BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii26–ii30
doi:10.1093/bioinformatics/bti1103

Databases

DPDB: a database for the storage, representation and analysis of polymorphism in the *Drosophila* genus

Sònia Casillas, Natalia Petit and Antonio Barbadilla*

Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

ABSTRACT

Motivation: Polymorphism studies are one of the main research areas of this genomic era. To date, however, no comprehensive secondary databases have been designed to provide searchable collections of polymorphic sequences with their associated diversity measures.

Results: We define a data model for the storage, representation and analysis of genotypic and haplotypic data. Under this model we have created DPDB, '*Drosophila* Polymorphism Database', a web site that provides a daily updated repository of all well-annotated polymorphic sequences in the *Drosophila* genus. It allows the search for any polymorphic set according to different parameter values of nucleotide diversity, linkage disequilibrium and codon bias. For data collection, analysis and updating we use PDA, a pipeline that automates the process of sequence retrieval, grouping, alignment and estimation of nucleotide diversity from Genbank sequences in different functional regions. The web site also includes analysis tools for sequence comparison and the estimation of genetic diversity, a page with real-time statistics of the database contents, a help section and a collection of selected links.

Availability: DPDB is freely available at http://dpdb.uab.es and can be downloaded via FTP.

Contact: antonio.barbadilla@uab.es

1 INTRODUCTION

Drosophila is the most intensively studied genus for DNA polymorphism, since current population genetics models on nucleotide variation have been tested using the extensive sequence data gathered for this genus (Aquadro *et al*., 2001; Powell, 1997). Each polymorphic study releases groups of homologous sequences (or haplotypes) for a given DNA region and species. The haplotypic information of a polymorphic set allows the estimation of both the one-dimensional and multi-dimensional components of nucleotide diversity in the studied regions. One-dimensional measures, such as the distribution of PI values [Nei's diversity index, (Nei, 1987)] along sliding windows, allow the detection of differently constrained regions (Vilella *et al*., 2005). Multi-dimensional diversity measures search for association among variable sites, as summarized by linkage disequilibrium estimators, and provide key information on the history and evolution of a DNA region, including the effective recombination rate underlying the region (Hudson, 1987; McVean *et al*., 2004; Nordborg and Tavare, 2002). Both diversity components are necessary for a complete description of nucleotide variation at the DNA level. To date, however, no comprehensive secondary database provides searchable collections of polymorphic sequences with their associated diversity measures. '*Drosophila* Polymorphism Database' (DPDB) is a database aimed to fill this vacuum, and allows the search of polymorphic sequences in the *Drosophila* genus according to different measures of nucleotide diversity.

2 DPDB APPROACH

The creation of a secondary database on DNA variation requires the development of a set of modules of data mining and analysis which operate together to automatically extract the available sequences from public databases, align them and compute the diversity estimates. A priori, the automation of this process seems destined to fail, since variation estimates usually require a careful manual inspection. Especially critical is the alignment of sequences (which is sensitive to the input parameters and the intrinsic characteristics of the sequences) and the sample stratification of aligned sequences (because any non-controlled heterogeneity will invalidate the estimates). On facing this 'manual versus automatic' dilemma, a first option would consist of giving up the automation and limiting the analyses to our own data. However, automation is nowadays an aspiration that cannot be waived. Therefore, while conscious of the limitations, we have tackled the bioinformatics automation of genetic diversity.

Our approach to build DPDB is outlined in Figure 1. We define a data model for the storage, representation and analysis of haplotypic variability based on the 'polymorphic set' as the basic storing unit: a group of homologous sequences for a given gene and species. Polymorphic sets are created by grouping by gene and species all the *Drosophila* sequences available in Genbank that are well annotated. From the sequence annotations, homologous subgroups are created for each polymorphic set corresponding to different functional regions (genes, CDSs, exons, introns, UTRs and promoters). Every subgroup is then aligned and selected according to different quality criteria. The selected alignments form the 'analysis units' of DPDB, on which the commonly used diversity parameters are computed. The results of the estimations are annotated together with the corresponding polymorphic set.

