Nucleic Acids Research, 1999, Vol. 27, No. 1 123–127

IXDB, an X chromosome integrated database (update)

Ulf Leser, Hugues Roest Crollius[1], Hans Lehrach[2] and Ralf Sudbrak[2,*]

Technische Universität Berlin, Fachbereich 13, CIS Group, Einsteinufer 17, D-10587 Berlin, Germany, [1]Centre National de Séquencage, Genoscope, BP191, 2 rue Gaston Crémieux, 91000 Evry Cedex, France and [2]Max-Planck-Institut für Molekulare Genetik, Ihnestrasse 73, D-14195 Berlin, Germany

Received October 5, 1998; Accepted October 7, 1998

ABSTRACT

Chromosome specific databases are an important research tool as they integrate data from different directions, such as genetic and physical mapping data, expression data, sequences etc. They supplement the genome-wide repositories in molecular biology, such as GenBank, Swiss-Prot or OMIM, which usually concentrate on one type of information. The Integrated X Chromosome Database (IXDB, http://ixdb.mpimg-berlin-dahlem.mpg.de/) is a repository for physical mapping data of the human X chromosome and aims at providing a global view of genomic data at a chromosomal level. We present here an update of IXDB which includes schema extensions for storing submaps and sequence information, additional links to external databases, and the integration of an increasing number of physical and transcript mapping data. The gene data was completely updated according to the approved gene symbols of the HUGO Nomenclature Committee. IXDB receives over 1000 queries per month, an indication that its content is valuable to researchers seeking mapping data of the human X chromosome.

INTRODUCTION

The human X chromosome is associated with a large number of human diseases, partly due to its hemizygosity in males which reveals all recessive disorders. Therefore among all other chromosomes the X chromosome has been one of the most intensively studied. A high resolution clone and transcript map is of prime interest to localise, identify, and characterise such genes and to assist in the analysis of the genomic sequence.

Next to the chromosome spanning YAC maps constructed as part of whole genome mapping projects we constructed the first X chromosome specific YAC map (1) using a large scale hybridisation strategy, which serves as a backbone for the generation of the bacterial clone and transcript map. Based on radiation hybrid mapping, ESTs have been placed on the framework of the genetic map (2), and many groups are involved in large scale transcript mapping. In addition to these chromosome wide approaches a large number of regional mapping efforts generally carried out as part of positional cloning projects produce a vast amount of data. Also the sequencing of the chromosome is progressing in selected regions and involves, for instance, the Sanger Centre, Baylor College of Medicine and our group (see http://www.sanger.ac.uk/sHGP/ChrX/xchrom1.shtml).

The different data schemas and file formats used by these sources make it extremely difficult to get an integrated view on a comprehensive subset of the available data. IXDB has been developed to serve as a repository for the integration of this heterogeneous data. Therefore it is based on a flexible and easily extendible relational data schema and uses sophisticated data integration mechanisms (3).

IXDB was made publicly accessible in October 1996 in order to facilitate access to X chromosome specific data (4). The present report concentrates on improvements to and developments of IXDB in the last 12 months.

DATA REPRESENTATION

The data schema of IXDB is centred around genomic objects in a variety of classes including clone types, markers, ESTs, loci and genes. Several types of information are attached to each of these classes, such as class-specific annotations, experimental results, typed relations to other objects, external references to other databases and map positions. Each piece of information is tagged with its source, and different sources can provide contradicting values. If required, data can be kept confidential.

During the last year the IXDB data schema has experienced a number of smaller and two major changes.

As a first step towards integrating sequence information into the database the schema now allows the storage of external references to web pages containing sequence information. Currently IXDB stores WWW links to EMBL, Entrez, dbEST and the GSC in Jena. Hence it is now only one mouse click from an object to its sequence, assuming it is available.

Secondly the schema of IXDB was extended to store submaps. Our current definition of a map is recursive in the sense that a map itself can contain maps. The need for this emerged from the assembly of a chromosome-wide transcript map which is comprised of 18 different groups of the European X chromosome transcript consortium, each focusing on their region of interest. We model this situation in two phases: first, the data of each participating group is collected in a group-specific map. Additionally, all these partial maps are placed on a chromosome-wide map, where markers in common with a reference map are used to calculate the position. Compared with the conventional approach,

*To whom correspondence should be addressed. Tel: +49 30 8413 1612; Fax: +49 30 8413 1380; Email: sudbrak@mpimg-berlin-dahlem.mpg.de

