

The Radiation Hybrid Database

Patricia Rodriguez-Tomé* and Philip Lijnzaad

EMBL Outstation, Hinxton—The European Bioinformatics Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 3, 1996; Accepted October 8, 1996

ABSTRACT

Since July 1995, the European Bioinformatics Institute (EBI) has maintained RHdb, a public database for radiation hybrid data. Radiation hybrid data are used in the generation of alternative genetic maps as they can include non-polymorphic markers and are also powerful enough to order unresolved genetic clusters of polymorphic STSs. The EBI is an Outstation of the European Molecular Biology Laboratory (EMBL).

INTRODUCTION

The radiation hybrid mapping technique orders genetic loci along a chromosome and estimates physical distances between adjacent loci. Radiation hybrids are produced by fusing irradiated donor cells with recipient rodent cells. These hybrid cells lines (grouped in 'panels' of identical radiation doses) each contain many large chromosome fragments produced by radiation breakage and are then screened by PCR amplification (producing 'scoring data') to identify those hybrids that have retained a given locus. Nearby loci will tend to show similar retention patterns thus allowing proximity to be inferred. This RH linkage method is a statistical procedure that requires some measure of the relative likelihood of the localization and order of the loci. This likelihood is expressed as a LOD score (1,2).

Radiation hybrid methods exploit differences between species (donor and recipient) and can be used to map both polymorphic markers such as sequence tagged sites (STS) or expressed sequence tags (EST). These maps are indispensable in the study of multifactorial diseases.

THE RADIATION HYBRID DATABASE

The Radiation Hybrid Database can be accessed on the World Wide Web at the URL <http://www.ebi.ac.uk/RHdb> (Fig. 1). This page provides information about the database and access to reports and query tools.

The first release of RHdb in July 1995 contained 1115 scoring data, and release 6 in August 1996 contained 28516 scoring entries, which represent a 25-fold increase in 1 year.

The RH mapping project

The radiation hybrid mapping project is a collaboration of a consortium of European and US Genome Centres that aims to

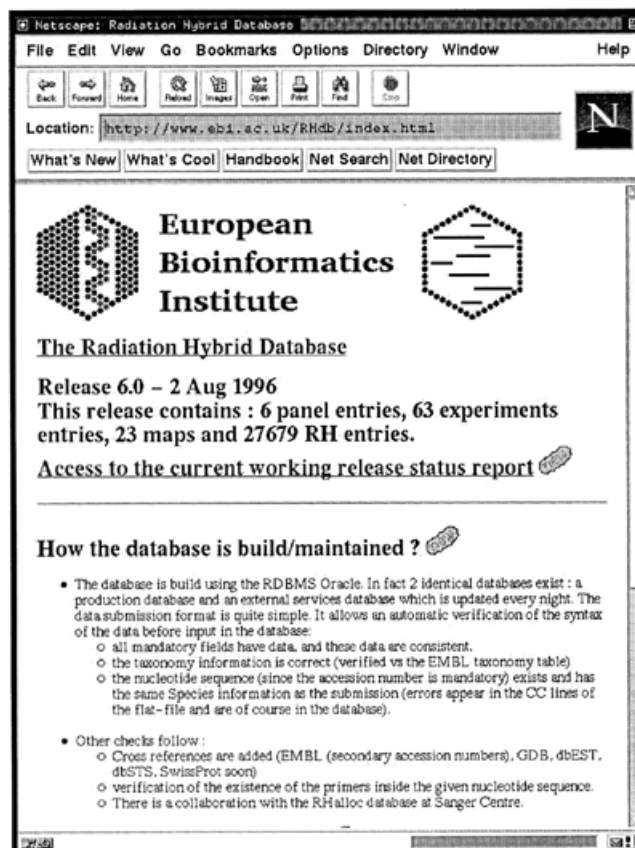


Figure 1. RHdb WWW home page: <http://www.ebi.ac.uk/RHdb>

build high resolution maps of the human genome integrating human cDNA with meiotically ordered polymorphic markers. In January 1995, this consortium asked the EBI to build a database for collecting the raw scoring data on radiation hybrid panels and the resulting maps (3,4). STSs belonging to one(or more) of the following categories are being used for mapping.

