The Eukaryotic Promoter Database EPD

Rouaïda Cavin Périer, Thomas Junier and Philipp Bucher*

Swiss Institute for Experimental Cancer Research, Ch. des Boveresses 155, 1066-Epalinges s/Lausanne, Switzerland

Received September 25, 1997; Revised and Accepted October 3, 1997

ABSTRACT

The Eukaryotic Promoter Database (EPD) is an annotated non-redundant collection of experimentally characterised eukaryotic POL II promoters. The underlying definition of a promoter is that of a transcription initiation site. All information presented in EPD results from an independent evaluation of primary experimental data shown in the biological literature. Sequences flanking transcription initiation sites are indirectly given by pointers to EMBL sequences. The annotation part of a promoter entry includes description of the promoter-defining evidence, cross-references to other databases, and bibliographic references. Being designed as a resource for comparative sequence analysis, EPD is structured in a way that facilitates dynamic extraction of biologically meaningful promoter subsets. The database is available through the World Wide Web at URL http://cmpteam4.unil.ch

BACKGROUND

EPD originated as a by-product of a comparative sequence analysis project aimed at characterising transcriptional control signals near transcription start sites. Efforts to systematically compile data for such a project began in 1981 starting with a published list of 60 promoters (1). In 1986, an early version of this collection containing 173 entries appeared in this Journal (2). The first machine-readable version was released two years later in a format jointly designed with the EMBL Data Library (3) staff.

EPD was conceptually defined as a database of gene function, not as a sequence database. An important consideration was that promoter sequences (as any other type of sequences) will become available in the nucleotide sequence database anyway, and therefore need not be duplicated in a specialised database. However, it was also assumed that promoters would not be automatically annotated as promoter-defining evidence is often not published together with the sequences. Therefore, there is a need for a database which keeps track of promoter-defining evidence and links this information to sequences. These considerations led to the design of EPD as an annotated list of machine-readable pointers to transcription initiation sites in the EMBL Data Library. Obviously, such a design required co-ordinated updating procedures by the two databases involved, as the position-sensitive sequence pointers in EPD had to be adjusted, each time a corresponding EMBL sequence was modified. In this respect, EPD was an early experiment of biomolecular database interconnection which went beyond mere cross-referencing of documents from different collections (4).

EPD has been designed as a resource for comparative sequence analysis and so far has mainly served this purpose. For instance, it has played an instrumental role in the development of eukaryotic promoter prediction algorithms (5). Very recently, the scope of EPD has been extended in the framework of a European collaboration (the TRADAT project funded by the EU Biotechnology programme) aimed at developing an integrated sequence analysis system for transcription regulatory regions. This development also led to the design of a new format productively used since June 1997. The extensions include new links to other databases which can be exploited by novel data access procedures and user interfaces. With these improvements, EPD may serve a more diverse user community in the future.

LEADING CONCEPTS

Promoter definition

The underlying promoter definition of EPD is that of a transcription start site. Possible alternatives would have been to define promoters genetically as *cis*-acting elements determining the site and rate of transcription initiation, or biochemically as target sites of transcription factors. There are other databases providing information about promoters in the latter sense, for instance TRANSFAC and COMPEL (6). The experimental evidence required for a transcription start site consists of data that characterise the structure of the 5′ end of an RNA in enough detail such that it can unambiguously be mapped onto the genomic sequence. The underlying and broadly accepted assumption is that 5′ ends of eukaryotic transcripts are generated by transcription initiation rather than endonucleolytic cleavage.

Entry concept

An entry in EPD corresponds to a single biological object, not to an individual data report. The redundancy policy applied permits only one entry per genetic map position and organism. The taxonomic resolution is generally at the species level. All information pertaining to the same transcription initiation site in a genome is thus combined in a single entry. Note that there is no one-to-one relationship between EPD promoter entries and genes. In eukaryotic organisms, the same initiation site may be used for transcription of multiple genes and vice versa.

*To whom correspondence should be addressed. Tel: +41 21 692 5892; Fax: +41 652 6933; Email: pbucher@isrec-sun1.unil.ch

Organisation as a functional position set

A functional position set (FPS) is a machine readable list of pointers to positions in DNA sequences stored elsewhere, which can be used for automatic retrieval of fixed-length sequence segments around physiological sites (7). There were several motivations for such an approach. For one thing, it avoids data redundancy and thereby helps to maintain global data coherence. Verifying and adjusting position pointers to a new sequence database release is a relatively straightforward way to keep track of corrections and extensions of the promoter sequence data. More importantly, incorporation of sequence data into a promoter database would have implied an arbitrary choice of the 5′ and 3′ borders of a promoter region not based on experimental criteria. Access to sequences through an FPS enables the user to customise the length and location of the extracted sequence segments with regard to a particular biological question or data analysis procedure.