Besides such filtering during the processing, information on the data source and the quality of the alignments is given with the query output to allow the user assessment of the confidence on the estimated values. Furthermore, any subset can be directly reanalyzed using PDA (Casillas and Barbadilla, 2004) by adding or deleting sequences, or changing default parameters.

*To whom correspondence should be addressed.

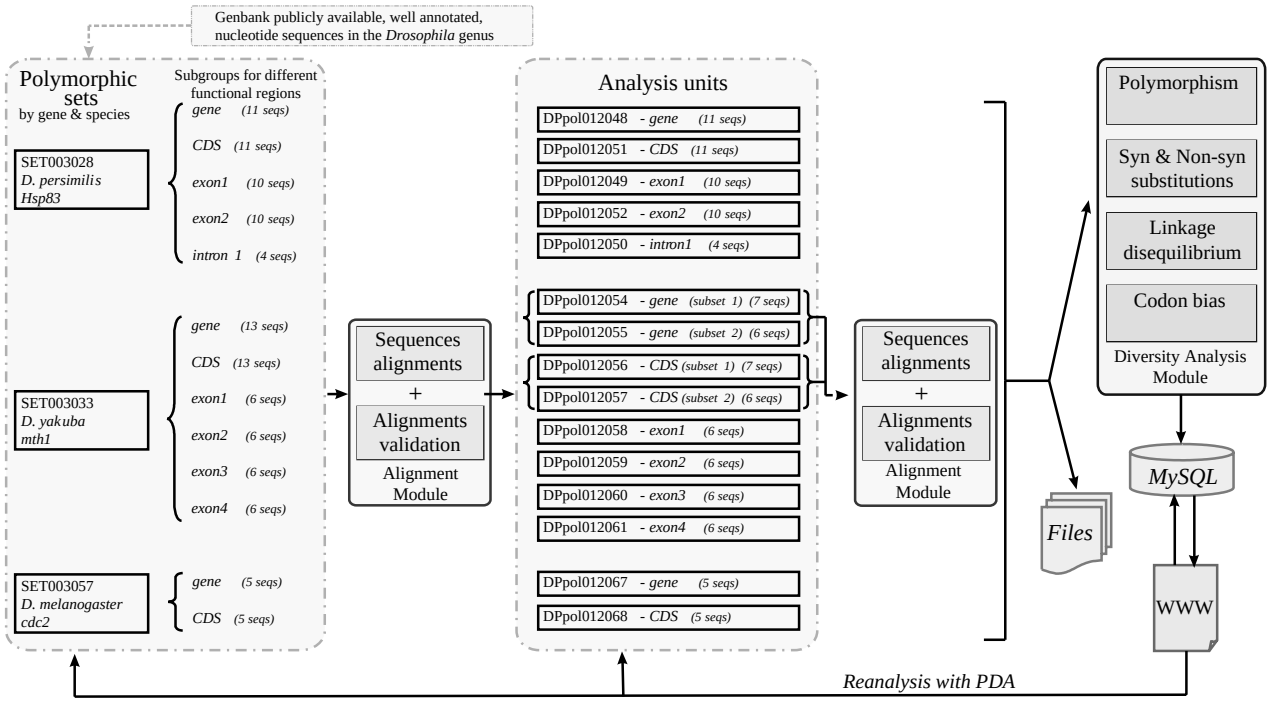

Fig. 1. The DPDB approach defines two basic data units: the 'polymorphic set' (a group of homologous sequences for a given gene and species) and the 'analysis unit' (on which diversity estimates are carried out). An analysis unit is a subset of sequences from a polymorphic set obtained by functional annotation and alignment filtering. Sequences are extracted from Genbank, and any subset can be directly reanalyzed using PDA. See text for details.

3 IMPLEMENTATION OF THE DATA MODEL IN DPDB

3.1 Overview of DPDB

DPDB is a secondary database which provides the collection of well-annotated polymorphic sequences in the *Drosophila* genus. DPDB allows, for the first time, the search for any polymorphic set according to different parameter values of nucleotide diversity, such as the PI value, the degree of linkage disequilibrium or the codon bias.