- (i) Random genomic human clones.
- (ii) CpG islands.
- (iii) Expressed sequences: STS were developed from cDNA sequences taken from the UniGene and UniEST collections (5), from Expressed Sequence Tags (6) in the dEST database (7) or the Genexpress index (8).

*To whom correspondence should be addressed. Tel: +44 1223 494409; Fax: +44 1223 494468; Email: tome@ebi.ac.uk

```

email : john.doe@ebi.ac.uk
//
DATA:
exp_id: E11
contact_id: C11
rh_species: Homo sapiens
rh_primer_a: GGCCCTCCTCCTATCTGACT
rh_primer_m: GGAAGCGGAGATAAATGC
rh_primer_length: 151
rh_embl_acc: Z56830
rh_db: RHalloc SC16370
rh_chrom: 22
rh_chrom_cc: Assigned with Coreill_1
rh_panel_name: Standford_G3
rh_test: 2
rh_score: 0100010000 0000001010 1010000010 0000111001
010100010 0000111001 0000011001 0101000101 0001111100
010101000 010
//

```

Figure 2. An example of an input file.

```

AC RH27783
XX
OS Homo sapiens
XX
DT 03-JUL-1996 (Rel. 5, Created)
XX
RN [1]
RA John Doe
RL Submitted (03-JUL-1996) to EBI by: John Doe, EBI
RL Hinxton, Cambridge CB10 1SD, UK
RL Email: john.doe@ebi.ac.uk
XX
DR EMBL; Z56830
DR RHalloc; SC16370
XX
PS GGCCCTCCTCCTATCTGACT
PS GGAAGCGGAGATAAATGC
PL 151 bp
PC E11
XX
CH 22
XX
SC Standford_G3_2
0100010000 0000001010 1010000010 0000111001 010100010 0000111001
0000011001 0101000101 0001111100 0101010000 010

```

Figure 3. A flat-file output example.

- (iv) Genetic markers from the Génethon genetic map (9) or developed at the Cooperative Human Linkage Center (10).
- (v) Chromosome specific libraries.

To avoid duplicating effort between mapping groups selecting the same ESTs sequences from the above resources, a database has been created at the Sanger Centre (UK). This database, RHalloc (11), can detect cases where different groups have selected the same EST sequence for mapping and can notify those groups.

While the data in RHdb comes mainly from this consortium, the database is built to be species independent.

The Relational Database

RHdb is maintained in the relational database management system (RDBMS) ORACLE. The schema of the database can be accessed at the URL <http://www.ebi.ac.uk/RHdb/schema.html>

The data

Each entry in the database is assigned an accession number which is a permanent unique identifier.

The main type of entry in the database contains the scoring results of a specific PCR hybridization on given PANELS using a specific STS (for which the primers sequences are given). This

type of entry is given an accession number of the form RHn, which serves as a permanent unique identifier.

There are three other primary types of entries in the database, for which accession number are widely used:

- (i) the hybrid cell lines grouped in PANELs (the name is used as accession number), with information about the authors, distributors and the clones used
- (ii) the maps constructed using the RH linkage programs (having an accession number of the form FwN) with the STSs order and LOD scores
- (iii) the PCR experimental conditions (accession number of the form EIn) in free text.

An RH entry also includes other types of information, each assigned a unique identifier:

- (i) the author identification (of form CIn)
- (ii) the corresponding bibliographic information (of form PIn) if available
- (iii) a 'flag' describing the type of STS used (EST, genetic marker, CpG island . . .).

Cross-references to other databases are systematically added when they are given by the submitter or can be inferred. To date, these databases include:

- (i) the nucleotide sequences databases (EMBL/GenBank/DBJ) accession number is mandatory (12)
- (ii) dbEST and dbSTS
- (iii) GDB (the Genome Database) (13)
- (iv) the submitting group's own unique identifiers
- (v) RHalloc (from the RH mapping consortium)
- (vi) Medline.

Data submission

The first purpose of this database is to collect large amounts of data to allow the construction of high resolution maps. Thus there is a need for effective acquisition methods. A data submission format of the form '*tag: text*' (Fig. 2) has been developed to allow fast formatting of the data by the submitter and fast input on the RHdb maintainer side. Each entry in the form ends with '//'. A full description of the format for each type of entry can be found at URL http://www.ebi.ac.uk/RHdb/rh_formats.html. A syntax verification program for this format is provided on the EBI anonymous FTP server at the URL ftp://ftp.ebi.ac.uk/databases/RHdb/softs/rh_submit.tar.gz

Entries should be submitted by e-mail to rhdb@ebi.ac.uk with 'datasub' as subject. Customised procedures are implemented for groups submitting very large amounts of data.

