ONTOLOGIES IN BIOLOGY: DESIGN, APPLICATIONS AND FUTURE CHALLENGES

Jonathan B. L. Bard and Seung Y. Rhee‡*

Biological knowledge is inherently complex and so cannot readily be integrated into existing databases of molecular (for example, sequence) data. An ontology is a formal way of representing knowledge in which concepts are described both by their meaning and their relationship to each other. Unique identifiers that are associated with each concept in biological ontologies (bio-ontologies) can be used for linking to and querying molecular databases. This article reviews the principal bio-ontologies and the current issues in their design and development: these include the ability to query across databases and the problems of constructing ontologies that describe complex knowledge, such as phenotypes.

"If the nineteenth century was the century of chemistry and the twentieth century the century of physics, the twenty-first century promises to be the century of biology"[1]. Determining factors in the success of the biological sciences have been the advances in technology and communications: these have enabled data to be generated in a high-throughput manner and to be distributed to scientists across the globe. Until recently, the most important task of bioinformatics was thought to be the storage, retrieval and analysis of molecular data, such as nucleotide sequences and protein structures[2]. However, as experimental technologies move from producing relatively simple data, such as nucleotide sequences, to more complex data, such as that for microarray results, images and molecular interactions, we need comparable advances in bioinformatics to manage and relate these data.

There is also a great deal of sophisticated biological knowledge, often hierarchical in nature, that needs to be integrated with molecular data: obvious examples include anatomies, signal-transduction pathways and, of particular current importance, phenotypic data. One way to do this is to represent such biological knowledge as ontologies: the resulting 'bio-ontologies' are formal representations of areas of knowledge in which the essential terms are combined with structuring rules that describe the relationship between the terms. Knowledge that is structured within a bio-ontology can then be linked to the molecular databases.

This review aims to cover the essential features of bio-ontologies, a relatively new area of bioinformatics. It deals only briefly with the formal study of ontologies in computer science, as this is an established and well-documented field[3]. We discuss a few bio-ontologies that are in use, focusing on key examples that will be relevant for the description of phenotypes. We then consider several important issues for the development of bio-ontologies, such as the production of ontologies for complex areas of knowledge (including phenotypic data), the ability to query across databases and the use of ontologies to analyse large data sets. The review concludes with a brief discussion of what the field can expect in the near future.

It is first worth pointing out that, for any ontology to be of public value, it has to be widely disseminated and accepted by the field that it aims to summarize. Sociological factors are important in ontology production and acceptance, and a strong community involvement is crucial to ensure that only single ontologies for each area are placed in the public domain. In this respect, the important standard is the *Open Biological Ontologies* (*OBO*) web site, in which many bio-ontologies are archived in a standard format (TABLE 1).

**Bioinformatics, Biomedical Sciences University of Edinburgh, Edinburgh EH8 9XD, UK.*
‡Plant Biology, Carnegie Institution of Washington, Stanford, California 94305, USA.
Correspondence to J.B.L.B. e-mail: j.bard@ed.ac.uk and S.Y.R. e-mail: rhee@acoma.stanford.edu
doi:10.1038/nrg1295

Table 1 | **Some principal biological ontologies and other web sites**

Web site name	URL	Function
AmiGO	www.godatabase.org/cgi-bin/go.cgi	Web application for browsing and searching gene ontology and gene associations
Cell Ontology (OBO)	obo.sourceforge.net/list.shtml	β-version containing >600 cell types
Directed Acyclic Graph (DAG)	www.nist.gov/dads/HTML/directAcycGraph.html	Provides definitions for *DAG*
DAG-edit	sourceforge.net/project/showfiles.php?group_id=36855	An ontology editor: a program for creating, editing and visualizing ontologies
Common Ontology Browser for Anatomy — COBrA (XSPAN)	www.xspan.org/applications/cobra/	Ontology editor that allows links to be made between ontologies
A Cross-Species Anatomy Network (XSPAN)	www.xspan.org	Project for linking anatomies of model organisms
Digital Anatomist[30]	depts.washington.edu/ventures/pfolio/fma.htm	Complex and rich ontology of human anatomy
Edinburgh Mouse Atlas Project (EMAP)	genex.hgu.mrc.ac.uk	Graphical database of mouse gene expression
Galen[31]	www.opengalen.org	A management architecture for clinical information that includes an ontology for human anatomy
Genecensus[17]	bioinfo.mbb.yale.edu/genome	Provides access to pathway-analysis tools
Gene Ontology (GO) Consortium[32]	www.geneontology.org	Resource for molecular function, biological process and cellular component ontologies
Gene Expression Database (GXD)[33]	www.informatics.jax.org/searches/expression_form.shtml	Mouse gene-expression database
GO-TermFinder	search.cpan.org/~sherlock/GO-TermFinder-0.5	Programs to facilitate analysis of *GO* annotations
Human Developmental Anatomy	genex.hgu.mrc.ac.uk/Databases/HumanAnatomy	Human anatomy of Carnegie stages* 1–20
Interoperable Informatics Infrastructure Consortium (I3C)	www.i3c.org	Provides standards for interoperability in bioinformatics
MetaCyc[13]	metacyc.org	Database of metabolic pathways
Mouse Developmental Anatomy	genex.hgu.mrc.ac.uk/Databases/Anatomy	Anatomy of Theiler stages‡ 1–26
Mouse Phenotype Ontology (MGI)	www.informatics.jax.org/searches/Phat.cgi?id=MP:0000001	Terms for describing mutant mice
Open Biological Ontologies (OBO)	obo.sourceforge.net	Umbrella web site for open bio-ontology projects
Ontology Markup Language (OML)	xml.coverpages.org/oml9808.html	XML§ markup language for ontologies
OntoExpress[18]	vortex.cs.wayne.edu/projects.htm	Tools for exploring microarray data
Ontology Web Language (OWL)	www.w3.org/TR/2003/PR-owl-ref-20031215	Specification of a semantic markup language to formalize ontologies on the web
Pathbase[12]	www.pathbase.net	Database of mouse pathology mages
Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB)[14]	pharmgkb.org/index.jsp	Resource on how genetic variability links to variability in drug responses
Plant Ontology (PO) Consortium	plantontology.org	Resource for anatomy and developmental stages ontologies for flowering plants
Protégé[34]	protege.stanford.edu/index.html	Program for making ontology frames
PubMed	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed	Database of the biomedical literature
Renamed ABox and Concept Expression Reasoner (RACER)	www.sts.tu-harburg.de/~r.f.moeller/racer	An ontology reasoner‖
Resource Description Framework Specification (RDFS)	www.w3.org/RDF	Ontology language for the exchange of knowledge on the web using XML and URI technologies¶
Semantic Web	www.w3.org/2001/sw	RDF-based representation of data on the web
Unified Medical Language System (UMLS)	www.nlm.nih.gov/research/umls	Project for handling medical concepts
WordNet	www.cogsci.princeton.edu/~wn	Lexical database of the English language

*Carnegie stage, a staging system for the embryological (as opposed to the growth) stages of human development; ‡Theiler stage, a staging system for mouse development; §XML technology, the programs associated with the extensible markup language for the exchange of structured data; ‖Reasoner, a program that explores logical relationships; ¶URI technology, universal resource indicator (a URL with additional pointer information).

