eVOC: A Controlled Vocabulary for Unifying Gene Expression Data

Janet Kelso,[1] Johann Visagie,[2] Gregory Theiler,[3] Alan Christoffels,[1,6] Soraya Bardien,[1] Damian Smedley,[4] Darren Otgaar,[2] Gary Greyling,[2] C. Victor Jongeneel,[3] Mark I. McCarthy,[4,5] Tania Hide,[2] and Winston Hide[1,7]

[1] South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa; [2] Electric Genetics PTY Ltd. Bellville, South Africa; [3] Office of Information Technology, Ludwig Institute for Cancer Research and Swiss Institute of Bioinformatics, Lausanne, Switzerland; [4] Genetics and Genomics Research Institute, Imperial College Faculty of Medicine, Hammersmith Hospital, London, W12 0NN, UK; [5] Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX37BN, UK

Expression data contribute significantly to the biological value of the sequenced human genome, providing extensive information about gene structure and the pattern of gene expression. ESTs, together with SAGE libraries and microarray experiment information, provide a broad and rich view of the transcriptome. However, it is difficult to perform large-scale expression mining of the data generated by these diverse experimental approaches. Not only is the data stored in disparate locations, but there is frequent ambiguity in the meaning of terms used to describe the source of the material used in the experiment. Untangling semantic differences between the data provided by different resources is therefore largely reliant on the domain knowledge of a human expert. We present here eVOC, a system which associates labelled target cDNAs for microarray experiments, or cDNA libraries and their associated transcripts with controlled terms in a set of hierarchical vocabularies. eVOC consists of four orthogonal controlled vocabularies suitable for describing the domains of human gene expression data including Anatomical System, Cell Type, Pathology and Developmental Stage. We have curated and annotated 7016 cDNA libraries represented in dbEST, as well as 104 SAGE libraries, with expression information, and provide this as an integrated, public resource that allows the linking of transcripts and libraries with expression terms. Both the vocabularies and the vocabulary-annotated libraries can be retrieved from http://www.sanbi.ac.za/evoc/. Several groups are involved in developing this resource with the aim of unifying transcript expression information.

[Supplemental material is available online at www.genome.org.]

Mining of large volumes of transcriptome data is currently frustrated by an inability to relate sequence and descriptive information. In part, this is due to the absence of a common structured vocabulary to describe the source of the biological sample materials.

Recent years have seen a growing trend toward the adoption of ontologies for the management of biological knowledge. In Computer Science, an ontology is defined as an "explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest, and the relationships that hold among them" (The Free Online Dictionary of Computing http://wombat.doc.ic.ac.uk/foldoc/foldoc.cgi?query=ontology).

Biological ontologies aim to overcome the semantic heterogeneity commonly encountered in molecular biology databases, and to provide a common terminology for the description of a focused aspect of biology. One such resource, TAMBIS (Stevens et al. 2000), implements ontologies for both bioinformatics tasks and molecular biology to provide users with transparent access to multiple heterogeneous bioinformatics resources. Other ontologies focusing on specific aspects of biology include the Gene Ontology Consortium (Ashburner et al. 2000), which provides vocabularies that can be used to describe gene products in any organism, the EcoCyc ontology (Karp et al. 2002b), which represents important metabolic and signal-transduction events in *Escherichia coli* and the MetaCyc (Karp et al. 2002a) and KEGG (Kanehisa et al. 2002) ontologies, which describe aspects of the relationships between the chemical reactants, catalysts, substrates, and products. Numerous other ontologies representing a wide array of biological phenomena exist or are under development.

Although several ontologies for the formal description of sample materials exist or are under development (Table 1), these are not suitable for querying gene expression data. For example, clinical ontologies including anatomical, pathological, and developmental stage-specific concepts have been available for some time (ICD-9-CM, SNOMED, GALEN, MeSH), but these have not been widely adopted for describing human gene expression profiles. A major reason why clinical ontologies are not widely used for describing gene expression is that they are extremely detailed and often tangled (Rector et al. 2001), with distinct concepts with varying relationship

[6] Present address: Molecular Genetics/Fugu informatics, Institute of Molecular and Cell Biology, Singapore.
[7] Corresponding author.
E-MAIL winhide@sanbi.ac.za; FAX 27-21-959-2512.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.985203.

13:1222–1230 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

Table 1. Existing Ontologies That Are Relevant to Human Expression Data

	Website	Scope
CBIL	http://www.cbil.upenn.edu/anatomy.php3	Adult anatomy
Cytomer	http://www.biobase.de/pages/products/cytomer.html	Human developmental anatomy
HUMAT	http://www.ana.ed.ac.uk/anatomy/database/humat/	Human developmental anatomy
EPOdb	http://www.cbil.upenn.edu/EpoDB/release/version_2.2/ controlled.vocab.html	Human anatomy, developmental stage, cell type
GeneX	http://www.ncgr.org/genex/	Human gene expression
MeSH	http://www.nlm.nih.gov/mesh/meshhome.html	Clinical ontology
UMLS	http://www.nlm.nih.gov/research/umls/umlsmain.html	Clinical ontology
GALEN	http://www.opengalen.org/	Clinical ontology
SNOMED	http://www.snomed.org/main.html	Clinical ontology
ICD-9-CM	http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm	Clinical ontology

types mixed together, making them unwieldy and difficult to adopt for general use. An example is the mixing of anatomical and pathological terms in ICD-9-CM, for example, benign neoplasm of the stomach. The complexity of the concepts represented by these ontologies makes them unsuitable for the computational interrogation of gene expression data to determine simple and complex expression profiles.

Implementing multiple ontologies with simple concepts in orthogonal domains provides a preferable solution, as it enables users to produce logical ontology cross-products. Cross-products are hybrid ontologies that can be constructed through the combination of simple ontologies. For example, the ICD-9-CM term mentioned above could have been constructed through the combination of terms from an anatomical and a pathological ontology by producing the cross-product of the terms "stomach" and "neoplasm | benign" from the respective ontologies.

Ideally, ontologies for gene expression should reflect a level of detail appropriate to the data being classified and the level at which queries are likely to be performed while simultaneously providing sufficient flexibility to enable regular updating without needing to significantly restructure the hierarchies.

For the extensive description of gene expression and to provide maximum flexibility in querying, we have developed eVOC—four orthogonal ontologies that aim to provide an appropriately detailed set of terms for describing the sample source of cDNA and SAGE libraries and labeled target cDNAs for microarray experiments. We have taken a data-driven approach to determining the level of granularity required.

We have annotated all publicly available human cDNA and SAGE libraries as extensively as possible. This is achieved by the assignment of terms from each of the four ontologies to the libraries. Initial assignment of terms to libraries was performed computationally, with curators who are domain experts performing assessment of annotation quality and further manual assignment. Where information was lacking in the library record, the original submitters were contacted where possible to provide more extensive information.

The most widely used ontology for keywording human SAGE and EST libraries is the CGAP/UniLib vocabulary (ftp://ftp.ncbi.nlm.nih.gov/pub/bioannot/info/keys) currently used by the National Cancer Institute to categorize libraries for CGAP (http://www.ncbi.nlm.nih.gov/CGAP/).

