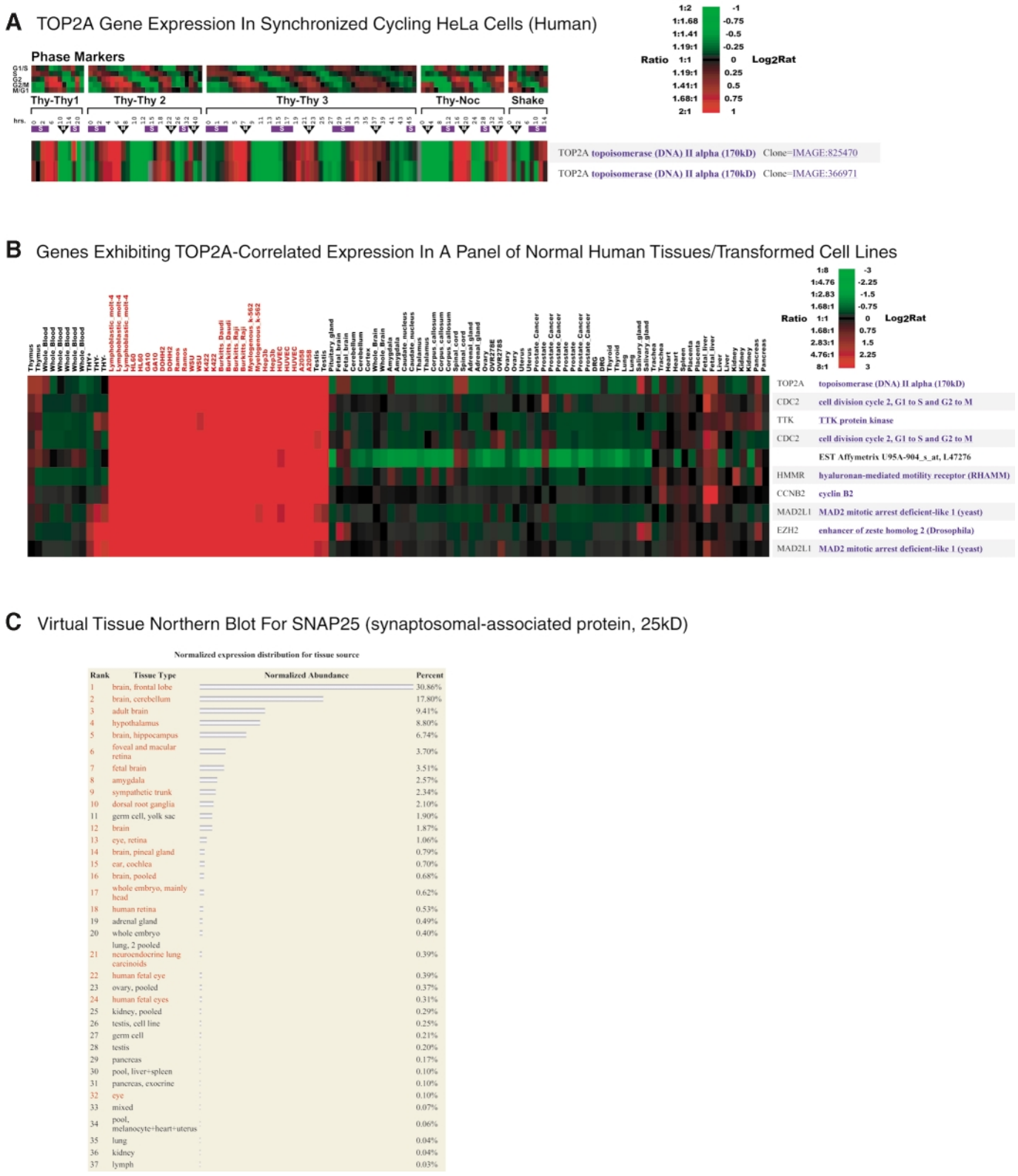


Figure 2. SOURCE gene expression tools. (**A**) SOURCE microarray data display for TOP2A's expression across the cell cycle of HeLa cells. The measurements are displayed as a temporally ordered matrix of gene expression data where rows represent genes (unique cDNA elements) and columns represent experimental samples. Colored pixels capture the magnitude of the response for any gene. Shades of red and green represent induction and repression, respectively. (**B**) Most highly correlated gene expression neighbours of TOP2A in a dataset of normal human tissues and cell lines. This figure only depicts the top 9 of the 47 neighbours with a minimum Pearson correlation coefficients 0.4. (**C**) Virtual Tissue Northern Blot for SNAP25 (synaptosomal-associated protein, 25 kD). Relative expression of SNAP25 across a variety of tissues was calculated using EST abundance data. Libraries stemming from neuron-enriched or -related libraries are highlighted in red for emphasis.

within clusters of genes (17,18). Since the data in SOURCE are refreshed weekly, users can also use this utility to regularly update annotations associated with genes or cDNA clones of interest. Input can be via a text file uploaded to the server or by pasting the queries into a text box. Batch SOURCE can be searched by clone identifier, accession number, gene name, gene symbol, UniGene identifier, or LocusLink identifier. Retrieval options include gene name, aliases, LocusLink ID, chromosome location, subcellular localization, representative accessions (protein or mRNA) and Gene Ontology annotations.