Ontology basics

Although there are more technical definitions, here we can consider an ontology to be an area of knowledge that is formalized, such that the individual terms (or concepts) are defined by a set of assertions that connect them to other terms. In an anatomy ontology, for example, the developing humerus might be defined as: *part of* (in the sense of a component piece of) the arm; *has cell type* osteoblast; *has adhesion points for* muscles; and *is-a* bone. Note that the terms do not represent an individual item but the associated set — that is, not the particular humerus of Eve Smith, but all humeri. As well as being described by their relationships, terms in an ontology also contain a unique identifier (ID) (such as GO:0019505),

Box 1 | **Ontologies: rules and representation**

Representing ontologies

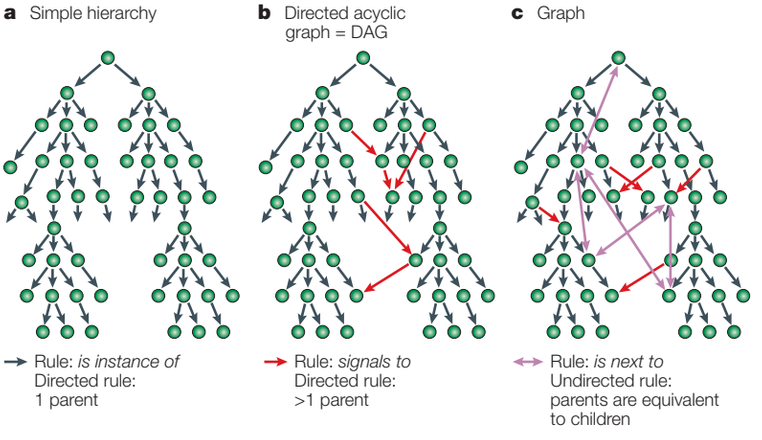
Although ontologies might seem to be abstract entities, it is usually possible to illustrate them as graphs in which vertices (nodes, leaves) and edges (lines connecting the nodes) represent the terms and the rules of the ontology[29]. For bio-ontologies, this graph is usually no more than a hierarchy: this will be simple if each term has a single parent (such as in taxonomy; panel **a**) and more complicated if a term has two or more parents or relationships (panel **b**). An example of the latter would be the *Gene Ontology* (*GO*) (see TABLE 1).

The details of the graph also depend on whether the relationships are 'directed' or not. 'Directed' relationships (as shown in panels **a** and **b**) imply a parent–child linking between the concepts: if A is *child of* B, then we would typically expect that B *is not a child of* A. By contrast, 'undirected' rules carry no such implication: if A is *next to* B, then B is also *next to* A (panel **c**). If all the relationships in a valid ontology are directed, it is not possible to make closed loops, and the ontology can be represented by a directed acyclic graph (DAG; panel **b**).

The transitivity rule

One important aspect of the assertions and rules that together define the ontology is that they can be used to make logical inferences about the terms and their associated properties. An assertion that connects C to B together with one that connects B to A implies that the same relationship connects C to A; the logic of this inference process is defined by the 'transitivity' rule. To illustrate this with the anatomical example given in the text, the humerus is: *part of* the arm; *has cell type* osteoblast; *has adhesion points* for muscles; and *is a* bone. In this example, *part of* is transitive and the properties *has cell type* and *has adhesion points* can be inferred to hold for the whole, B, if they hold for the part, C. By transitivity, these properties will also hold for A if B is *part of* A; that is, the arm includes all the cell types and expressed genes for each of its constituent tissues. By contrast, *descends from* is not transitive and no deduction about the child can be made on the basis of the parent. (The reader should note that this analysis of the *part-of* relationship (or 'mereology') is highly simplified[5].)

The *is-a* rule is also transitive but in the opposite direction: for example, individual bones have specific features that are not common to all bones (only the humerus has a radial groove). In terms of the previous example, if A, B and C are linked by *is-a* relationships, the appropriate properties of A can be associated with B and the properties of both B and A with C. Figure reproduced with permission from REF. 29 © (2003) Wiley.



a name ('resorcinol metabolism', for example), a textual definition (such as 'the chemical reactions and physical changes involving resorcinol ($C_6H_4(OH)_2$), a benzene derivative with many applications (including dyes, explosives, resins and as an antiseptic)') and synonyms (such as '1,3-benzenediol metabolism' or '1,3-dihydroxybenzene metabolism').

Ontologies are different from annotations (descriptions of data objects) in that they formalize the meaning of terms through a set of assertions and rules that are collectively known as a 'description logic'. An advantage of ontologies is that the description logic can be used both for querying an information set and for facilitating analyses across information sets that are not traditionally accessible to searching and comparing. If, for example, a database stores gene-expression data for the mouse forelimb skeleton under its individual parts (the humerus, radius, ulna, carpal bones, and so on), then a query on gene expression in the forelimb skeleton can automatically use the *part of* relationship to identify the constituent tissues, search on their database entries for expression and then combine them to list all the genes that are expressed in the forelimb skeleton. Furthermore, the structure of ontologies can be represented and viewed as graphs (see BOX 1 for more details).