CGAP provides a single integrated hierarchy of keywords that includes terms from multiple classification domains (including tissues, developmental stage, library preparation, and chemical agents among others). There are many different relationships between parent and child terms in different sections of the hierarchy. eVOC, in contrast, provides completely orthogonal ontologies covering four distinct domains. There is a single implied type of relationship between the terms within each of the eVOC ontologies.

The structure of the CGAP ontology enables rapid keyword searching, whereas the eVOC data structure, by incorporating the rigorous separation of classification terms into orthogonal domains and the formalization of relationships between terms, allows for a degree of computer reasoning to be applied. This facilitates a wide range of query types. For example, a comparison of eVOC and UniLib querying shows clearly that both eVOC and UniLib allow querying for multiple terms combined with "AND" (the intersection set), and yield comparable results in terms of the libraries returned. However, UniLib is unable to support more complex queries incorporating "OR" and "NOT", which are possible with eVOC. eVOC therefore provides users with greater flexibility, as more complex biological queries can be formulated. Whereas this may be a simple implementation issue, it is one that directly affects the user interaction with the data.

A major distinction between CGAP and eVOC is that the CGAP hierarchy is cancer specific by design. The terms included are therefore those of interest in cancer, whereas eVOC is designed for more general application. Specifically, CGAP lacks the comprehensive pathology terminology that is necessary for a broadly applicable human expression ontology.

METHODS AND DISCUSSION

The design and creation of the expression ontologies is distinct from the annotation of cDNA and SAGE libraries by use of each of the ontologies. These processes will be discussed separately.

Development of a Data Structure for Expression Ontologies

The expression ontologies have been developed in four orthogonal (mutually exclusive) knowledge domains including Anatomical System, Cell Type, Developmental Stage, and Pathology. Anatomical System and Cell Type describe where a gene is expressed, Developmental Stage describes the timing of gene expression during development, and Pathology describes the disease state in which the gene is expressed. These four ontologies were found to represent the vast majority of the expression data currently under classification. The addition of further ontologies may be appropriate in the future.

The expression ontologies are independent pure hierarchies (or trees). In a pure hierarchy, each node has only one parent but may have multiple children. Each node is associated with a specific concept in the knowledge domain represented by the hierarchy through the association of each node with one or more synonymous terms. For example, the terms "nasal" and "nose" are synonyms attached to a single node in the ANATOMY ontology.

In these pure hierarchies, there is only a single type of relationship between the nodes in each hierarchy, although the nature of the relationship is not defined explicitly. For each ontology, the nature of the expression domain imposes an implicit type on the relationship between the nodes. For instance, in the Anatomical System ontology, the relationships are of the "part-of" type. In the Cell Type and Pathology ontologies, they are of the "subclass" type, and in the Developmental Stage ontology, the relationships are of the "is-a" variety.

Pure hierarchies have a number of advantages over the more complex data structures often used to represent ontologies (Rector et al. 2001). They are easy to maintain and expand and can be visualized easily. Moreover, it is possible to construct a simple, yet extremely powerful and flexible mechanism to query data across multiple hierarchies.

In cases in which terms appear to have more than one parent, two options are available: migration to a directed acyclic graph (DAG), or untangling of the hierarchy to yield a pure hierarchy (Fig. 1). To handle multi-parent terms and different parent-child relationships, the GO project (Gene Ontology Consortium 2001) has implemented a DAG structure. During the development of the eVOC ontologies, and on the basis of available cDNA and SAGE libraries, we have found that where it appears there is a need to represent multiple relationship types in one hierarchy, it is possible to untangle the hierarchy further by splitting it into separate hierarchies with more narrowly defined relationship types.

The disadvantage of maintaining untangled orthogonal ontologies is that the volume of work involved in curation increases linearly with the number of hierarchies. It is therefore necessary to strike a balance between keeping the number of ontologies manageable, and representing relationships in as fine grained a fashion as possible. The sort of queries the ontologies are required to accommodate dictates where this balance is found. In other words, the ontology design should be data driven.

Each of the terms in the ontologies has a numeric identifier that uniquely identifies the term and that can be used as an unambiguous database cross-reference. Definitions of each of the terms are to be provided as part of the ongoing development. The source of each definition will be made available, along with the definition.

Development of the Four Expression Ontologies

The four expression ontologies (Fig. 2) currently implemented are shown below.

Anatomical System Ontology

The Anatomical System ontology provides a controlled vocabulary for the description of the anatomical system or organ in which a gene is expressed. It is based on the controlled vocabulary used in the Computational Biology and Informatics Laboratory's (CBIL) databases (www.cbil.upenn.edu/anatomy.php3), but with modifications including the removal of all references to tissue type, cell type, or developmental stage. Organization of the Anatomical System hierarchies is currently systems based. Examples of broad Anatomical Systems are digestive system or nervous system, with more specific anatomical terms within these systems being pancreatic islets or retina. Future developments to eVOC will include the creation of an Anatomical Site ontology which will extend the current Anatomical System ontology by dividing anatomical parts according to their spatial position, rather than according to the system to which they belong. This is of particular value in describing libraries from spatially distinct anatomical sites containing multi-system anatomical sites. For example, "head" is a distinct anatomical site, but includes both nervous and circulatory systems. The Anatomical System ontology contains 372 terms.

Cell Type Ontology

The Cell Type ontology provides a fine-grained description of where a gene is expressed. It is a listing of human cell types extracted from Gray's Anatomy (Gray et al. 1995). The Cell Type ontology includes 153 different cell types.

Because various cell types are represented across many anatomical systems, cell types could have