Physiological viewpoint

EPD is a database on gene function. Promoters are viewed as physiological sequence objects dependent on the correct interpretation by a *trans*-acting environment. The content of an entry is therefore restricted to information connected to the transcription initiation process. In accordance with this policy, the literature references given in EPD refer to transcript mapping data or reports on regulatory properties of the promoter, but never to mere sequence data. Furthermore, promoters are classified according to their cognate *trans*-acting environments (which make them promoters) rather than by the organisms which replicate them. Therefore, a promoter on a viral genome may be classified as a vertebrate promoter if it is interpreted by the host transcription machinery rather than virus-encoded factors.

Relationship to primary information sources

All information in EPD derives from independent examination and interpretation of experimental data presented in the cited research publications. The same standards are applied to data from different sources. As a consequence, the interpretations presented in EPD may differ from the conclusions reached by the authors of the data. Note that many transcription initiation sites described in the literature have not been included in EPD because the underlying experimental evidence did not meet the minimal requirement for inclusion, as stated in the user manual.

Maintenance policy

EPD entries are dynamic by design, not only because the pointers to transcription initiation sites in EMBL need to be verified and adjusted on a regular basis, but also because biological information on particular promoters accumulates continually. Changes in existing EPD entries are therefore not confined to error corrections. With the recent format extensions, maintenance of complete lists of cross-references to other databases has become a major objective.

Priorities

Being maintained with very limited resources, completeness was never considered a realistic objective. It is estimated that EPD presently contains only about half of the reported promoters formally qualifying for inclusion. By contrast, a great effort has been made to ensure accuracy of all information in the data collection. This remains a promise of EPD.

CONTENTS

In order to maintain a high quality standard, the scope of EPD has been limited in two ways: (i) by excluding certain classes of eukaryotic promoters, and (ii) by covering only certain aspects of promoters.

Admission criteria

To be included in EPD, a promoter has to fulfil all of the following five conditions:

(i) Polymerase system: the promoter must be recognised *in vivo* by RNA POL II.
(ii) Taxonomic range: the promoter must be active in a higher eukaryote; excluded are *phycophyta*, *fungi*, *myxomycetes* and *protozoa*.
(iii) Experimental data: the transcription initiation site(s) must be mapped with an accuracy of ±5 bp.
(iv) Physiology: the promoter must be biologically functional.
(v) Sequence data: the DNA sequence surrounding the transcription start site must be available from the EMBL Data Library.

The reasons for some of the restrictions may not be immediately obvious. The decision to exclude promoters from lower eukaryotes was taken because transcript mapping experiments were rarely carried out in these organisms, especially in yeast. The exclusion of POL I and POL III promoters seemed reasonable 10 years ago but is no longer justified now as we know that the same type of *cis*-acting elements may interact with more than one polymerase system. The experimental data criterion is strictly handled. For instance, single nuclease protection or primer extension data are generally not accepted, unless they are supported by additional experiments, e.g. demonstration of promoter activity in transiently transfected cells. On the other hand, data pertaining to the orthologous promoter of a closely related species is accepted, if the homologous promoters meet a stringent sequence similarity criterion. The requirement of function helps to deal with a number of pathological situations such as transcribed pseudogenes and promoters accidentally created by insertion of retroviral DNA.

Information content of an entry

An EPD entry contains the following types of information:

- Promoter identification and description.
- Machine-readable pointers to the transcription initiation site in corresponding sequence entries.
- Description of the experimental evidence defining the transcription start site.
- Various kinds of promoter classifications useful for extraction of biologically meaningful promoter subsets.
- Information on regulatory properties.
- Cross-references to other databases.
- Bibliographic references.

Each entry has an ID and an accession number. The description typically includes a gene or a gene product name. The exceptions are promoters used for transcription of several genes, such as the Adenovirus major late promoter.