Ontology IDs. Each term in the ontologies that are associated with the *OBO* has an ID that has two components: a letter code that specifies the ontology type and a number. For example, CL:0000188 represents a skeletal muscle cell in the *Cell Ontology* (*OBO*): the ontology type is defined by the prefix CL and the number represents a unique entity in the CL ontology. IDs can be used in two ways: to link a biological database to ontologies and to connect different biological databases (interoperability). If a database, such as a sequence repository, associates its data objects with ontology IDs, a user can query the database for data that is associated with a particular ontology ID

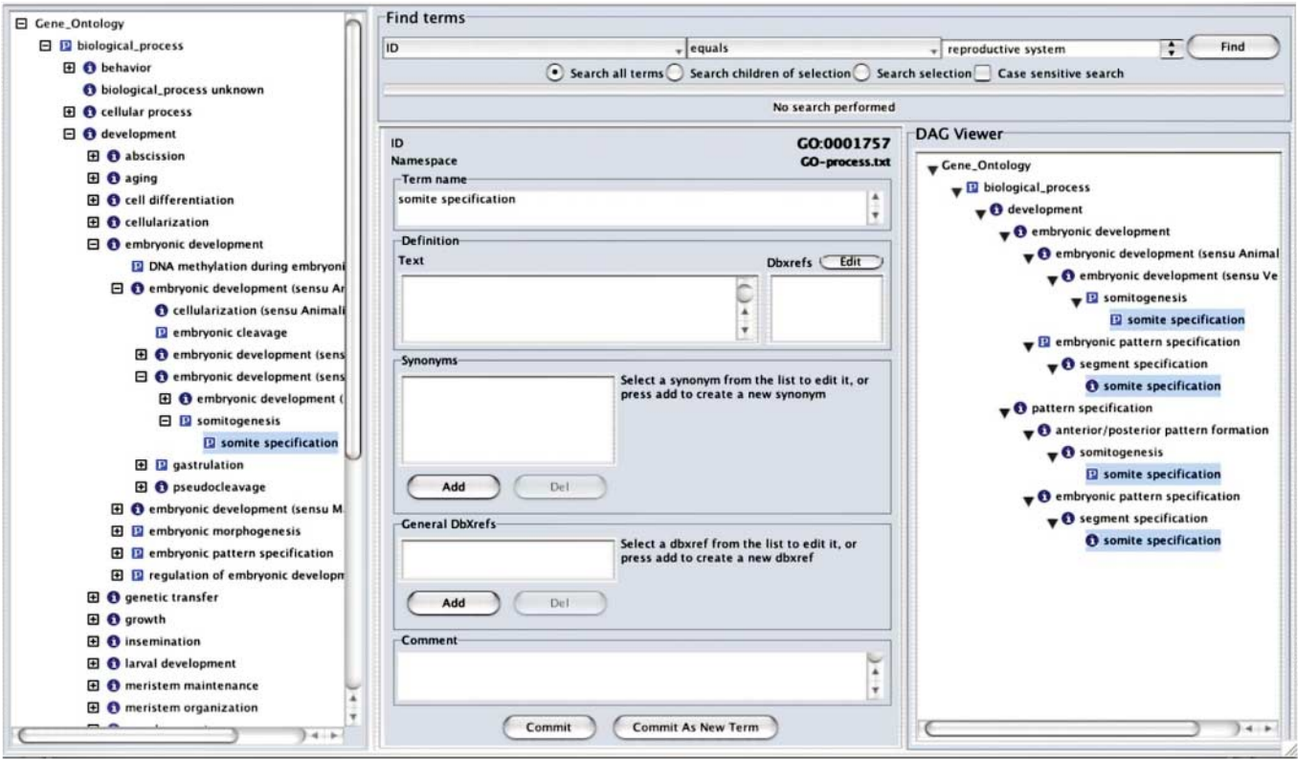


Figure 1 | **The *GO* process ontology visualized with the *DAG-edit* program (version 4).** This *Gene Ontology* (*GO*) screen shot illustrates the mechanics of ontologies. The left panel shows the expanded ontology with the process 'somite specification' highlighted; the middle panel provides the unique identifier (ID) together with search and other facilities; the right panel shows all four hierarchies in which somite specification is found. DAG, directed acyclic graph.

and also use the logic of the rules in the ontology to ask further questions about the associated data. Ontology IDs can also be used to allow one database to query another directly. If different biological databases use the same ontologies to describe their data objects, the ontology IDs can be used as the currency with which associated data in individual databases can be retrieved[4].

Examples of bio-ontologies

Ontologies have been used in biology for some time, although they have not necessarily been recognized as such. Indeed, the field of systematics could be considered to be a classic example of ontologies in biology. An important large-scale example for molecular biology is the *Unified Medical Language System* (*UMLS*) (see TABLE 1) hierarchy of terms and relationships, which is used for searching *PubMed* and many other resources. A typical, small-scale example of an ontology would be the controlled vocabulary for sexual phenotype, in which male, female and hermaphrodite are the three obvious *classes of* gender (the *is-a* relationship). Here, we focus on three ontologies that should be instrumental in describing phenotypic data.

Gene Ontology. The *Gene Ontology* (*GO*) is by far the most widely used bio-ontology. It aims to formalize our knowledge about biological processes (FIG. 1), molecular functions and cell components, in three orthogonal (mutually independent) hierarchies. The *GO* has reached a substantial size, containing approximately 16,500 terms, with the nodes and leaves within each hierarchy being connected by *is-a* or *part of* relationships. As the terms can have more than one parent, the structure is represented graphically as a directed acyclic graph (DAG, see BOX 1b).

A practical importance of the *GO* is that it is linked to a database of more than 120,000 gene products from almost 20 experimental organisms — including animals, plants, fungi, bacteria and viruses — in which the proteins are tagged with *GO* IDs. This means that a user can identify both the proteins associated with a specific *GO* term (there are, for example, 16 proteins associated with virus–host interactions) and all of the *GO* terms associated with a given protein by using an appropriate browser, such as *AmiGO* (see TABLE 1). For each gene, the user is directed to the database that contributed the annotation to find more detailed information about that gene. This type of infrastructure provides a simple way of traversing between areas of knowledge.

Anatomical ontologies. These ontologies include the supracellular physical structures that make up a particular organism. They can be organized using appropriate rules, such as relative location (the atrium *is part* of the heart), lineage (the gut *is derived from* the endoderm) and class (the cardiovascular system *is-a* organ system).

Although it might seem that making such ontologies is straightforward, the handling of anatomies actually highlights some interesting problems. For example, it is worth considering the requirements for two very different users interested in human anatomy. A surgeon will want an ontology to specify those tissues he might have to cut through if he enters the body from a particular angle to operate on the diaphragm. This ontology would have to include both tissues and their spatial relationships (such as *next to*) and can be self-contained. A developmental biologist who wants to identify the genes associated with a particular tissue at a particular developmental stage would require an ontology of tissue names (ordered by a *part-of* rule) that were linked to a database that contains gene-expression data, and spatial relationships might not be essential.