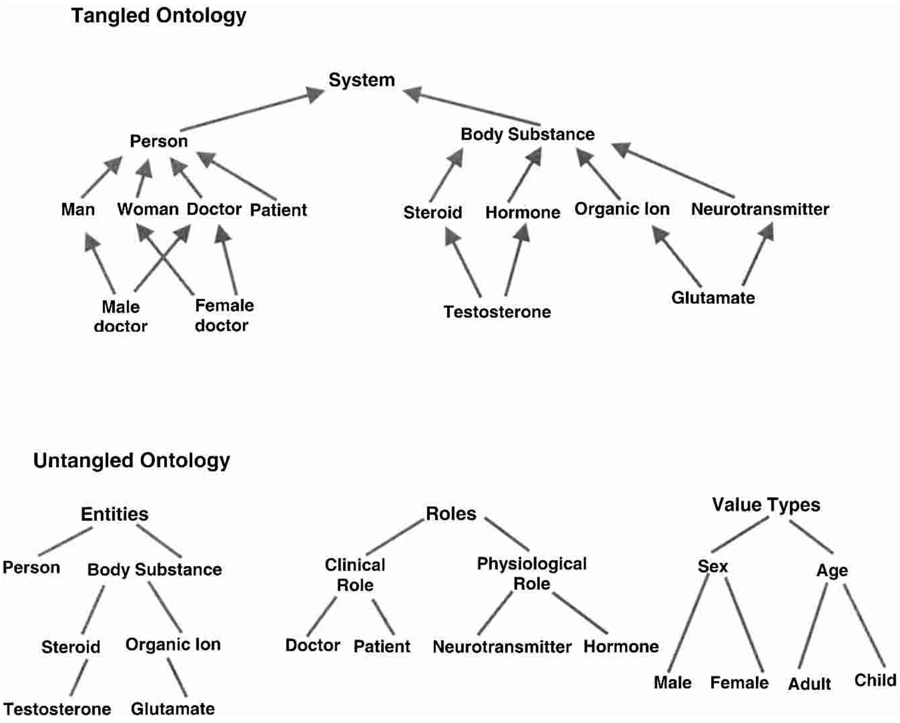
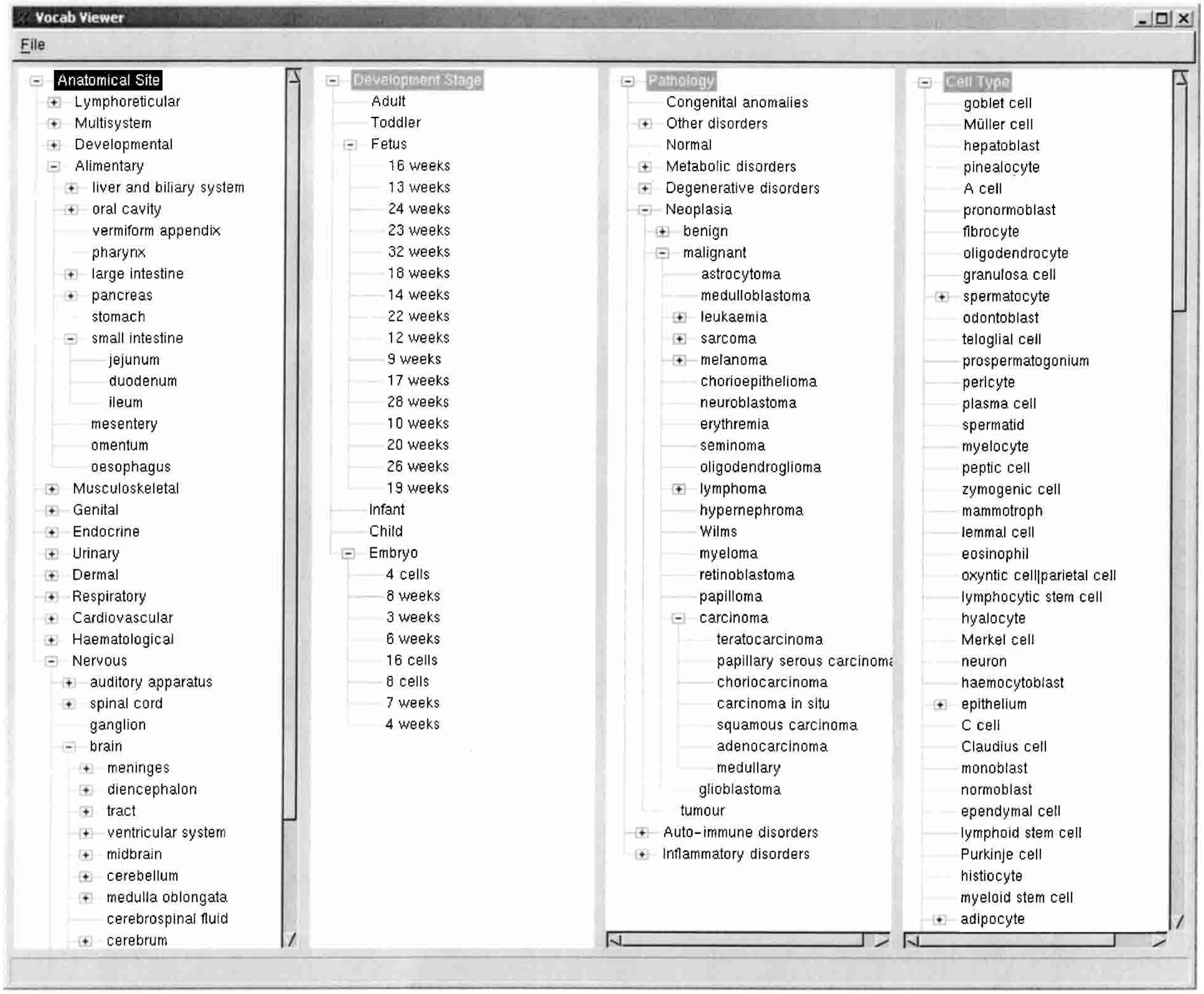


Figure 1 Untangling a tangled ontology (modified from Kemp and Gray [2002]) A complex mixed ontology can be simplified by creating simpler ontologies representing distinct domains.



- ⊟ Anatomical Site
  - ⊞ Lymphoreticular
  - ⊞ Multisystem
  - ⊞ Developmental
  - ⊟ Alimentary
    - ⊞ liver and biliary system
    - ⊞ oral cavity
    - vermiform appendix
    - pharynx
    - ⊞ large intestine
    - ⊞ pancreas
    - stomach
    - ⊟ small intestine
      - jejunum
      - duodenum
      - ileum
    - mesentery
    - omentum
    - oesophagus
  - ⊞ Musculoskeletal
  - ⊞ Genital
  - ⊞ Endocrine
  - ⊞ Urinary
  - ⊞ Dermal
  - ⊞ Respiratory
  - ⊞ Cardiovascular
  - ⊞ Haematological
  - ⊟ Nervous
    - ⊞ auditory apparatus
    - ⊞ spinal cord
    - ganglion
    - ⊟ brain
      - ⊞ meninges
      - ⊞ diencephalon
      - ⊞ tract
      - ⊞ ventricular system
      - ⊞ midbrain
      - ⊞ cerebellum
      - ⊞ medulla oblongata
      - cerebrospinal fluid
      - ⊞ cerebrum

- ⊟ Development Stage
  - Adult
  - Toddler
  - ⊟ Fetus
    - 16 weeks
    - 13 weeks
    - 24 weeks
    - 23 weeks
    - 32 weeks
    - 18 weeks
    - 14 weeks
    - 22 weeks
    - 12 weeks
    - 9 weeks
    - 17 weeks
    - 28 weeks
    - 10 weeks
    - 20 weeks
    - 26 weeks
    - 19 weeks
  - Infant
  - Child
  - ⊟ Embryo
    - 4 cells
    - 8 weeks
    - 3 weeks
    - 6 weeks
    - 16 cells
    - 8 cells
    - 7 weeks
    - 4 weeks

- ⊟ Pathology
  - Congenital anomalies
  - ⊞ Other disorders
  - Normal
  - ⊞ Metabolic disorders
  - ⊞ Degenerative disorders
  - ⊟ Neoplasia
    - ⊞ benign
    - ⊟ malignant
      - astrocytoma
      - medulloblastoma
      - ⊞ leukaemia
      - ⊞ sarcoma
      - ⊞ melanoma
      - chorioepithelioma
      - neuroblastoma
      - erythremia
      - seminoma
      - oligodendroglioma
      - ⊞ lymphoma
      - hypernephroma
      - Wilms
      - myeloma
      - retinoblastoma
      - papilloma
      - ⊟ carcinoma
        - teratocarcinoma
        - papillary serous carcinoma
        - choriocarcinoma
        - carcinoma in situ
        - squamous carcinoma
        - adenocarcinoma
        - medullary
      - glioblastoma
    - tumour
  - ⊞ Auto-immune disorders
  - ⊞ Inflammatory disorders