DPDB is searchable through a web site and also includes: (1) an Analysis section with tools for sequence comparison and the estimation of genetic diversity; (2) a daily updated Statistics page with the database contents; (3) a comprehensive Help section for the whole site; and (4) a page with selected links for the study of *Drosophila* polymorphism.

DPDB aims to be the reference site for DNA polymorphism in *Drosophila* (Galperin, 2005; Matthews *et al*., 2005), spanning studies that try to describe and explain the underlying causes of polymorphic patterns found in these species, such as recombination rate (Begun and Aquadro, 1992; Betancourt and Presgraves, 2002), gene density in different genomic regions (Payseur and Nachman, 2002), chromosomal inversions (Navarro *et al*., 2000), sequence complexity (Nelson *et al*., 2004) or demographic history (Glinka *et al*., 2003). DPDB has already been successfully used to study the association between coding polymorphism levels and gene structure in *Drosophila melanogaster* (Petit *et al*., unpublished data).

We want to guarantee long-term support for this database by including updating of the interface and new data processing and representation. We also aim to extend the database to other species groups.

3.2 Primary data source and processing

For data collection, diversity measures and updating we use PDA (Casillas and Barbadilla, 2004), a pipeline made of a set of Perl modules that automates the process of sequence retrieval, grouping, alignment and estimation of diversity parameters from sequences in large DNA databases. Using PDA we get all the publicly available *Drosophila* nucleotide sequences (excluding ESTs, STSs, GSSs, working draft and patents) with their annotations and references from Genbank (Benson *et al*., 2005), additional information of genes and aberrations from Flybase (Drysdale *et al*., 2005) and the cross-references to Popset [from NCBI (Wheeler *et al*., 2005)].

Polymorphic sets of two or more sequences are created by grouping sequences by gene and species. For each polymorphic set, subgroups of homologous sequences are created for the different functional regions (genes, CDSs, exons, introns, UTRs and promoters), as defined in the Features section of the Genbank format files. Note that those sequences lacking these annotations, even though coming from polymorphic studies, are not included in the analyses, so only well-characterized sequences are used. Every subgroup is then aligned with ClustalW (Chenna *et al*., 2003). After a manual inspection of hundreds of ClustalW alignments, we defined an optimal ClustalW parameter setting for *Drosophila* polymorphic data. Likewise, we fixed 95% as the minimum percentage of similarity between each pair of sequences within an analysis unit (excluding gaps), so that different analysis units can be obtained for a given functional region. Diversity measures are estimated on these analysis units, including polymorphism at synonymous and non-synonymous sites, linkage disequilibrium and codon bias [see Table 1 in Casillas and Barbadilla (2004) for a detailed description of all the estimations].

Sequences belonging to a polymorphic set can be either: (1) from previous polymorphism studies, or (2) from independent studies of the same gene and species, possibly not primarily focused on polymorphism. This second subset of sequences increases significantly the amount of polymorphic sets (figures are given in the Statistics section of the Web site); although it raises the question whether the estimations are reliable. We assess the confidence on each polymorphic set by taking into account the data source and the quality of the alignment. According to the data source, we use the following four criteria to determine if the study had a polymorphism goal: (1) one or more sequences from the alignment are stored in the Popset database; (2) all the sequences have consecutive Genbank accession numbers; (3) all the sequences share at least one reference; and (4) one or more references are from journals that typically publish polymorphism studies. To assess the quality of an alignment we use three other criteria: (1) the number of sequences included in the alignment; (2) the percentage of gaps or ambiguous bases within the alignment; and (3) the percentage of difference between the shortest and the longest sequences. Three qualitative categories are defined for each criterion: high, low and medium quality (for details on these criteria see the Help section in the Web site). Finally, alignments giving extreme polymorphism values are routinely checked. This allows us to continue improving the default parameters and to check data consistency.