For human clinical anatomy, two sophisticated ontology frameworks are available: *Galen* and the *Digital Anatomist* (TABLE 1). Both handle geographical and other knowledge about adult tissues and include many thousands of terms in a wide variety of relationships. There are, for example, several relationships that can be subsumed under *part of*: the coronoid process *is a physical component of* the ulna bone, marrow *is contained within* the ulna bone and the pancreas is *a member of* the glandular system (these types of relationship are the subject of an area of logic called 'mereology'[5]; see BOX 1). Many such relationships are included in both of these ontologies, which set out to be comprehensive. However, they are not always easy to use.

By contrast, the developmental biologist will probably find it useful to consult the ontology of *Human Developmental Anatomy*, which is modelled on the *Mouse Anatomy Ontology* (see TABLE 1) and is designed for archiving gene-expression data on human embryos in their first seven weeks of development. It has several thousand tissues that are linked by a simple *part-of* rule that usually means *is a physical component of*. This ontology is intuitive to use but includes no spatial rules; its use is therefore limited to handling tissue-associated data and to defining the tissues that are present at a given developmental stage.

The *OBO* web site provides access to a further ten anatomies for common plants and animals, with all using *part-of*, *is-a* and, in some cases, *is-derived-from* relationships. All are embedded in the core databases for their species, and several are now linked to gene-expression data.

Cell Ontology. This new and still unfinished ontology is being designed to provide all model species with a common language and ID set for cell phenotypes. Its production illustrates some of the core aspects of ontology design. First, its conceptual framework required an analysis of the contexts in which cells are used and described (morphology, function, species, and so on) and the sorts of relationships required (*is-a* and *is-derived-from*). Second, in attempting to make it useful for all organisms, ubiquitous cell types are at a much higher level in the hierarchy than those restricted to specific families of organisms. Third, it required the garnering of data from a range of standard textbooks. Fourth, it involved people from widely different fields freely giving expert knowledge: Michael Ashburner and J.B.L.B. initiated the ontology and provided invertebrate and vertebrate data, whereas David States and S.Y.R. provided blood-cell and plant data, respectively. The prototype ontology is now publicly available for comment and is being improved by the community. In the end, the ontology will reflect the expertise of the community that uses it rather than that of any individual and it will be available for anyone who wishes to code cell-type identities in a standard way.

Creating and displaying ontologies

There are several tools for editing and viewing ontologies (TABLE 1). To the user with the appropriate viewing software, an ontology appears as a tree or network (FIG. 1). However, the underlying textual syntax is far more opaque, as an inspection of any ontology in a text editor rapidly makes clear. Ontologies can be written in either frames or flat files. Programs such as *Protégé* or *Ontolingua* generate a separate frame (or page) for each knowledge item that contains information such as its links, definitions and its relationships to other items. By contrast, flat files include knowledge of all the items of an ontology within a single file, and are the more commonly used format. Flat files can be written in a range of functionally similar formats, such as *OWL*, *GO*, *RDF* and *XML*, by using ontology editors (see TABLE 1). The *GO* flat file, for example, provides a common format that can be edited and viewed using several editors: the standard editor for *GO* is *DAG-edit*, which allows a user to create, edit and view an ontology (FIG. 1). A recently developed tool, *COBrA*, can also translate one format into another and allows a user to make links between two ontologies.

Applications

Bio-ontologies now have a wide variety of uses, the most important of which is the representation of knowledge in a computer-comprehensible way, interoperability across databases, and the annotation and analysis of large-scale data with ontology IDs. Here, we consider each of these roles, starting with how the field is approaching the representation of complex knowledge in which a single, simple ontology is inadequate.

Handling complex areas of knowledge — the phenotype example. Most bio-ontologies are relatively simple in that they describe the essential features of well-defined and local domains of knowledge. However, there are complex areas of knowledge — a challenging example being the description of mutant phenotypes — that cannot easily be described in this way. 'Phenotype' can be defined as the observable and measurable characteristics of an organism, which result from the interaction of the organism's genetic 'blueprint' (its genotype) and the environment. Phenotype information is currently described as free-text in most biological databases[6–8], although efforts have been made to store the information in more structured ways[9,10]. Free-text, database-specific

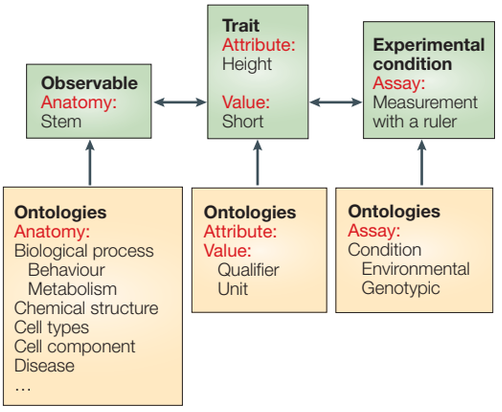


Figure 2 | **Coding phenotype.** The diagram shows the three domains of information that are used to describe a phenotype: the observable, the trait and the experimental condition (upper, green boxes). The set of ontologies that would be needed to code each domain are given in the lower, yellow boxes. The example describes a dwarf plant phenotype for which the ontologies (red text) and terms in the ontologies (standard text in upper boxes) are provided. The observable quantity or character can be anything for which characteristics can be observed or measured — in this case, a stem.

phenotypic descriptions cannot be queried and compared easily, especially if they lie outside a researcher's immediate research focus.

Phenotypic descriptions can be handled using ontologies in several ways. The first and most straightforward is to make a dedicated ontology specific for an organism. The *Mouse Genome Database* (the Jackson Laboratory, USA; see TABLE 1) has taken this approach for the mouse, and its ontology is being used to code phenotypes of mutant mice[11]. One caveat with this approach is that terms might be needed to represent all of the variety of phenotypes under different conditions. This could result in a rapid increase in the size of the ontology, making it laborious to maintain. Another problem lies in the difficulty of extending the mouse phenotype ontology to any other organism.