- ⊟ Cell Type
  - goblet cell
  - Müller cell
  - hepatoblast
  - pinealocyte
  - A cell
  - pronormoblast
  - fibrocyte
  - oligodendrocyte
  - granulosa cell
  - ⊞ spermatocyte
  - odontoblast
  - teloglial cell
  - prospermatogonium
  - pericyte
  - plasma cell
  - spermatid
  - myelocyte
  - peptic cell
  - zymogenic cell
  - mammotroph
  - lemmal cell
  - eosinophil
  - oxyntic cell|parietal cell
  - lymphocytic stem cell
  - hyalocyte
  - Merkel cell
  - neuron
  - haemocytoblast
  - ⊞ epithelium
  - C cell
  - Claudius cell
  - monoblast
  - normoblast
  - ependymal cell
  - lymphoid stem cell
  - Purkinje cell
  - histiocyte
  - myeloid stem cell
  - ⊞ adipocyte

Figure 2 A screenshot of the four ontologies. Anatomical System, Developmental Stage, Pathology, and Cell Type hierarchies are displayed with indications where the tree can be expanded.

been included in the Anatomical Site ontology, with cell type terms having multiple parents. Instead, we have separated the Anatomical System and Cell Type ontologies in order to maintain pure trees. This separation provides users with greater flexibility, as they can query on specific cell types, regardless of the anatomical location, and can also perform combined queries across Cell Type and Anatomical System terms to yield results for a cell type in a specified location.

Developmental Stage Ontology

The Developmental Stage ontology provides an ordered timeline of human development for the description of gene expression in temporal space. Examples of terms in the current hierarchy include embryo and adult. Embryogenesis is further divided into the standard Carnegie stages (www.ana.ed.ac.uk/anatomy/database/humat/), which define the first two months of human development. Each of the major stages of development is further divided into appropriate weekly and yearly categories (Supplemental Table 1C). The Developmental Stage ontology contains 132 distinct terms.

Pathology Ontology

The Pathology ontology is loosely based on the World Health Organisation's ICD-9-CM (www.mcis.duke.edu/standards/termcode/icd9/1tabular.html). ICD-9-CM is designed for the classification of morbidity and mortality information for statistical purposes and for the indexing of hospital records by disease and surgical operations. We have implemented a modified version of the first two levels of this hierarchy, and have incorporated terms that are used widely in sample descriptions, but which are not present in ICD-9-CM, for example, Wilm's tumor. We have also removed terms that refer to systems, organs, tissues, and cell types, as these are already included in the Anatomical System and Cell Type ontologies. The Pathology ontology contains 141 terms.

Species-Specific Considerations

The broad domains covered by eVOC's four orthogonal hierarchies are sufficiently generic to be applicable to a wide and diverse variety of eukaryotic organisms. However, given that each organism has unique tissue organization, development, and disease processes, organism-specific ontologies are appropriate for expression data. For instance, an extensive mouse-specific expression ontology, the Mouse Anatomical Dictionary, has been collaboratively developed by the Jackson Laboratories and the Edinburgh Mouse Atlas project (http://www.informatics.jax.org/searches/anatdict_form.shtml).

There is significant value in being able to identify and relate equivalent tissues in different species, and to compare gene expression patterns in these tissues. Although it is not clear that it will always be possible to identify these equivalent tissues in the model organisms, the production of species-specific ontologies to form the basis of these comparisons is the first step. To facilitate interoperability between species-specific ontologies, these need to be in a compatible, accessible format (Bard and Winter 2001). The eVOC human expression ontologies are provided in a format that promotes easy adoption and that will facilitate the interrogation of cross-species ontologies from different sources.

Curation of the eVOC Ontologies

We maintain a central, versioned database of eVOC ontologies that are updated, modified, and released publicly, by domain experts on an ongoing basis. The curators have the ability to add or delete terms and synonyms and to make changes to the hierarchies.

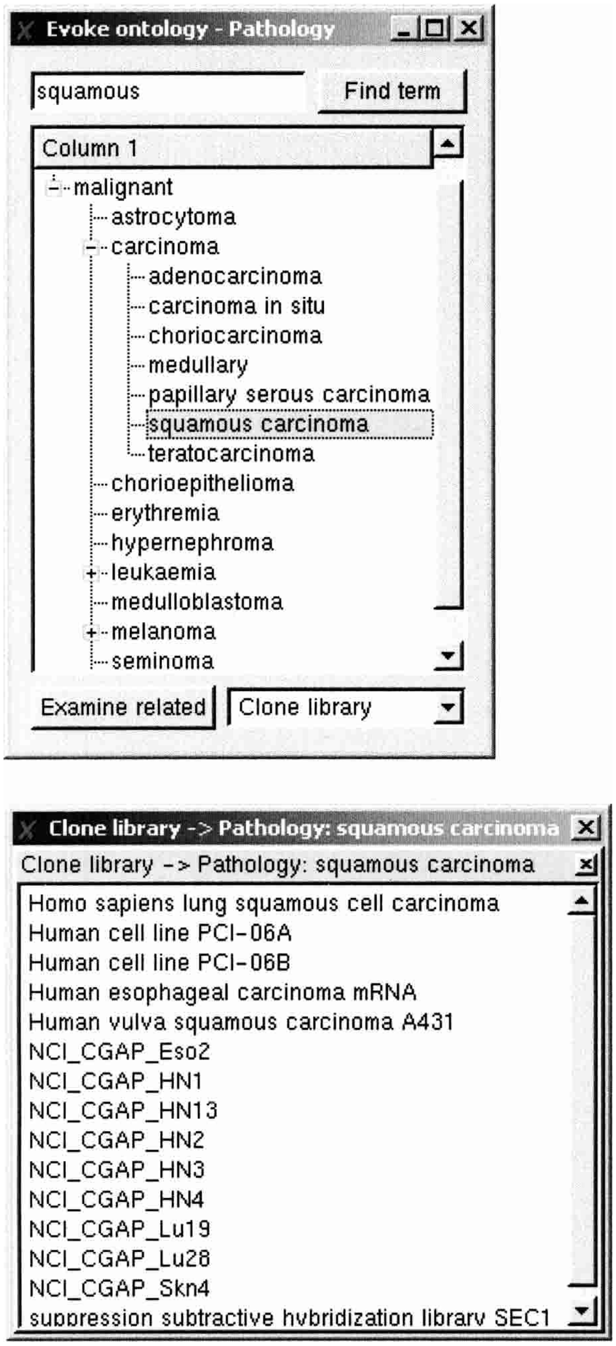
Groups that choose to modify the ontologies for their own purposes are encouraged to contribute their modifications and corrections to the curators for inclusion. A mailing list, evoc@sanbi.ac.za, has been established for this purpose.