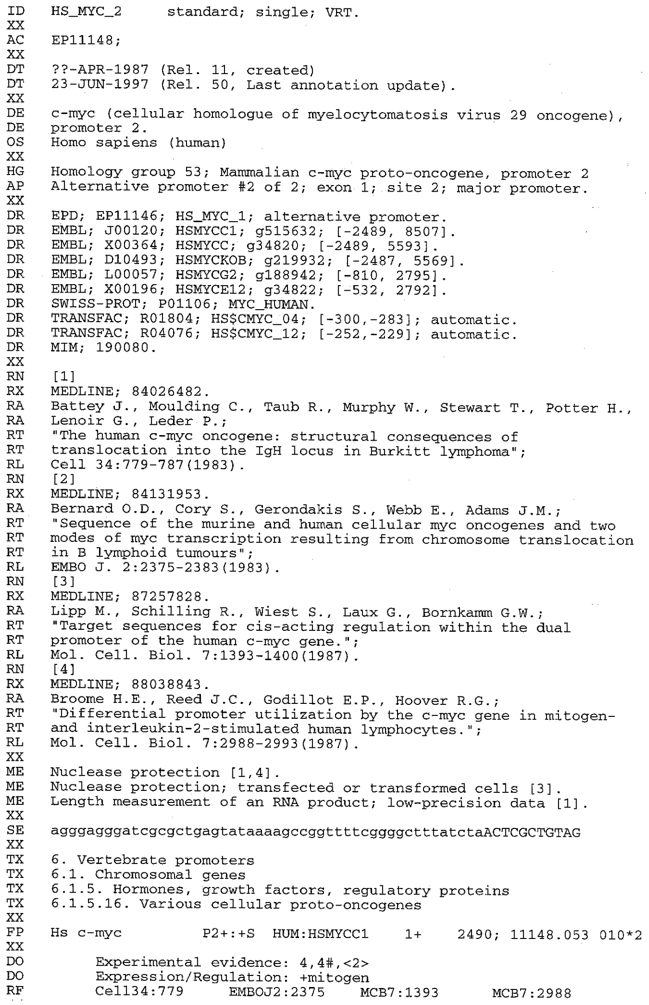
Table 1. Taxonomic breakdown of EPD Release 51

Total # of entries (# of independent entries)	1308	(861)
1. Plant promoters	179	(128)
1.1. Chromosomal genes	167	(117)
1.2. Prokaryotic plasmid DNA	8	(7)
1.3. Viral genes	4	(4)
2. Nematode promoters	11	(10)
2.1. Chromosomal genes	11	(10)
3. Arthropod promoters	163	(108)
3.1. Chromosomal genes	156	(101)
3.2. Transposable elements and retroviruses	2	(2)
3.3. Viral genes	5	(5)
4. Mollusc promoters	3	(3)
4.1. Chromosomal genes	3	(3)
5. Echinoderm promoters	42	(24)
5.1. Chromosomal genes	42	(24)
6. Vertebrate promoters	910	(588)
6.1. Chromosomal genes	750	(472)
6.2. Transposable elements and retroviruses	31	(13)
6.3. Viral genes	129	(103)

The machine-readable pointers to sequence data have four parts: a reference to an EMBL entry, a sign indicating plus or minus strand, a symbol indicating the topology of the sequence (linear or circular), and a sequence position number (the sequence topology is not redundant because the notion of a circular sequence is not exactly the same as in EMBL).

Based on the initiation patterns, three types of promoters are distinguished: (i) single initiation sites, (ii) clustered multiple initiation sites, and (iii) transcription initiation regions. Over 90% of all entries belong to the single initiation site class. Groups of promoters sharing over 50% sequence identity among each other are identified by a so-called homology group number. A subset of not closely related promoters (less than 50% identity) recommended for statistical analyses is marked by a special flag. All entries are embedded in a hierarchical classification system which helps to locate promoters of interest. The top level distinguishes between phyla (vertebrates, echinoderms, etc.); the second level between replicon types (chromosomes, transposable elements, viruses, see Table 1).