The second approach to describing phenotype is to make a composite annotation using several simpler ontologies. We can deconstruct a phenotype into three independent components: the observable character, the trait and the experimental condition. Each of these components can then be described using one or more ontologies (FIG. 2). The observable character or quantity can be anything for which characteristics can be observed or measured. Examples include an anatomical structure of an organism, such as a stem or a metabolite, the amount of which can be quantified. Some existing ontologies, such as *GO* or the *Cell*, *Anatomy* and *Biochemical substance* ontologies at *OBO*, can be used to describe such an observable character. The trait is the attribute or characteristic that is being measured; examples of traits include height, weight, viability and enzymatic activity. In addition to the *GO* function ontology, there are a couple of ontologies for trait (in development phase) at *OBO*, which can be found under the *Attribute_and_value* and *Plant trait* categories. Finally, the experimental condition under which the trait is measured can be described by considering the assay method and the environmental and genotypic conditions under which the measurement is taken. There are several ontologies that deal with experimental methods at *OBO*, and these fall under *Experimental methods* and a preliminary *Environmental condition* ontology. FIG. 2 shows how the 'dwarf plant' phenotype can be translated into a plant that has 'short height' (*Trait* ontology) of the stem (*Anatomy* ontology) that is assayed using a 'ruler measurement' (*Assay* ontology).

Making composite annotations using these ontologies requires that the relationship context of each ontology to the data object (for example, *having* a trait, *of* a tissue, *measured by* an assay) be included. TABLE 2 illustrates some simple examples of composite annotations to describe a gene's expression patterns and biological roles using multiple ontologies. It should be pointed out that the difficulty with this approach is that assigning a unique code for a mutant phenotype becomes impossible and the code is actually the set of IDs from the relevant ontologies. This set is, however, easy to search.

Several databases have used multiple ontologies to describe complex information (see TABLE 1). For example, *Pathbase* is a database of mouse pathological images that uses separate anatomy, pathology, cell and other ontology IDs to access the relevant image[12]. *MetaCyc*, which handles metabolic pathways[13], has been available for several years. It uses ontologies for metabolic pathways, reactions, compounds and cellular components to describe metabolism, but does not yet link metabolism with anatomy and developmental stages. A further example is *PharmGKB*, which handles PHARMACOGENETIC information[14]. *PharmGKB* aims to represent the relationship between genotype and phenotype for drug response in humans, but its individual component ontologies are not yet enumerated in detail. Although

Table 2 | **Annotations of *Arabidopsis* genes using multiple ontologies**

Gene	Relationship	Primary ontology	Context	Qualifier ontologies
AOC1	*is expressed in*	*Anatomy*: leaf	1:during	*Temporal*: senescence
OST1	*exhibits*	*Function*: protein kinase activity	1: in 2: during	*Anatomy*: guard cell *Process*: response to drought
AG	*is involved in*	*Process*: specification of organ identity	1: of 2: in	*Anatomy*: petals *Taxonomy*: *Arabidopsis thaliana*

The example shows how composite annotations can be used to describe the roles of a gene by using several terms from different ontologies. The first term describes a primary relationship to the gene product and subsequent terms act as qualifiers within a certain context.

PHARMAKOGENETICS
The study of drug responses related to inherited genetic differences.

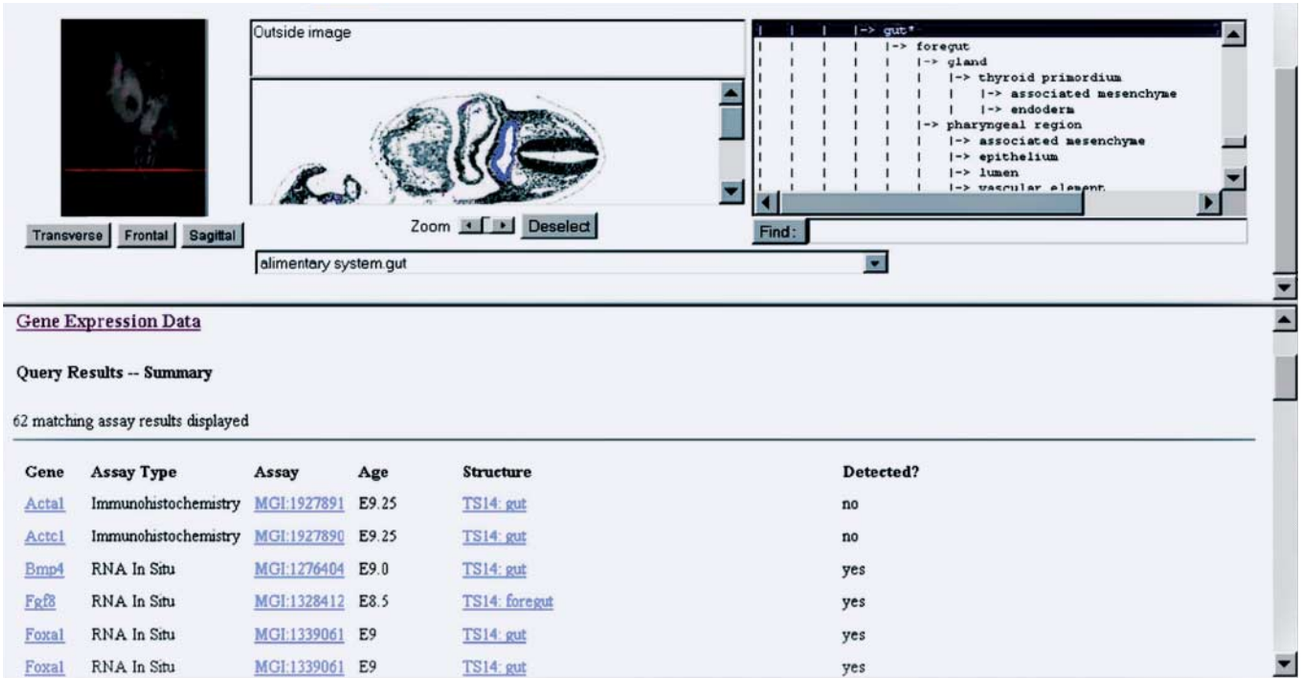


Figure 3 | **Interoperability between mouse anatomy from digital sections and a gene-expression database.** The top part of the picture shows a section of an embryonic stage (E)9-mouse embryo from the *Edinburgh Mouse Atlas Project* (*EMAP*; see TABLE 1), with the gut filled in blue; the lower half shows the associated gene-expression data sent from the *Mouse Gene Expression Database* (*GXD*; see TABLE 1), hosted at the Jackson Laboratory, USA.

this approach has been used to describe complex information and is being used to describe phenotype, powerful query and analysis tools that take advantage of such structured knowledge have yet to come.