Annotation of cDNA and SAGE Libraries Using eVOC

The ontologies presented here are independent of the expression data that they are used to annotate. We have already annotated publicly available cDNA and SAGE libraries using these expression ontologies; Supplemental Table 1, A–D (available from http://www.sanbi.ac.za/evoc/) provides statistics for the number of libraries and ESTs annotated with specific terms in each of the ontologies. Figure 3 provides an example of the annotation cDNA libraries in a subsection of the Pathology ontology. The eVOC ontologies are also highly appropriate for the annotation of labeled target cDNAs for microarray experiments.

cDNA and SAGE libraries are collections of the transcribed sequences expressed in the biological sample material from which the library is prepared. Information about the source of the sample is stored with the library information. The amount and quality of the source information provided varies depending on the source of the library. Libraries submitted to public databases are described by use of highly inconsistent terminology. Here, curators have manually translated the unstructured terms used in the library records into standardized terms selected from the four ontology domains, and have applied these to each of the libraries. Ideally, an ontology-based form would guide submitters in selecting appropriate terms for the description of their libraries. This would reduce the curation required and facilitate querying of the public databases in a manner not currently possible.

Each of the cDNA and SAGE libraries was assigned computationally to the most specific possible terms in each of the four ontologies. Manual curation and annotation of the computational assignments was then performed. Libraries are annotated with terms in each of the four hierarchies if sufficient information is available in each of the ontology domains. Annotation of a library in one ontology is completely independent of annotation in another ontology. Each annotation


Evoke ontology - Pathology
squamous
Find term
Column 1
Examine related
Clone library
Clone library -> Pathology: squamous carcinoma

- malignant
  - astrocytoma
  - carcinoma
    - adenocarcinoma
    - carcinoma in situ
    - choriocarcinoma
    - medullary
    - papillary serous carcinoma
    - squamous carcinoma
    - teratocarcinoma
  - chorioepithelioma
  - erythremia
  - hypernephroma
  - leukaemia
  - medulloblastoma
  - melanoma
  - seminoma

Clone library -> Pathology: squamous carcinoma

Homo sapiens lung squamous cell carcinoma
Human cell line PCI-06A
Human cell line PCI-06B
Human esophageal carcinoma mRNA
Human vulva squamous carcinoma A431
NCI_CGAP_Eso2
NCI_CGAP_HN1
NCI_CGAP_HN13
NCI_CGAP_HN2
NCI_CGAP_HN3
NCI_CGAP_HN4
NCI_CGAP_Lu19
NCI_CGAP_Lu28
NCI_CGAP_Skn4
suppression subtractive hybridization library SEC1

Figure 3 A screenshot of the Pathology ontology with the term "squamous cell carcinoma" selected. Selection of a term displays the libraries that are annotated with that term (squamous cell carcinoma in this case) in the lower window. Using this GUI (developed by Electric Genetics), users can view the ontologies, browse the hierarchical trees, and perform set operations on the annotated cDNA library data. The user is able to obtain the list of cDNA libraries or ESTs returned by a query, or provide a list of libraries or EST accessions and obtain the associated expression profile.

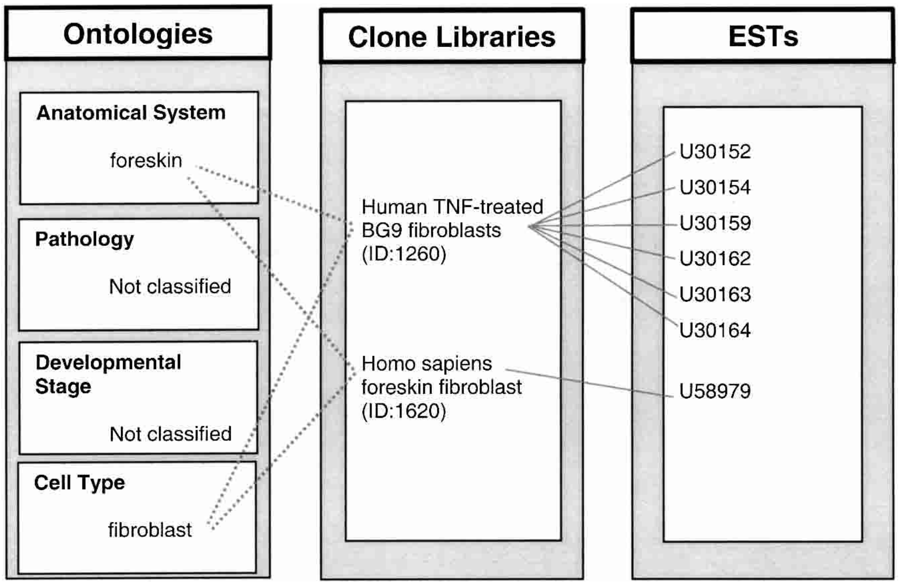


Figure 4 The four expression ontologies are used to annotate cDNA clone libraries. ESTs can be transitively associated with ontology terms via their association with a unique clone library. Clone libraries are generated from biological sample materials representing specific expression states (e.g., human foreskin fibroblasts). All of the genes/transcripts expressed in the original biological sample are captured in the clone library and can be sequenced as ESTs from the library. By mapping the clone libraries to a set of controlled terms (the ontologies), all of the ESTs from each clone library can be transitively linked to these same standardized terms in the relevant ontology via their association with their parent clone library.

is transferred from the library information provided by the original submitter. Whereas the curators exercise domain expertise in assigning libraries to specific terms within each hierarchy, they derive no new information. This process is therefore largely objective. Evidence for annotations is primarily based on the original submission record for both cDNA and SAGE libraries.

In most instances, annotation of data from existing databases is performed following the development of ontologies. Appropriate terms are assigned to data points on the basis of information already present in the database. This post-facto approach results in an often-imperfect mapping between data and terms, as much of the sample information is not provided in the original submission and is therefore lost. The Ontologies Working Group of the MGED Consortium is building ontologies for use in data submission forms for the microarray databases. This will allow subsequent database queries to take advantage of the standardized terms provided by the ontologies. The implementation of a similar ontology-based data entry system for the public nucleotide databases would be of immense value for the submission of cDNA and SAGE library information.

The clone libraries annotated here are generated from biological sample materials representing specific expression states (e.g., infant lung). These libraries represent a snapshot collection of the transcripts expressed in the original sample. The transcripts expressed in the original biological sample can therefore be sequenced as ESTs from the clone library. By mapping the clone libraries to a set of controlled terms (the ontologies), all of the ESTs from each clone library can be transitively linked to these same standardized terms in the relevant ontology via their association with their parent clone library. In the case of ESTs, we maintain a database for the bidirectional accession to clone library lookup, which in turn allows us to link vocabulary terms directly to ESTs (Fig. 4).