3.3 Database structure, querying and output

The storage of diversity estimates in databases makes them permanently available and allows the reanalysis of all or part of the sequences. With this perspective in mind, we have created a relational MySQL database (see its structure in the Help section of the Web site) to store the results of the analyses. This database is centered on the two main storing units: the polymorphic set and the analysis unit (Fig. 1), and all the subsequent diversity data are annotated into different joined tables. The database also includes the *Drosophila* primary information retrieved from different external sources (Genbank, Flybase and Popset).

The database contents are updated daily, and records are assigned unique and permanent DPDB identification numbers to facilitate cross-database referencing: an increasing six-digit number is preceded by the string *SET* for polymorphic sets, *DPpol* for analysis units, *DPseq* for individual sequences, or *DPref* for references. Earlier analysis units are stored in separate tables when they are updated, and the later ones are assigned new identification numbers, so that the user can trace the history of a polymorphic set.

DPDB is accessible via web at http://dpdb.uab.es using a query interface based on SQL (Structured Query Language) searches (Fig. 2). The interface facilitates data interrogation by diversity estimates and the results can be filtered according to different confidence criteria established in DPDB (Fig. 2A). The first output page lists all the polymorphic sets by organism, gene, analyzed region and analysis unit showing additional information about the quality of the alignment, the confidence on the data source and the date of the last update (Fig. 2B). A complete report for each analysis unit can then be obtained through the corresponding link (Fig. 2C), as well as access to the primary database (individual sequences, genes, aberrations, references and polymorphic studies in the Popset database). Note that the alignment can be obtained in different formats, as well as the DND tree file, so that the user can revise it and decide if the estimates are reliable.

Furthermore, any analysis unit can be interactively reanalyzed using PDA. On using this option, the set of sequences is taken as input in the PDA submission page. Any subset of sequences can then be included or excluded from the analysis or the default parameters modified.

A Graphical search can also be performed for the different diversity values. A histogram is displayed on which any category can be queried to the database.

3.4 Analysis tools

The DPDB web site includes a set of analysis tools organized in different modules for sequence comparison and the estimation of genetic diversity. On the first module, three programs are available: (1) the Blast package (McGinnis and Madden, 2004) is implemented to search for homologous sequences in the primary DPDB database or in the *D.melanogaster* genome; (2) the ClustalW software (Chenna *et al*., 2003) is available with default parameters optimized for alignments of *Drosophila* polymorphic sequences (as manually checked); and (3) Jalview (Clamp *et al*., 2004) is implemented on the web to visualize and edit sequences alignments. The second module includes two other tools: (1) SNPs-Graphic allows performing analyses by the sliding window method, obtaining both the estimations in different regions of the alignment and graphic representations; and finally, (2) the PDA pipeline (Casillas and Barbadilla, 2004).

3.5 Statistics

The Statistics section summarizes the contents of both the primary and secondary databases. It is updated on a daily base, and includes tabular and graphic information.

The distributions of polymorphic sets according to different parameters, such as the species, genes and classes of genes [GO categories (Ashburner *et al*., 2000)] are shown. Then, the analysis units are classified according to the gene region, the quality of the alignments and the confidence on the data source. Average diversity estimates by gene region are also shown. The number of analysis units per taxon can be viewed in the 'Phylogeny of the *Drosophila* genus' graph (categories are based on the NCBI's taxonomy browser). Finally, some important statistics on the primary database are displayed, such as the total number of sequences, genes, aberrations and references, in different classifications.

At the time of writing this article, DPDB contained 1082 polymorphic sets, corresponding to 119 different species of the *Drosophila* genus and 587 different genes. A total of 2879 analysis units on these polymorphic sets were analyzed, most of them corresponding to the gene (1177), CDS (769), exon (473) or intron (435) regions.

The statistics on the quality of the alignments show that a high percentage of analysis units have <6 sequences, but that most of them have few gaps within the alignment, and that sequences are generally of similar length. Finally, according to the data source confidence, $\sim$50% of the analysis units come from sequences where polymorphism was the primary focus of the study.