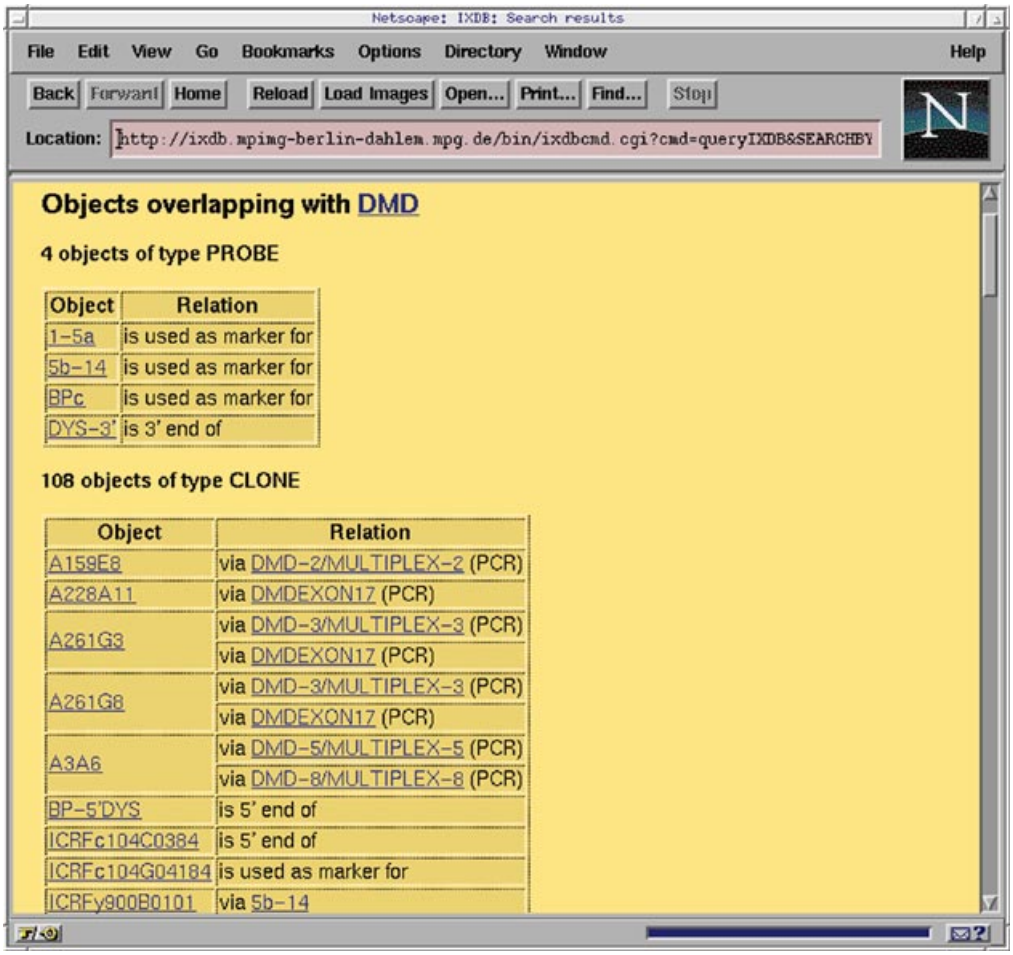

Figure 1. Objects overlapping with the gene DMD. For each overlapping object the relation to the queried object is shown. This can be either a direct connection or an overlap via an intermediate object.

this has several advantages: data is kept together by origin, entire submaps can be moved with one command, and contradictions can be detected without forcing a compromise in the global map. Recursive map definitions are also helpful if, for instance, different contig maps are to be assembled.

WORLD WIDE WEB ACCESS

IXDB is accessible via a World Wide Web (WWW) interface (http://ixdb.mpimg-berlin-dahlem.mpg.de/). Queries can be performed by name, keyword, external identification number (e.g. a GDB accession id), or by map position. The result will be the detailed IXDB object report on the appropriate object which contains all information stored in IXDB including general annotations, sequence information, external references, experimental results, related objects and map positions. By clicking on hyperlinks the user can easily navigate to connected objects or related data in other databases.

Most mapping experiments yield essentially the result that one object is contained in or overlapping with another one. Maps and contigs are calculated using large amounts of such information 'bits'. In general, the quality and level of detail of the maps is the better the more data is available, supposing a high level of quality. Using IXDB, the researcher already has the opportunity to access the largest set of mapping data for the human X chromosome that is publicly available. We now provide a method to exploit this information more easily. For each object, we offer the option to calculate the total set of other overlapping objects by automatically searching through all data sets stored in IXDB. This set is calculated as follows: given a starting object, X, first all markers are retrieved which are directly contained in or overlapping with X. This evidence can for instance be based on an experimental result, such as a PCR or a hybridisation, or on an observation, such as the fact that a probe is derived from the end of a clone. This set already contains the data from all sources in IXDB. For all markers in this set, such as ESTs, STSs, plasmids etc., we determine in a second step all other objects which also contain any of them. The union of both sets is the maximal set of objects which definitely have a physical overlap with X. An example WWW report of this procedure is given in Figure 1.