We have annotated 7016 human cDNA and 104 human SAGE libraries with the eVOC expression ontologies. These represent all of the human cDNA and SAGE libraries that were available publicly in April 2002. The amount of information provided for each library varies widely. In some cases, extensive information about the anatomical system, developmental stage, and pathological state of the sample source is provided, whereas in other cases, only a subset of this information is provided. The majority of the cDNA libraries (94.8%) have the information required for classification in the Anatomical System ontology, and most have information required for annotation with Pathology and Developmental Stage terms (Table 2). Where libraries were unable to be annotated, this was because the library information provided by submitters did not capture the relevant information. As a result of the fact that cDNA and SAGE libraries are largely derived from whole organs and tissues rather than from individual cell types, the majority of the libraries (94.2%) could not be annotated using the Cell Type ontology.

Using the Ontologies

Querying

Untangled hierarchies allow for the implementation of a very simple query schema. A query for a particular term returns the node with which that term is associated, as well as all of the nodes in the entire subtree (branch) rooted at that node. For instance, a query for the term neoplasia returns a particular node in the Pathology ontology, as well as all of its children, recursively. The next step in building a useful querying system lies in utilizing the mappings from nodes to public databases (for example, cDNA libraries). In this way, a query for a

Table 2. Total Number of Annotated cDNA and SAGE Libraries in Each Ontology

	Total libraries	Annotated libraries	Not annotated
Anatomical system	7120	6752	5.2%
Cell type	7120	410	94.2%
Developmental stage	7120	5891	17.3%
Pathology	7120	6401	10.1%

Most libraries can be annotated with Anatomical System terms, as these are generally present in the library record. Less information is available for Cell Type and Developmental Stages, as these are not consistently captured during the capture of library information.

particular term is translated first to a node, then expanded to a set of nodes, and then translated to a set of cDNA libraries. The set of libraries includes all of the libraries associated with all of the nodes in the branch rooted at the node that was originally associated with this node.

This simplistic query methodology can be the basis of an enormously powerful query infrastructure if the ability to perform basic set algebra (union and intersection) operations on the returned sets of cDNA libraries is used.

Consider, for instance, the query "liver AND neoplasia" (Fig. 5). A query on liver resolves to a node in the Anatomical System ontology, which in turn results in a set of cDNA libraries (all of the libraries associated with the liver node and all its subnodes). Similarly, a query on neoplasia returns the set of cDNA libraries associated with a subtree of the Pathology ontology. The combined query—"liver AND neoplasia"—returns the intersection of these two sets of cDNA libraries. In other words, it will return only libraries that were constructed from neoplastic liver samples.

Example Applications

The ontologies and the associated annotated cDNA and SAGE libraries have a wide array of applications.

By simply curating dbEST using the eVOC ontologies, users are provided with the ability to perform queries on the basis of location, state, and timing of expression on human ESTs or cDNA libraries. Querying using terms from any combination of the ontologies, both libraries and transcripts can be selected from the database on the basis of their expression patterns. Moreover, the differential expression of genes or gene isoforms on the basis of EST data can be determined swiftly and accurately by providing a list of EST accessions and analyzing the distribution of terms attached to each EST.

Laboratory-based applications of eVOC include the selection of clone libraries relevant to laboratory research projects; for example, a simple query that returns the total number of publicly available retinal cDNA libraries yields 22 results (Fig. 6). To select suitable libraries for the comparison of gene expression in adult and fetal retina, further refined queries can be used to show that 7 libraries are derived from adult retina, 3 are derived from fetal retina, and 12 libraries do not have information about the developmental stage from which the retinal tissue was isolated.

Similarly, the number of cDNA libraries available for pancreatic tissue yields 31 results. To determine how many of these are pancreatic islet libraries, a second query is performed and yields a total of 10 pancreatic islet libraries that have source descriptions as diverse as Human insulinoma and HR85 islet.

Additionally, the ability to identify cDNA and SAGE libraries from similar expression states provides access to an increased resource for data mining, and allows users to identify and analyze genes that are differentially expressed both in their expression location and their expression level. We have used the system to identify neoplastic and normal cDNA libraries, and have identified differential gene expression and alternative splicing in these expression states (H. Brentani, O.L. Caballero, A.A. Camargo, A.M. da Silva, W.A. da Silva, E. Dias Neto, M. Grivet, A. Gruber, P.E.M. Guimaraes, W. Hide, et al., in prep.).

To illustrate the power of expression ontologies in determining the tissue specificity of alternatively spliced transcripts, we have analyzed the data produced by Xu et al. (2002), who performed a genome-wide detection of alternatively spliced transcripts and identified those that show tissue specificity. To determine the tissue specificity of the splice-

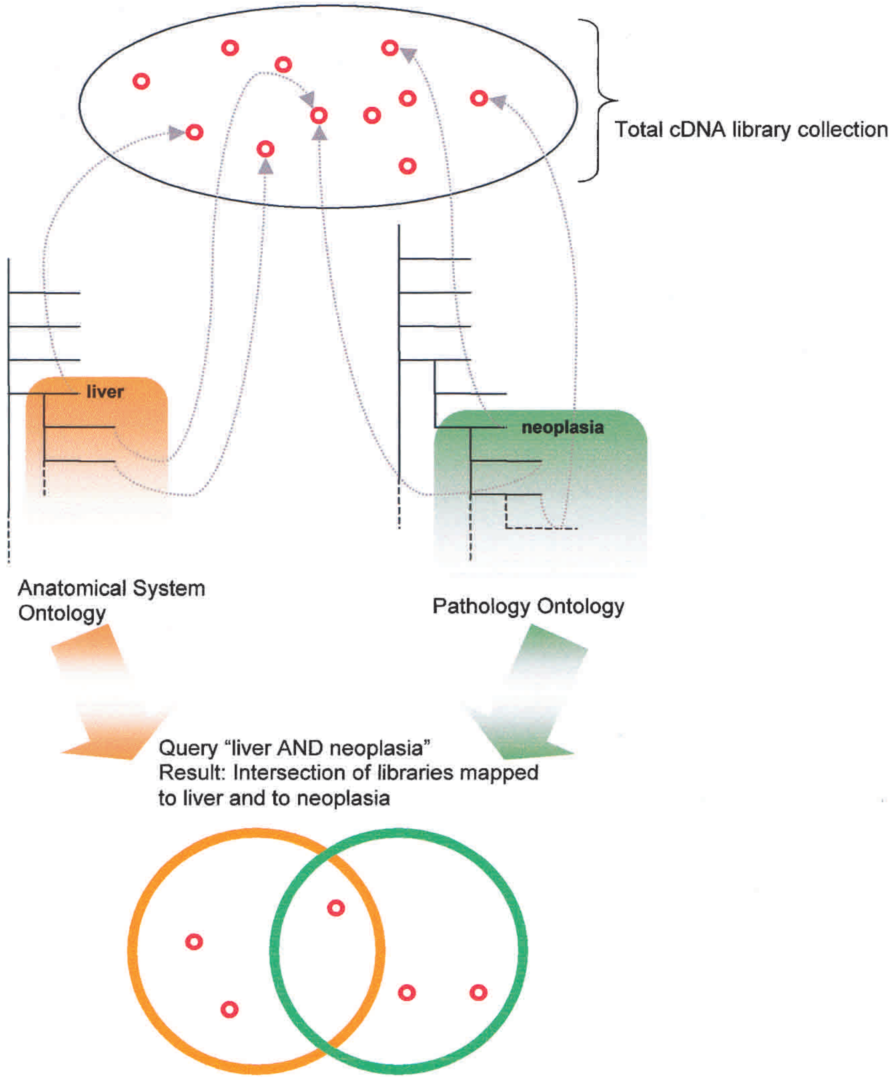
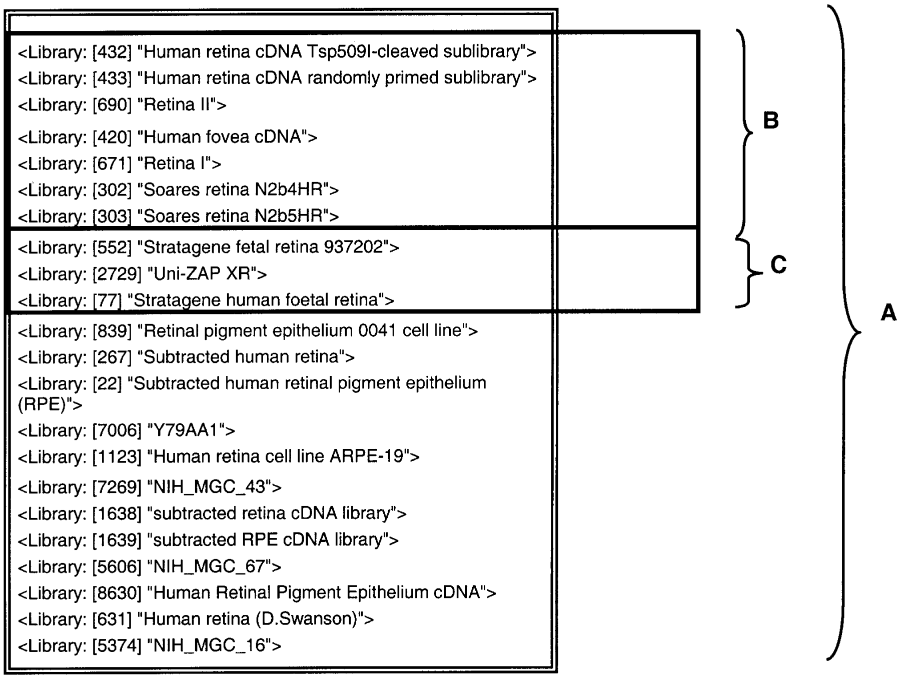