The information on regulatory properties was never collected in a systematic way and thus remained fragmentary. With the old format, there was space only for one database cross-reference, the machine-readable pointer to the corresponding EMBL sequence entry. A major objective of the current reorganisation is the introduction of many new cross-references to other data collections. So far, we have incorporated links to TRANSFAC (6), SWISS-PROT (8), Flybase (9) and MIM (10). In addition, MEDLINE identifiers have been added to the bibliographic references. In the future, we plan also to provide exhaustive cross-referencing to sequences in the EMBL Library as already exemplified by the entry shown in Figure 1.



```
ID   HS_MYC_2     standard; single; VRT.
XX
AC   EP11148;
XX
DT   ??-APR-1987 (Rel. 11, created)
DT   23-JUN-1997 (Rel. 50, Last annotation update).
XX
DE   c-myc (cellular homologue of myelocytomatosis virus 29 oncogene),
DE   promoter 2.
OS   Homo sapiens (human)
XX
HG   Homology group 53; Mammalian c-myc proto-oncogene, promoter 2
AP   Alternative promoter #2 of 2; exon 1; site 2; major promoter.
XX
DR   EPD; EP11146; HS_MYC_1; alternative promoter.
DR   EMBL; J00120; HSMYCC1; g515632; [-2489, 8507].
DR   EMBL; X00364; HSMYCC; g34820; [-2489, 5593].
DR   EMBL; D10493; HSMYCKOB; g219932; [-2487, 5569].
DR   EMBL; L00057; HSMYCG2; g188942; [-810, 2795].
DR   EMBL; X00196; HSMYCE12; g34822; [-532, 2792].
DR   SWISS-PROT; P01106; MYC_HUMAN.
DR   TRANSFAC; R01804; HS$CMYC_04; [-300,-283]; automatic.
DR   TRANSFAC; R04076; HS$CMYC_12; [-252,-229]; automatic.
DR   MIM; 190080.
XX
RN   [1]
RX   MEDLINE; 84026482.
RA   Battey J., Moulding C., Taub R., Murphy W., Stewart T., Potter H.,
RA   Lenoir G., Leder P.;
RT   "The human c-myc oncogene: structural consequences of
RT   translocation into the IgH locus in Burkitt lymphoma";
RL   Cell 34:779-787(1983).
RN   [2]
RX   MEDLINE; 84131953.
RA   Bernard O.D., Cory S., Gerondakis S., Webb E., Adams J.M.;
RT   "Sequence of the murine and human cellular myc oncogenes and two
RT   modes of myc transcription resulting from chromosome translocation
RT   in B lymphoid tumours";
RL   EMBO J. 2:2375-2383(1983).
RN   [3]
RX   MEDLINE; 87257828.
RA   Lipp M., Schilling R., Wiest S., Laux G., Bornkamm G.W.;
RT   "Target sequences for cis-acting regulation within the dual
RT   promoter of the human c-myc gene.";
RL   Mol. Cell. Biol. 7:1393-1400(1987).
RN   [4]
RX   MEDLINE; 88038843.
RA   Broome H.E., Reed J.C., Godillot E.P., Hoover R.G.;
RT   "Differential promoter utilization by the c-myc gene in mitogen-
RT   and interleukin-2-stimulated human lymphocytes.";
RL   Mol. Cell. Biol. 7:2988-2993(1987).
XX
ME   Nuclease protection [1,4].
ME   Nuclease protection; transfected or transformed cells [3].
ME   Length measurement of an RNA product; low-precision data [1].
XX
SE   agggagggatcgcgctgagtataaaagccggttttcggggctttatctaACTCGCTGTAG
XX
TX   6. Vertebrate promoters
TX   6.1. Chromosomal genes
TX   6.1.5. Hormones, growth factors, regulatory proteins
TX   6.1.5.16. Various cellular proto-oncogenes
XX
FP   Hs c-myc         P2+:+S  HUM:HSMYCC1    1+    2490; 11148.053 010*2
XX
DO        Experimental evidence: 4,4#,<2>
DO        Expression/Regulation: +mitogen
RF        Cell34:779     EMBOJ2:2375    MCB7:1393      MCB7:2988
```

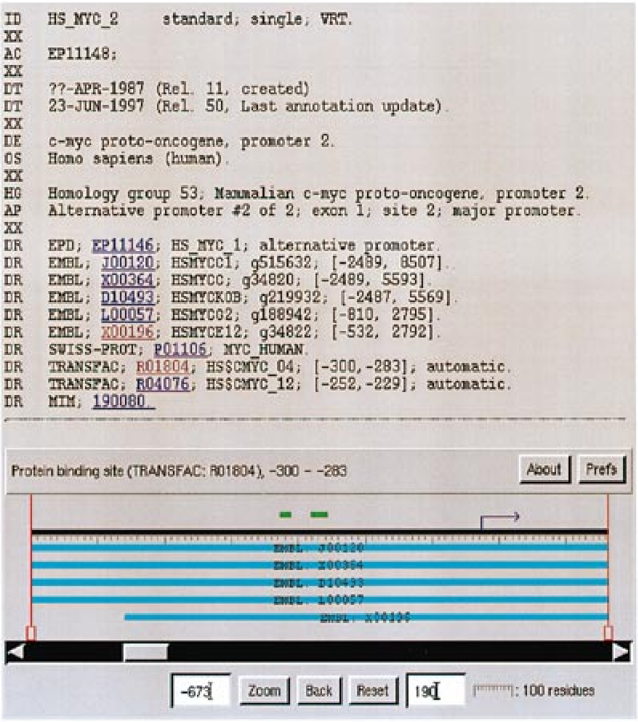
Figure 1. Example of an EPD entry.