Once received, the data are processed through verification and checking programs:

- (i) the syntax of the taxonomy provided is verified
- (ii) the coherence of this taxonomy with the nucleotide sequence from which the primers have been constructed (as stated above, this sequence accession number is mandatory)
- (iii) the primers should be in the sequence, if not, the correct commentary is added
- (iv) cross-references between RHdb entries are added.

Well formatted submissions are processed within a day, and entries are immediately created. Accession numbers are returned to the author by e-mail.

For small amounts of data, a WWW submission form is being developed in collaboration with the HGMP resource centre (Hinxton, UK) that will allow automatic data verification during

Netscape: Radiation Hybrid Database : how to query ?

File Edit View Go Bookmarks Options Directory Window Help

Location: <http://www.ebi.ac.uk/RHdb/RHdbquery.html>

What's New What's Cool Handbook Net Search Net Directory

How to query the database ?

1. Searching the database using SRS
2. Query the relational database directly using forms :
 - o The maps
 - Chromosome :
 - o Get all entries which are and on chromosome (numeric value)
 - o Get an entry by using a cross-reference ID :
 - Database name :
 - Entry accession number :
3. Get the score data in various mapping programs input format.

Last updated: 03 June 1996 (PRT - Patricia.Rodriguez-Toms@ebi.ac.uk)
Send comments and questions to rhdb@ebi.ac.uk

Figure 4. An example of a query form.

Netscape: Radiation Hybrid Database : mapping program output

File Edit View Go Bookmarks Options Directory Window Help

Location: http://www.ebi.ac.uk/RHdb/vers_soft.html

What's New What's Cool Handbook Net Search Net Directory

Query the database and get the score data in various mapping programs input format.

- This form generates a script that queries directly the RH database and sends back the results by email.
- You can select for a given panel particular RH entries. The output will be sorted in the given order of the RH entries.
- If an RH entry does not correspond to the panel, it will be ignored in the search.
- Full chromosomes files per panel can be found in the [RHdb.FTP directories](#).

- Please give your email address :
- What program input format do you want ?
- Choose a Panel :
- Which RH entries ? (separate the entries with a space)

Figure 5. Getting the score data form.

the submission process, with respect to the data already in the database.

Data access

Each type of entry (RH, experimental conditions, panels and maps) is represented externally as an ASCII 'flat-file'. The flat-file (Fig. 3) is composed of lines beginning with a two character tag and followed by an associated text. The entry ends with '/'. These files are available on the EBI anonymous FTP server at URL <ftp://ftp.ebi.ac.uk/databases/RHdb> with the filenames:

- rh.dat* for the RH entries
- panel.dat* for the panel information
- exp.dat* for the experimental conditions
- map.dat* for the maps.

A full release of the database is made every two months, and new files replace the previous ones in the FTP directory. Between consecutive releases, files containing updates or new entries are also made available in the same directory.

Data query/retrieval

The EBI provides a query/retrieval system using SRS, the Sequence Retrieval System (14). This system allows entries to be retrieved based on a number of keywords. Specific query forms are accessible at the URL: <http://www.ebi.ac.uk/srs/src>. RHdb flat-files are indexed using SRS and can be queried that way (Fig. 4).

World Wide Web

Reports are automatically generated every night from the relational database. Written in HTML format, they can be accessed through the EBI WWW server at URL http://www.ebi.ac.uk/RHdb/rh_reports.html and give the most up to date view of the data. We are developing query forms that will provide answers to more specific questions (Fig. 4).

Mapping software

RHdb also provides the scoring data in the format used by the most common RH linkage programs: RHMAP (15,16), RHMAPPOR [Kruglyak,L., Slonim,D.K., Stein,L.D. and Lander,E.S. unpublished] and MULTIMAP (17). At each release, files are provided per chromosome, panel and program. A mail service with a WWW form front-end allows the user to specify entries to be formatted (Fig. 5). The result is sent back to the user by e-mail.