The third, and still new, way of handling complex data, such as phenotype, is to combine terms in multiple, orthogonal ontologies to create a single new ontology[15]. The process of heart development can, for example, be described as a combination of the relevant terms in the anatomy (for example, heart) and the *GO* process (such as development) ontologies. Although this set of joint terms might provide some novel concepts that are worth investigating, it suffers from the problem that many of the terms might not be biologically valid (for example, 'heart' plus 'photosynthesis'). Deciding which terms should be excluded in the primary ontologies to make the cross-product ontology can be time-consuming, to the extent that if more than two ontologies are needed for the description, the task of examining and validating all the cross products will simply become impractical.

Much work will be needed to optimize the way in which ontologies handle complexity such as phenotypes. A good solution will ensure that implementation will be organism-independent as much as possible to facilitate interoperability across databases. During the past two years, the curators of approximately 15 biological databases (see online link box) have met to discuss the problem of representing phenotype information. This resulted not only in identifying the issues at hand for each community, but also in defining the ontologies that are needed for interoperability and for describing phenotype robustly (FIG. 2). Such a forum will be crucial to the success of handling complex information across many biological databases.

Interoperability. Interoperability, or the querying of one database by another, is becoming increasingly important. Ontologies are beginning to be valuable here through the use of the unique IDs that are associated with each of their terms. A simple example typifies the importance of this: the *Edinburgh Mouse Atlas Project* (*EMAP*) web site (Edinburgh, UK) includes two-dimensional section sets of many three-dimensional models of early mouse development in which the individual tissues have been delineated and assigned *EMAP* IDs (in essence, this is a graphical ontology in which the knowledge is visual rather than textual). In the user interface, a tissue is highlighted when the cursor reaches it and a click of the mouse sends the ID of that tissue to *GXD*, the *Mouse Gene Expression Database* (Jackson Laboratory, USA), as a query. As *GXD* uses the same anatomy IDs as *EMAP*, it responds to the query by producing a table of all the genes that are expressed in that tissue and returns them to the user's screen through *EMAP* (FIG. 3). *GXD* also carries *GO* IDs and, therefore, searches can also be made on the basis of gene annotations to *GO* terms. For example, the query can be given as 'return only those genes expressed in the developing heart and have transcription-factor activities'.

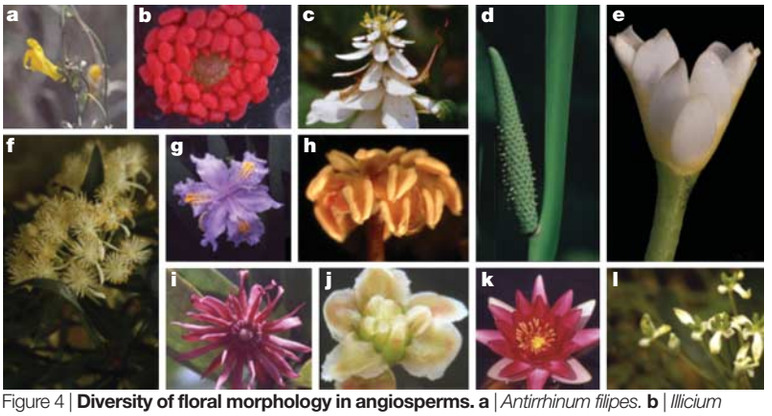


Figure 4 | **Diversity of floral morphology in angiosperms. a** | *Antirrhinum filipes.* **b** | *Illicium floridanum* (sexual organs only). **c** | *Houttuynia cordata* filled cultivar. **d** | *Acorus calamus* s.S. **e** | *Aponogeton distachyos.* **f** | *Tasmania moorei* male. **g** | *Iris japonica* (petaloid stigma). **h** | *Amborella trichopoda* male. **i** | *Illicium floridanum.* **j** | *Amborella trichopoda* female. **k** | *Nymphaea hybrida* var. *escarbuncle.* **l** | *Tasmania moorei* female. Images courtesy of M. Buzgo, University of Florida, USA.

Exploring large data sets. An important application of bio-ontologies is their use in investigating gene function. Several biological databases now use the *GO* terms to assign functions, biological roles and sub-cellular locations of proteins. These annotations can be used in combination with sequence-similarity analysis to infer the function, role and location of proteins in, say, agronomically important animals and plants, even when their genomes have not been fully sequenced. For example, to identify candidate genes that correspond to QUANTITATIVE TRAIT LOCI in swine and cattle, Harhay and Keele annotated expressed sequence tags (ESTs) from swine and cattle with the *GO* terms by sequence comparison with model species for which genomes have been annotated with *GO*[16]. Similarly, groups of annotated genes can be compared to determine over- or under-representation of the annotated terms. Several bioinformatics tools, such as *GeneCensus*[17], *OntoExpress*[18] and *TermFinder*[19] (see TABLE 1), can compare the statistical significance of the representation of *GO* terms between two sets of genes (for example, from a pair of expression clusters identified from microarray analysis). Furthermore, annotations of genes using ontologies can lead to the development of algorithms that can use these annotations to predict function[20].

Mapping knowledge domains. Different ontologies can be mapped to each other and these links provide hooks from one expert domain of knowledge to another, thereby creating an ontology network that allows a user working in one area to take advantage of knowledge from a related area[21–23] (see TABLE 1). For example, Bodenreider and colleagues mapped *UMLS*, a highly specialized medical ontology, onto *WordNet*, an electronic lexical database for the English language, in an attempt to identify an overlap between the two[24]. Their work shows how the knowledge domains of two different types of community — medical specialists and the general public — can be linked. For instance, a patient can search for a common disease name in *WordNet* and then be linked to the comparable term in *UMLS* and therefore to the details of the disease. More advances on such mapping analysis between ontologies using, for example, SEMANTICS, or even NATURAL LANGUAGE PROCESSING, could be a key factor in closing the distances between different experts, disciplines and even sociological boundaries. At a more biological level, the *XSPAN* project seeks to make mappings across the anatomies of the main model organisms on the basis of cell type, homology and analogy. These links could be useful in identifying related mutant phenotypes in different model organisms.