FORMAT

EPD is distributed and maintained as a single ASCII flatfile. The format resembles those of the EMBL and SWISS-PROT sequence files. Each line starts with a line code identifying the type of information presented. The original EPD format, in which the database was distributed for almost 10 years, was very concise, relying on many abbreviations and providing certain types of information by alphanumeric codes rather then free text. As mentioned before, this format has recently been extended to allow representation of new types of information, but also to make it more human readable. An example of an entry in the new format is shown in Figure 1.

The old representation of an entry, which comprises the lines starting with FP, DO and RF, is included at the bottom of the new format ensuring that all software written for the old format will continue to work. The line types of the new format will briefly be explained now. (For a description of the old format, see the EPD user manual.)

The ID line type contains a unique entry identifier, specification of the initiation site type (single, multiple or region), and a



```
ID   HS_MYC_2     standard; single; VRT.
XX
AC   EP11148;
XX
DT   ??-APR-1987 (Rel. 11, created)
DT   23-JUN-1997 (Rel. 50, Last annotation update).
XX
DE   c-myc proto-oncogene, promoter 2.
OS   Homo sapiens (human).
XX
HG   Homology group 53; Mammalian c-myc proto-oncogene, promoter 2.
AP   Alternative promoter #2 of 2; exon 1; site 2; major promoter.
XX
DR   EPD; EP11146; HS_MYC_1; alternative promoter.
DR   EMBL; J00120; HSMYCC1; g515632; [-2489, 8507].
DR   EMBL; X00364; HSMYCC; g34820; [-2489, 5593].
DR   EMBL; D10493; HSMYCK0B; g219932; [-2487, 5569].
DR   EMBL; L00057; HSMYC02; g188942; [-810, 2795].
DR   EMBL; X00196; HSMYCE12; g34822; [-532, 2792].
DR   SWISS-PROT; P01106; MYC_HUMAN.
DR   TRANSFAC; R01804; HS$CMYC_04; [-300,-283]; automatic.
DR   TRANSFAC; R04076; HS$CMYC_12; [-252,-229]; automatic.
DR   MIM; 190080.
```

Protein binding site (TRANSFAC: R01804), -300 - -283

Figure 2. Elements of an EPD entry as represented by the *SEView* graphical sequence element viewer. The displayed objects correspond to cross-references with positional information. The blue arrow indicates the location and direction of the transcription initiation site corresponding to this promoter entry, green boxes are transcription factor binding sites described by TRANSFAC, and cyan lines represent EMBL entries whose sequences contain this region. It is possible to zoom on a particular region for greater detail, and to slide the display laterally. A simple click on an element brings up a short description in the upper panel (here, the user has just clicked on the first of the two green boxes), and a double click retrieves the corresponding entry as an HTML document. The applet can be tested live at http://www.epd.unil.ch/epd/doc_server.HTML

taxonomic division code (e.g. VRT for vertebrates). The AC, DE and OS lines, as well as the line types containing the bibliographic references, carry the same type of information as in EMBL or SWISS-PROT sequence entries. Note that an EPD accession number consists of the character string ‘EP’ followed by 5 digits. The HG line is optional. It contains a homology group number that allows identification of all sequence-wise similar promoters in EPD. The AP line provides information on alternative promoters of the same gene.

The DR lines contain cross-references to other databases. The precise format of these lines depends on the target database. Note that some cross-references include numbers indicating the relative position of a linked sequence object, or keywords characterising the nature of the relationship between the entries. For instance, the ranges associated with cross-references to EMBL entries define the extensions of the EMBL sequences relative to the initiation site described by the EPD entry. The multiplicity of EMBL cross-references in the example shown mirrors the redundancy of the sequence database. The positional information given on DR lines can be used for graphic display of the various sequence objects along the chromosome axis (Fig. 2).