Use of SOURCE has steadily grown over the past two years. Today, thousands of researchers query the system on a daily basis, totaling over 100 000 hits per month. Individual GeneReports make up the majority of accesses, with the gene expression browser and the batch retrieval utility being extremely popular as well. Reciprocal links now exist to and from a number of databases, including SwissProt, GeneCards, and the UCSC Genome Browser.

FUTURE DIRECTIONS

We plan to continue to add new features to SOURCE, including more gene expression data sets as they are published and other useful resources that we and others develop as the field of functional genomic analysis continues to advance. We are planning on transitioning from a purely UniGene-based mapping of clones to genes, to one based on a combination of UniGene and the genome scaffold. We are also planning on adding additional model organisms and allowing users to navigate orthologies through a simple interface. As genome-scale gene expression datasets continue to amass for these organisms, this will allow SOURCE users to rapidly identify groups of orthologs that are similarly regulated in diverse organisms. Furthermore, we are hoping to provide developers access to SOURCE through data integration tools such as BioMoby (http://www.biomoby.org/) in order to further enhance the ability of researchers to extract and manipulate data in batch. The need for central and publicly available resources which curate biological data will only continue to grow and we feel that SOURCE and resources like it will be critical in enabling biologists to efficiently analyze genome-scale datasets.

ACKNOWLEDGEMENTS

We wish to thank members of the Stanford Microarray Database and the Brown and Botstein laboratories for helpful discussions and advice. This work was supported by N.I.H. grant CA85129-04 (P.O.B. and D.B.) and National Institute of General Medical Sciences training grant GM07365 (A.A.A. and M.D.). P.O.B. is an associate investigator of the Howard Hughes Medical Institute.

REFERENCES

1. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.
2. Gasteiger,E., Jung,E. and Bairoch,A. (2001) SWISS-PROT: connecting biomolecular knowledge via a protein database. *Curr. Issues Mol. Biol.*, **3**, 47–55.
3. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. *et al.* (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.*, **30**, 13–16.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
5. Povey,S., Lovering,R., Bruford,E., Wright,M., Lush,M. and Wain,H. (2001) The HUGO Gene Nomenclature Committee (HGNC). *Hum. Genet.*, **109**, 678–680.
6. Blake,J.A., Richardson,J.E., Bult,C.J., Kadin,J.A. and Eppig,J.T. (2002) The Mouse Genome Database (MGD): the model organism database for the laboratory mouse. *Nucleic Acids Res.*, **30**, 113–115.
7. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
8. Hodges,P.E., Carrico,P.M., Hogan,J.D., O'Neill,K.E., Owen,J.J., Mangan,M., Davis,B.P., Brooks,J.E. and Garrels,J.I. (2002) Annotating the human proteome: the Human Proteome Survey Database (HumanPSD) and an in-depth target database for G protein-coupled receptors (GPCR-PD) from Incyte Genomics. *Nucleic Acids Res.*, **30**, 137–141.
9. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
10. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
11. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet.*, **4**, 332–333.
12. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
13. Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.
14. Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M., Moqrich,A. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
15. Oyler,G.A., Higgins,G.A., Hart,R.A., Battenberg,E., Billingsley,M., Bloom,F.E. and Wilson,M.C. (1989) The identification of a novel synaptosomal-associated protein, SNAP-25, differentially expressed by neuronal subpopulations. *J. Cell Biol.*, **109**, 3039–3052.
16. 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.
17. Boldrick,J.C., Alizadeh,A.A., Diehn,M., Dudoit,S., Liu,C.L., Belcher,C.E., Botstein,D., Staudt,L.M., Brown,P.O. and Relman,D.A. (2002) Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proc. Natl Acad. Sci. USA*, **99**, 972–977.
18. Diehn,M., Alizadeh,A.A., Rando,O.J., Liu,C.L., Stankunas,K., Botstein,D., Crabtree,G.R. and Brown,P.O. (2002) Genomic expression programs and the integration of the CD28 costimulatory signal in T cell activation. *Proc. Natl Acad. Sci. USA*, **99**, 11796–11801.