Once the ontologies are networked and data objects such as genes are annotated to the ontologies, we can start to ask questions about the genes that are involved in, say, a process such as the tricarboxylic acid (TCA) cycle in *Escherichia coli*, *Arabidopsis*, *Drosophila* and humans. We can, for example, address how the TCA cycle has evolved on the basis of the properties (and therefore mechanisms) of its constituent proteins. We might also be able to apply ontological approaches to address systematics, taxonomy and evolution. It is, for example, known that there was a large radiation of flowering plants (angiosperms) approximately 150 million years ago that diversified the morphology of flowering plants[25] (FIG. 4). When genes from a wide range of flowering plants are annotated to angiosperm anatomy and developmental stages ontologies, it will be possible to perform a systematic analysis of the genetic changes associated with taxonomic diversification. Recently initiated projects — such as the *Plant Ontology Consortium* (see TABLE 1), which attempts to develop a unified anatomy and developmental stages ontologies for angiosperms, and the *Floral Genome Project,* which attempts to identify genes involved in floral development in angiosperms — will facilitate such analyses.

Future projects, prospects and challenges

Bio-ontologies for the obvious knowledge domains are now in place and are under active curation. Attention is beginning to be focused on ontologies that describe *in vivo* cell imaging, molecular interactions and data that are linked to space rather than text[26]. For example, as gene-expression domains might not be restricted to tissue boundaries, it is better to represent them as volume units (VOXELS) in a three-dimensional model of the anatomy (see *EMAP* web site). As more knowledge of genetic networks is gleaned, we can also look forward to an ontology of signalling pathways. In addition, to fully correlate between phenotype and genotype, a systematic and standard way of describing genotype will be needed.

It cannot be emphasized too strongly, however, that the key to the general use of ontologies will be access to the data in biological databases that are annotated with the knowledge in these ontologies. Many biological databases are now incorporating ontology IDs (particularly those from the *GO*) and using them to annotate data objects. The more that this is done, the more useful these resources will be for the community. Unfortunately, most of the current search and analysis tools for mining

QUANTITATIVE TRAIT LOCUS
Genetic locus or chromosomal region that contributes to the phenotypic variation in continuously varying traits, such as weight.

SEMANTICS
The meaning of a string in some language; this is distinct from syntax, which describes how symbols can be combined independently of their meaning.

NATURAL LANGUAGE PROCESSING
Computer understanding, analysis, manipulation and/or generation of natural (human) language.

VOXEL
The three-dimensional, or volume, equivalent of a pixel (two-dimensional picture unit).

GRAPH THEORETICAL APPROACH
An approach to extracting meaning from ontologies that depends on using the intrinsic properties of graphs.

these data are not as powerful as might be liked. However, analysis of the ontologies using GRAPH THEORETICAL APPROACHES[27,28] might provide interesting insights about the representation of knowledge. We hope that more tools will be produced that exploit the ontologies and their associations with data objects.

A difficult problem that the field has yet to confront is how to deal with a term that is represented in several, possibly overlapping, ontologies. For example, *MetaCyc* contains a term for which the ID is 'NAD BIOSYNTHESIS III'. This term is synonymous to GO:0019360 in *GO*, which corresponds to nicotinamide nucleotide biosynthesis from niacinamide. A term having several apparently unique IDs cannot be fully interoperable on the basis of any one of them. *GO* provides a mapping of different ontologies to *GO*, but this is mostly a manual effort and keeping it updated is a major challenge. There is no easy answer to this problem, but one possibility is that the *OBO* (or a similar site) could hold a look-up table for all IDs and their alternatives that can be accessed automatically. This will of course only be achieved if there are both funding and a communal agreement to share codes, and even then, appropriate software needs to be implemented before such a system would itself be fully interoperable.

It is nevertheless reasonable to expect that the development of ontologies, annotation of data objects using the ontologies and sophisticated search tools should enable us to start to systematically address the missing gaps in our knowledge. For example, once genes with known function are linked to the ontologies, we can ask how many genes in a genome are not associated with a molecular function, biological process, expression pattern or cellular location. Similarly, we can examine which processes, functions and cellular components are described by known molecular entities and which aspects of biology have no, or unexpectedly few, genes associated with them. There is much to be explored.

Conclusions

An ontology makes explicit knowledge that is usually diffusely embedded in notebooks, textbooks and journals or just held in academic memories, and therefore represents a formalization of the current state of a field. Integrating this knowledge poses two problems. First, not everyone in a field agrees on either the facts or the relationships. Second, knowledge changes with time, even in apparently ossified subjects such as anatomy — for example, we still do not know all the cell-lineage relationships of human anatomy. The first problem can be adequately handled if ontologies are felt to be owned by the field rather than just the individual authors. Mechanisms for sharing the development with those in the field — including establishing a forum for those interested in similar areas of ontology development and soliciting or incorporating feedback from individual researchers — will facilitate public ownership. Public support is just as important for maintenance of ontologies as it is for databases. This will only happen if ontologies are actively curated and this, of course, is the solution to the second problem.

If ontologies are properly curated over the longer term, they will come to be seen as modern-day (albeit terse) textbooks providing online and up-to-date biological expertise for their area. In another sense, they will provide the common standards needed for producing a strong biological framework for integrating data sets. Ontologies therefore provide the formal basis for an integrative approach to biology that complements the traditional deductive methodology.