3.6 Software details

DPDB is stored locally in a MySQL relational database running on a Windows 2003 Server, using the software IIS (Internet Information Server). It can be freely downloaded via ftp at ftp://dpdb.uab.es.

The web interface is mainly implemented in ASP and offers constant interfaces, standard file formats and *ad hoc* queries. Programs for data manipulation and search are all implemented in Perl modules, and search and analysis results are given in HTML formats.

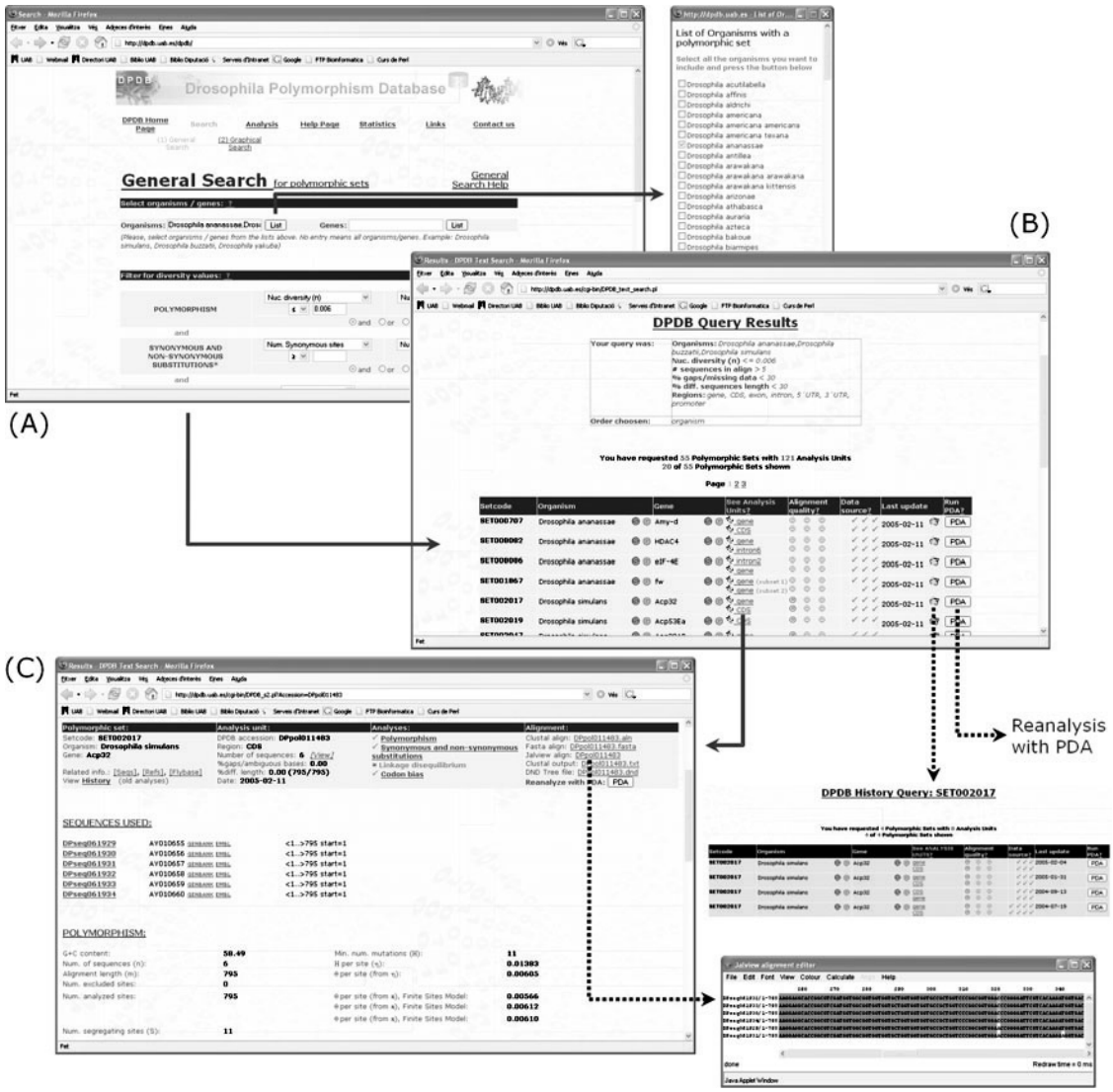

Fig. 2. DPDB interface: (**A**) the General Search page (with the species selector window), (**B**) the first output page of a query, and (**C**) a full report for an analysis unit. In this example we quired all the analysis units from the *Drosophila ananassae, Drosophila buzzatii* and *Drosophila simulans* species, having a nucleotide diversity value <0.006 and excluding lower quality alignments. At the time of making the figure, 55 polymorphic sets were found, each of them with different analysis units for the different gene regions. Part of the full report corresponding to the *CDS* region of gene *Acp32* in *D.simulans* in shown in (C), including general information about the analysis, the alignment in different formats, the corresponding sequences and some of the estimations. Sequences from the analysis units can be directly reanalyzed with PDA. Note that the history of a polymorphic set (the series of previous analysis that have been updated) can be queried from (B).

ACKNOWLEDGEMENTS

Authors would like to thank Raquel Egea and Jordi Pijoan for help in the manual revision of the alignments and testing the database, and Alfredo Ruiz for helpful discussions and comments. This work was funded by the Ministerio de Ciencia y Tecnología (Grants PB98-0900-C02-02 and BMC2002-01708). S.C. was supported by the Ministerio de Ciencia y Tecnología (Grant BES-2003-0416) and partially by the bioinformatics company Ebiointel, and N.P. by a grant from the Departament de Genètica i Microbiologia of the Universitat Autònoma de Barcelona.

Conflict of Interest: none declared.

REFERENCES

Aquadro,C.F. *et al.* (2001) Genome-wide variation in the human and fruitfly: a comparison. *Curr. Opin. Genet. Dev.*, **11**, 627–634.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Begun,D.J. and Aquadro,C.F. (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D.melanogaster*. *Nature*, **356**, 519–520.