Total cDNA library collection
liver
neoplasia
Anatomical System Ontology
Pathology Ontology
Query "liver AND neoplasia"
Result: Intersection of libraries mapped to liver and to neoplasia

Figure 5 Schematic of query system. Libraries are attached to terms that are nodes in the ontology trees. Boolean queries such as "liver AND neoplasia" are translated into set operations on the libraries below the nodes matching the query terms. The result is a list of libraries that meet the criteria set by the query.



<Library: [432] "Human retina cDNA Tsp509I-cleaved sublibrary">
<Library: [433] "Human retina cDNA randomly primed sublibrary">
<Library: [690] "Retina II">
<Library: [420] "Human fovea cDNA">
<Library: [671] "Retina I">
<Library: [302] "Soares retina N2b4HR">
<Library: [303] "Soares retina N2b5HR">
<Library: [552] "Stratagene fetal retina 937202">
<Library: [2729] "Uni-ZAP XR">
<Library: [77] "Stratagene human foetal retina">
<Library: [839] "Retinal pigment epithelium 0041 cell line">
<Library: [267] "Subtracted human retina">
<Library: [22] "Subtracted human retinal pigment epithelium (RPE)">
<Library: [7006] "Y79AA1">
<Library: [1123] "Human retina cell line ARPE-19">
<Library: [7269] "NIH_MGC_43">
<Library: [1638] "subtracted retina cDNA library">
<Library: [1639] "subtracted RPE cDNA library">
<Library: [5606] "NIH_MGC_67">
<Library: [8630] "Human Retinal Pigment Epithelium cDNA">
<Library: [631] "Human retina (D.Swanson)">
<Library: [5374] "NIH_MGC_16">

Figure 6 Sample query to determine suitable libraries for a laboratory research project on differential gene expression between adult and fetal retina. (*A*) The query: "retina" results in a list of the 22 libraries associated with the term retina in the Anatomical System ontology. (*B*) Further refining the query to: "retina & adult" results in a list of the seven libraries associated with the terms retina in the Anatomical System ontology and also with the term adult in the Developmental Stage ontology. (*C*) A list of the three libraries that represent fetal retina can be obtained using the query "retina & fetus".

forms Xu et al. (2002) classified 4271 (~60%) of the publicly available cDNA libraries according to a flat list of 46 human tissue classes. This classification was used to determine the tissue distribution of alternatively spliced transcripts, identifying 667 tissue-specific alternative spliceforms. Because in the eVOC system, cDNA libraries are classified according to a more detailed hierarchical vocabulary, and because the classification is according to four orthogonal ontologies, it is possible to extend the information already derived regarding the tissue-specific isoforms identified by Xu et al. (2002).

We submitted the isoform-specific EST lists provided for a subset of the genes identified by Xu et al. (2002) as having tissue-specific isoforms to eVOC in order to determine the expression profile of each isoform according to each of the four eVOC ontologies, Anatomical System, Cell Type, Developmental Stage, and Pathology (Table 3). We were able to duplicate the tissue-specificity results described previously by comparing the expression profiles of each isoform delivered by eVOC with the published tissue specificity. Additionally, we were able to derive more information about the Pathology and Developmental Stage specificity of these isoforms. For example, IRP3 was described by Xu et al. (2002) as having a brain-specific isoform. Additional information provided by the Developmental Stage ontology in eVOC showed that this isoform is, in fact, specific to the infant brain.

By implementing a set of orthogonal, hierarchical controlled vocabularies, eVOC provides a detailed and flexible system for the detection of expression-state-specific spliceforms. eVOC can be used to identify not only tissue-specific spliceforms, but also splicing that is specific to certain developmental stages, cell types, and pathological states, or any combination of these states.

Future Applications

The eVOC ontologies have been implemented as part of a candidate disease gene-profiling tool that uses expression information in conjunction with other evidence to prioritize disease gene candidates within specified regions of the genome (D. Smedley, P. Hüsler, J. Kelso, W. Hide, and M. McCarthy, in prep.).

Availability and Interfaces (Editing and Graphical Browsing)

eVOC is provided under a BSD-style license and is available for download free of charge from http://www.sanbi.ac.za/evoc/, and can be used and modified without restriction.