1. D'Souza, D. *The Virtue of Prosperity: Finding Values in an Age of Techno-Affluence* (Simon and Schuster, Inc., New York, 2000).
2. Baxevanis, A. D. (ed.). *Current Protocols in Bioinformatics* (Wiley, New York, 2002).
3. van Heijst, G., Schreiber, A. & Wielinga, B. Using explicit ontologies in KBS development. *Int. J. of Human-Computer Studies* **46**, 183–292 (1997).
4. Stein, L. D. Integrating biological databases. *Nature Rev. Genet.* **4**, 337–345 (2003).
5. Simons, P. *Parts: A Study in Ontology* (Oxford Univ. Press, Oxford, UK, 1987).
6. Twigger, S. *et al.* Rat Genome Database (RGD): mapping disease onto the genome. *Nucleic Acids Res.* **30**, 125–128 (2002).
7. Garcia-Hernandez, M. *et al.* TAIR: a resource for integrated *Arabidopsis* data. *Funct. Integr. Genomics* **2**, 239–253 (2002).
8. Lawrence, C. J., Dong, Q., Polacco, M. L., Seigfried, T. E. & Brendel, V. MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res.* **32**, D393–D397 (2004).
9. Drysdale, R. Phenotypic data in FlyBase. *Brief Bioinform.* **2**, 68–80 (2001).
An early example of the use of multiple ontologies to describe phenotype.
10. Ware, D. H. *et al.* Gramene, a tool for grass genomics. *Plant Physiol.* **130**, 1606–1613 (2002).
11. Blake, J. A., Richardson, J. E., Bult, C. J., Kadin, J. A. & Eppig, J. T. MGD: the Mouse Genome Database. *Nucleic Acids Res.* **31**, 193–195 (2003).
12. Schofield, P. N. *et al.* Pathbase: a database of mutant mouse pathology. *Nucleic Acids Res.* **32**, D512–D515 (2004).
13. Krieger, C. J. *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **32**, D438–D442 (2004).
14. Hewett, M. *et al.* PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res.* **30**, 163–165 (2002).
15. Hill, D. P., Blake, J. A., Richardson, J. E. & Ringwald, M. Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. *Genome Res.* **12**, 1982–1991 (2002).
Proposes a way to generate more specific ontologies by combining concepts from two orthogonal ontologies.
16. Harhay, G. P. & Keele, J. W. Positional candidate gene selection from livestock EST databases using Gene Ontology. *Bioinformatics* **19**, 249–255 (2003).
17. Lin, J. *et al.* GeneCensus: genome comparisons in terms of metabolic pathway activity and protein family sharing. *Nucleic Acids Res.* **30**, 4574–4582 (2002).
18. Draghici, S. *et al.* Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.* **31**, 3775–3781 (2003).
19. Christie, K. R. *et al. Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.* **32**, D311–D314 (2004).
20. King, O. D. *et al.* Predicting phenotype from patterns of annotation. *Bioinformatics* **19** (Suppl. 1), I183–I189 (2003).
Uses decision trees to predict phenotypes of yeast mutants on the basis of genes' annotations to *GO* and other phenotypic descriptions.
21. Tulipano, P. K., Millar, W. S. & Cimino, J. J. Linking molecular imaging terminology to the gene ontology (GO). *Pac. Symp. Biocomput.* 613–623 (2003).
22. Bodenreider, O., Mitchell, J. A. & McCray, A. T. Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. *Proc. AMIA Symp.* 61–65 (2002).
23. Leroy, G. & Chen, H. Meeting medical terminology needs — the Ontology-Enhanced Medical Concept Mapper. *IEEE Trans. Inf. Technol. Biomed.* **5**, 261–270 (2001).
Describes a query tool that involves the mapping of different concepts using human-created ontologies and natural language processing.
24. Bodenreider, O., Burgun, A. & Mitchell, J. A. Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study. *Stud. Health Technol. Inform.* **95**, 379–384 (2003).
Maps *GO* terms and NCBI's LocusLink terms to WordNet to determine the overlap between molecular biological and lay knowledge.
25. Judd, W. S., Campbell, C. S., Kellogg, E. A., Stevens, P. F. & Donoghue, M. J. *Plant Systematics: A Phylogenetic Approach* (Sinauer Associates, Inc., Sunderland, Massachusetts, 2002).
26. Cook, D. L., Farley, J. F. & Tapscott, S. J. A basis for a visual language for describing, archiving and analyzing functional models of complex biological systems. *Genome Biol.* **2**, RESEARCH0012 (2001).
Provides a lexicon of icons to graphically represent molecular biology information.
27. Sigman, M. & Cecchi, G. A. Global organization of the WordNet lexicon. *Proc. Natl Acad. Sci. USA* **99**, 1742–1747 (2002).

Applies graph theoretical calculations to analyse the organization of WordNet.
28. Ogata, H., Fujibuchi, W., Goto, S. & Kanehisa, M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res.* **28**, 4021–4028 (2000).
Uses graph comparison methods to correlate the genome locations of microbial genes and these organisms' metabolic pathways.
29. Bard, J. Ontologies: formalising biological knowledge for bioinformatics. *Bioessays* **25**, 501–506 (2003).
30. Rosse, C. *et al.* Motivation and organizational principles for anatomical knowledge representation: the digital anatomist symbolic knowledge base. *J. Am. Med. Inform. Assoc.* **5**, 17–40 (1998).
Proposes a human anatomy ontology that accommodates both the systemic and regional (topographical) views of anatomy.
31. Trombert-Paviot, B. *et al.* GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. *Int. J. Med. Inf.* **58–59**, 71–85 (2000).
Provides an information-management architecture for handling all types of clinical data in language-independent ways.
32. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
33. Hill, D. P. *et al.* The mouse Gene Expression Database (GXD): updates and enhancements. *Nucleic Acids Res.* **32**, D568–D571 (2004).
34. Noy, N. F. *et al.* Protege-2000: an open-source ontology — development and knowledge-acquisition environment. *Proc. AMIA Symp.* 953 (2003).

Acknowledgements
We thank the curators of the various animal, plant and prokaryote databases who participated in the mutant phenotype ontology meetings (see list of URLs in online links box for groups that participated). We are grateful to S. Aitkin for commenting on the material in box 1 and to M. Buzgo for providing the photographs in figure 4 and for helpful comments on the manuscript. S.Y.R. is supported in part by the National Science Foundation (NSF), and J.B.L.B. thanks the Biotechology and Biological Sciences Research Council (BBSRC) for funding. This is Carnegie publication 1680.
We dedicate this paper to the late Robin Winter who articulated much of our knowledge about human congenital dysmorphologies and who is sorely missed.

Competing interests statement
The authors declare that they have no competing financial interests.

Online links

FURTHER INFORMATION
Discussion paper by Michael Ashburner on phenotype and trait ontology:
http://obo.sourceforge.net/pheno/pheno.html

Minutes from phenotype meetings:
http://obo.sourceforge.net/pheno

Database groups that participated in phenotype meetings
The *Arabidopsis* Information Resource:
http://www.arabidopsis.org
Berkeley *Drosophila* Genome Project:
http://flybase.net/annot
DictyBase: http://dictybase.org
Flybase: http://flybase.org
Gramene: http://www.gramene.org
International Crop Information System:
http://www.icis.cgiar.org
The Institute for Genome Resources — microbial systems:
http://www.tigr.org
The London Dysmorphology Database:
http://www.hgmp.mrc.ac.uk/DHMHD/lddb.html
MaizeGDB: http://www.maizegdb.org
Mouse Anatomy:
http://genex.hgu.mrc.ac.uk/Databases/Anatomy
Mouse Genome Informatics:
http://www.informatics.jax.org
Mouse mutagenesis centres:
http://www.mgu.har.mrc.ac.uk
Nugene: http://www.nugene.org
OMIM: http://www.ncbi.nlm.nih.gov/omim
Rat Genome Database: http://rgd.mcw.edu
***Saccharomyces* Genome Database:**
http://genome-www.stanford.edu/Saccharomyces
Access to this interactive links box is free online.