The lines starting with ME describe experiments defining the transcription initiation site. In the new format, the experiments are individually linked to bibliographic references. The SE line shows a short sequence segment corresponding to the –49 to +10 region of the promoter. Transcribed and untranscribed nucleotides are represented by upper and lower case characters, respectively. This newly introduced line type is not meant to provide sequence data but serves as a control string for sequence extraction. The TX lines define a promoter’s location within EPD’s hierarchical classification system.

In the old format, all information that could be used for selection and retrieval of biologically meaningful promoter sequence subsets was given on a single line starting with the FP code. The current software behind our web pages still uses this part of the entry for FPS-dependent sequence retrieval. Future versions will use the new EMBL cross-references directly. The line on regulation and expression is the only part of the old format not yet replaced by a new representation.

During the last 5 years, EPD has not only been distributed in its original format but also in various alternative views. A ‘view’ is defined as a file that can be automatically derived from EPD and other public databases, and thus does not contain any genuinely new information. The distributed views included the widely used promoter sequence data files containing the –499 to +100 regions of all promoters in EPD.

ACCESS

EPD can be obtained via anonymous ftp from ftp.epd.unilich (directory: /pub/databases/epd). The following files are available:

- The EPD database in the original flatfile format
- Sequence containing views in EMBL and Fasta format. These files contain promoter sequences in the range from –499 to +100 relative to the transcription start site together with excerpts from the promoter annotation.
- A slightly reduced version of EPD in ASN.1 format designed for import into the GenBank-Entrez data environment (11).
- Documentation files including the EPD user manual and a formal data description of the ASN.1 version.

The URL for online access to EPD is http://www.epd.unil.ch/epd This site offers the following services:

- Access to EPD entries by ID or accession number. The following formats are offered: text only, HTML and HTML combined with a graphic representation of sequence objects by a Java applet (see Fig. 2).
- Access to EPD entries via a query form allowing for field-restricted character string searches.
- A page for downloading promoter sequence subsets defined in EPD (for instance all human promoters from 100 bases upstream to 100 bases downstream of the initiation site).

ACKNOWLEDGEMENT

EPD is funded in part by grant 95.0236-1 of the Swiss Federal Office for Education and Science.

REFERENCES

1 Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349–383.

2 Bucher,P. and Trifonov,E.N. (1986) *Nucleic Acids Res.*, **14**, 10009–10026.

3 Stösser,G., Sterk,P., Tuli,M.A., Stoehr,P.J. and Cameron,G.N. (1997) *Nucleic Acids Res.*, **25**, 7–13 [see also this issue (1998) *Nucleic Acids Res*. **26**, 8–15].

4 Fuchs,R. and Cameron,G.N. (1991) *Prog. Biophys. Mol. Biol.*, **56**, 215–245.

5 Fickett,J.W. and Hatzigeorgiou,A.G. (1997) *Genome Res.*, **7**, 861–878.

6 Wingender,E., Kel,A.E., Kel,O.V., Karas,H., Heinemeyer,T., Dietze,P., Knüppel,R., Romaschenko,A.G. and Kolchanov,N.A. (1997) *Nucleic Acids. Res.*, **25**, 265–268 [see also this issue (1998) *Nucleic Acids Res*. **26**, 362–367].

7 Bucher,P. and Bryan,B. (1984) *Nucleic Acids Res.*, **12**, 287–305.

8 Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res.*, **25**, 31–36 [see also this issue (1998) *Nucleic Acids Res*. **26**, 38–42].

9 Gelbart,W.M., Crosby,M., Matthews,B., Rindone,W.P., Chillemi,J., Russo Twombly,S., Emmert,D., Ashburner,M., Drysdale,R.A., Whitfield,E., *et al.* (1997) *Nucleic Acids Res.*, **25**, 63–66 [see also this issue (1998) *Nucleic Acids Res*. **26**, 85–88].

10 Pearson,P., Francomano,C., Foster,P., Bocchini,C., Li,P. and McKusick,V. (1994) *Nucleic Acids Res.*, **22**, 3470–3473.

11 Benson,D.A., Boguski,M., Lipman,D.J. and Ostell,J. (1994) *Nucleic Acids Res.*, **22**, 3441–3444 [see also this issue (1998) *Nucleic Acids Res*. **26**, 1–7].