Future developments

Once a map is built, the data in RHdb is used by researchers to place an STS of interest on the map. For that purpose, access to large amounts of data is necessary. SRS is a powerful query tool that allows retrieving the information from inter-linked databases but does not provide such facilities needed for mapping. To

directly query the relational database, an external user needs on-site access to an ORACLE RDBMS and its network tools or to database gateways. This user will also need a good knowledge of the relation schema of the database. Any modification in that schema requires a modification of the query programs.

The EBI is in the process of rationalising its internal infrastructure using the Common Object Broker Architecture (CORBA) (18). RHdb will be part of this rationalisation and we will develop the necessary tools and make them available to external users. The use of this standard will also facilitate third party developments.

How to contact the RHdb at the European Bioinformatics Institute

Network: <http://www.ebi.ac.uk/RHdb> (World Wide Web)
<ftp://ftp.ebi.ac.uk/pub/databases/RHdb>
 (anonymous FTP server)
 datalib@ebi.ac.uk (for general enquiries,
 with RHdb in the subject)
 RHdb@ebi.ac.uk (for data submission to
 RHdb, with 'datasub' in the subject)

Postal address: RHdb
 EMBL Outstation—the EBI,
 The Wellcome Trust Genome Campus,
 Hinxton,
 Cambridge CB10 1SD,
 UK

Telephone: +44 (1223) 494401
 Fax: +44 (1223) 494468

ACKNOWLEDGEMENTS

We would like to thank C. Rice and P. Deloukas (Sanger Centre) for their advice and suggestions during the creation of the database, and G. Cameron for critical reading of this manuscript.

REFERENCES

- 1 Cox,D.R., Burmeister,E., Price,R., Kim,S. and Myers,R.M. (1990) *Science*, **250**, 245–250.
- 2 Walter,M.A., Spillet,D.J., Thomas,P., Weissenbach,J. and Goodfellow,P.N. (1994) *Nature Genet.*, **7**, 22–28.
- 3 Hudson,T.J. *et al.* (1995) *Science*, **270**, 1945–1954.
- 4 Gyapey,G. *et al.* (1996) *Hum. Mol. Genet.*, **5**, 339–346.
- 5 Boguski,M.S. and Schuler,G.D. (1995) *Nature Genet.*, **10**, 369–371.
- 6 Adams,M.D. *et al.* (1991) *Science*, **252**, 1651–1656.
- 7 Boguski,M.S., Lowe,T.M.J. and Tolstoshev,C.M. (1993) *Nature Genet.*, **4**, 332–333.
- 8 Houlgatte,R., Mariage-Samson,R., Duprat,S., Tessier,A., Bentolila,S., Lamy,B. and Auffray,C. (1995) *Genome Res.*, **5**, 272–304.
- 9 Dib,D., Fauré,S., Fizames,C., Samson,D., Drouot,N., Vignal,A., Millasseau,P., Marc,S., Hazan,J., Seboun,E., Lathrop,M., Gyapey,G., Morissette,J. and Weissenbach,J. (1996) *Nature*, **380**, 152–154.
- 10 Murray,J. *et al.* (1994) *Science*, **265**, 2049.
- 11 Durbin,R., Rice,C.M. and Durham,J. (1995) The Sanger Centre.
- 12 Rodriguez-Tomé,P., Stoehr,P.J., Cameron,G.N. and Flores,T.P. (1996) *Nucleic Acids Res.*, **24**, 6–12.
- 13 Fasman,K.H., Letovsky,S.I., Cottingham,R.W. and Kingsbury,D.T. (1996) *Nucleic Acids Res.*, **24**, 57–63.
- 14 Etzold,T. and Argos,P. (1993) *Comput. Appl. Biosci.*, **9**, 49–57.
- 15 Lange,K., Boehnke,M., Cox,D.R. and Lunetta,K.L. (1995) *Genome Res.*, **5**, 136–150.
- 16 Boehnke,M., Lange,K. and Cox,D.R. (1991) *Am. J. Hum. Genet.*, **49**, 1174–1188.
- 17 Matisse,T.C., Perrin,M. and Chakravasti,A. (1994) *Nature Genet.*, **6**, 384–390.
- 18 <http://www.omg.org/>