Benson,D.A. *et al.* (2005) GenBank. *Nucleic Acids Res.*, **33** (Database issue), D34–D38.

Betancourt,A.J. and Presgraves,D.C. (2002) Linkage limits the power of natural selection in *Drosophila*. *Proc. Natl Acad. Sci. USA*, **99**, 13616–13620.

Casillas,S. and Barbadilla,A. (2004) PDA: a pipeline to explore and estimate polymorphism in large DNA databases. *Nucleic Acids Res.*, **32**, W166–W169.

Chenna,R. *et al*. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.

Clamp,M. *et al*. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.

Drysdale,R.A. *et al*. (2005) FlyBase: genes and gene models. *Nucleic Acids Res.*, **33** (Database issue), D390–D395.

Galperin,M.Y. (2005) The Molecular Biology Database Collection: 2005 update. *Nucleic Acids Res.*, **33** (Database issue), D5–D24.

Glinka,S. *et al*. (2003) Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics*, **165**, 1269–1278.

Hudson,R.R. (1987) Estimating the recombination parameter of a finite population model without selection. *Genet. Res.*, **50**, 245–250.

Matthews,K.A. *et al*. (2005) Research resources for *Drosophila*: the expanding universe. *Nat. Rev. Genet.*, **6**, 179–193.

McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20–W25.

McVean,G.A. *et al*. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.

Navarro,A. *et al*. (2000) Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in *Drosophila*. *Genetics*, **155**, 685–698.

Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nelson,C.E. *et al*. (2004) The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.*, **5**, R25.

Nordborg,M. and Tavare,S. (2002) Linkage disequilibrium: what history has to tell us. *Trends Genet.*, **18**, 83–90.

Payseur,B.A. and Nachman,M.W. (2002) Gene density and human nucleotide polymorphism. *Mol. Biol. Evol.*, **19**, 336–340.

Powell,J.R. (1997) *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press, New York.

Vilella,A.J. *et al*. (2005) VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics*, **21**, 2791–2793.

Wheeler,D.L. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33** (Database issue), D39–D45.