Table 3. eVOC Extends the Expression Information That Can be Obtained From Other Sources

Gene name	Isoform 1		Isoform 2	
	Xu et al.	eVOC	Xu et al.	eVOC
IRP3	Brain-specific	5 nervous → brain 1 respiratory → lung 4 infant	No specificity	2 urogenital → genital → female → uterus 1 urogenital → genital → female → placenta 1 haematological → blood 3 adult
WNK1	Kidney-specific	7 urinary → kidney	No specificity	2 urogenital → genital male → penis 1 alimentary → pancreas

IRP3, described by Xu et al. (2002) as having a brain-specific isoform, was shown to be infant brain specific by combining information gathered from the eVOC ontologies. The ESTs for each isoform were submitted to eVOC and the associated terms in each of the four ontologies were examined to identify expression state specificity. Five of the six ESTs from distinct cDNA libraries were found to support the brain specificity reported by Xu et al. (2002). Further, using eVOC, four of the six libraries had been annotated with developmental stage information, and this was used to confirm that isoform 1 of IRP3 is only observed in infant libraries.

From the Web site, users are also able to download the annotated datasets, join the Expression Vocabulary Consortium, and sign up to use the eVOC mailing list.

Although genomic information is not integrated directly into eVOC, users have the ability to integrate the expression information within eVOC with human genome information through the transitive mapping of ESTs (generated from the clone libraries that are mapped to eVOC) to the genome. This functionality is being provided through the integration of eVOC with the EnsemblMart data mining resource that is part of the Ensembl Project at EBI. The eVOC ontologies will be available in the January 2003 release of the EnsemblMart database (http://www.ensembl.org/Homo_sapiens/martview). EnsemblMart is a data retrieval tool that provides users with the ability to build queries of the biological data (including genome sequence and annotation data) present in the Ensembl genome database. Because ESTs have been mapped to the genome by Ensembl, eVOC terms can be linked transitively (via their parent clone library, which is mapped to the eVOC ontologies) to the genomic sequence. As a result, users will be able to perform expression-based queries in the context of genomic data and will be able to extract transcripts and genes on the basis of the location, state, and timing of their expression.

A graphical interface for querying eVOC has been developed by Electric Genetics (Fig. 3) and is available from info@egenetics.com. This interface provides users with the ability to view the ontologies, browse the hierarchical trees, and perform set operations on the annotated cDNA library data. Using this interface, it is possible to obtain the list of cDNA libraries or ESTs returned by a query, or to provide a list of libraries or EST accessions and obtain the associated expression profile. The interface will be extended to include curation facilities, simplifying the users ability to modify the existing eVOC ontologies or create de novo ontologies of their own. In addition, Electric Genetics has developed an API that provides the ability to develop custom software to interface eVOC with external data repositories and to perform complex ontological queries on that data.

Summary

We have presented here a set of ontologies for the description of gene expression data, and have provided a database of the mappings between these ontologies and public cDNA and SAGE libraries. These have been applied successfully in retrieving expression information about ESTs from public databases, selecting clone libraries from particular expression states, and in the detection of expression state-specific alternative spliceforms.

The simple orthogonal ontologies are flexible and extensible, making them applicable to real data and allowing them to be both machine and human readable. The ontologies are under continual development; existing ontologies are extended and altered, appropriate new ontologies are added, and the annotation of expression libraries is regularly updated. Both the ontologies and the annotated expression libraries are publicly available and able to be adopted freely, modified, and integrated for both novel and existing applications. The wide number of potential applications makes eVOC a valuable resource for the biologist.

ACKNOWLEDGMENTS

We thank Cathal Seoighe for comments and suggestions. This work had financial support from the South African Government through the Department of Arts, Culture, Science, and Technology initiated Innovation Fund Program, grant 32146 (J.V, D.O, G.G, T.H., and W.H.), and the South African National Research Foundation (J.K.). The work was supported in part by funds from the European Commission ([GIFT consortium QLG2-CT-1999-00546] M.M and D.S.) and the Wellcome Trust.

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

REFERENCES

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

Bard, J. and Winter, R. 2001. Ontologies of developmental anatomy: Their current and future roles. *Brief. Bioinform.* **2:** 289–299.

Gene Ontology Consortium. 2001. Creating the gene ontology resource: Design and implementation. *Genome Res.* **11:** 1425–1433.

Gray, H.L., Bannister, L.H., Williams, P.L., Collins, P., and Berry, M.M. 1995. Gray's Anatomy. **38.**

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30:** 42–46.

Karp, P.D., Riley, M., Paley, S.M., and Pellegrini-Toole, A. 2002a. The MetaCyc database. *Nucleic Acids Res.* **30:** 59–61.

Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C., and Gama-Castro, S. 2002b. The EcoCyc database. *Nucleic Acids Res.* **30:** 56–58.

Kemp, G. and Gray, P. 2002. Modelling biological data in hierarchies. Tutorial: *Intelligent systems for molecular biology*. Edmonton, Alberta, Canada.

Rector, A.L., Wroe, C., Rogers, J., and Roberts, A. 2001. Untangling taxonomies and relationships: Personal and practical problems in loosely coupled development of large ontologies. *K-CAP'01*. 139–146.

Stevens, R., Baker, P., Bechhofer, S., Ng, G., Jacoby, A., Paton, N.W., Goble, C.A., and Brass, A. 2000. TAMBIS: Transparent access to multiple bioinformatics information sources. *Bioinformatics*. **16:** 184–185.

Xu, Q., Modrek, B., and Lee, C. 2002. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.* **30:** 3754–3766.

WEB SITE REFERENCES

ftp://ftp.ncbi.nih.gov/pub/bioannot/info/keys; Ontology for keywording human SAGE and EST libraries.

http://wombat.doc.ic.ac.uk/foldoc/foldoc.cgi?query=ontology; The Free Online Dictionary of Computing.

http://www.ensembl.org/Homo_sapiens/martview; EnsemblMart is a data retrieval tool which provides users with the ability to build queries of the biological data (including genome sequence and annotation data) present in the Ensembl genome database.

http://www.informatics.jax.org/searches/anatdict_form.shtml; The Mouse Anatomical Dictionary, an extensive mouse-specific expression ontology.

http://www.ncbi.nlm.nih.gov/CGAP/; The Cancer Genome Anatomy Project (CGAP).

http://www.sanbi.ac.za/evoc/; Ontologies and associated expression data describing human anatomical systems, cell types, pathologies and developmental stages.

www.ana.ed.ac.uk/anatomy/database/humat; A human developmental anatomy ontology.

www.cbil.upenn.edu/anatomy.php3; A human anatomical ontology.

www.mcis.duke.edu/standards/termcode/icd9/1tabular.html; The World Health Organization's ICD-9-CM system for the classification of morbidity and mortality information.

http://www.biobase.de/pages/products/cytomer.html; Cytomer, a human developmental anatomy ontology.

http://www.cbil.upenn.edu/EpoDB/release/version_2.2/controlled.vocab.html; EPOdb, a human anatomy, developmental stage and cell type ontology.

http://www.ncgr.org/genex/; GeneX, a human gene expression ontology.

http://www.nlm.nih.gov/mesh/meshhome.html; MeSH, a clinical ontology.

http://www.nlm.nuh.gov/research/umls/umlsmain.html; UMLS, a clinical ontology.

http://www.opengalen.org; GALEN, a clinical ontology.

http://www.snomed.org/main.html; SNOMED, a clinical ontology.

http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm; The World Health Organization's ICD-9-CM system for the classification of morbidity and mortality information.

Received November 12, 2002; accepted in revised form February 25, 2003.