Maps can be either viewed as an HTML table or viewed and compared using derBrowser, a graphical map display tool implemented in JAVA. The applet is now also able to display submaps and by clicking on one the user can 'zoom' in and view the corresponding submap in the same way as the parent map. A new feature is the ability to compare two maps. Therefore, the applet reads all markers present on both maps, compares the

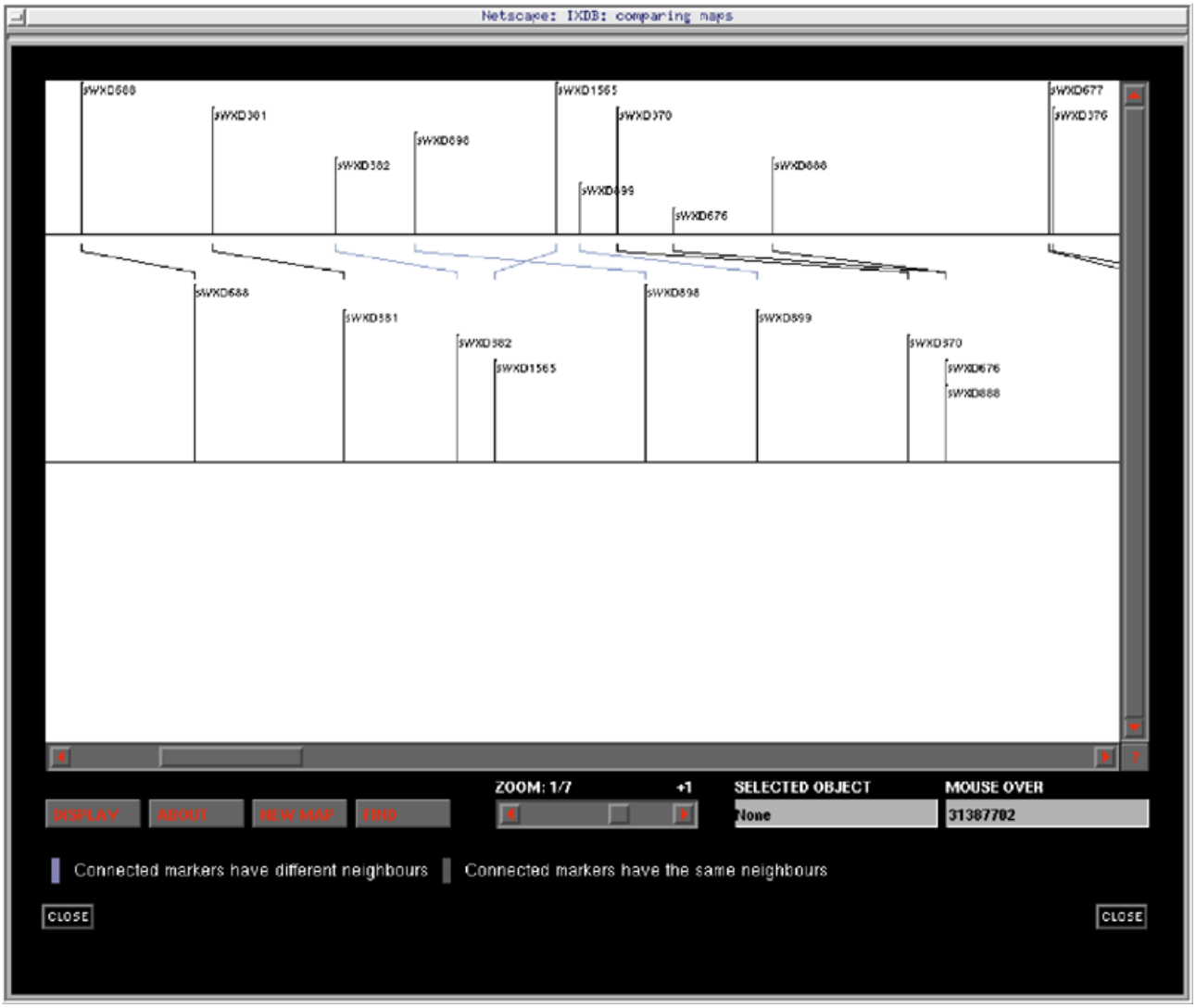

Figure 2. Comparison of the Washington University STS map and the Radiation Hybrid Transcript Map (2). Objects sWXD688 and sWXD381 each have the same neighbours to the left and right while the following four objects have different neighbours which is denoted through the color of their connecting line.

relative order of them according to their left and right neighbours, and marks differences (Fig. 2).

The web interface also offers the opportunity to display all objects stored in IXDB ordered by type. Each list consists of the alphabetically ordered object names as clickable hyperlinks and their descriptions (Fig. 3). Upon selecting an object the user is led to the IXDB object report.

CURRENT CONTENT

IXDB has integrated data from 16 regional and nine chromosome scale projects. During the last year a large amount of new data has been downloaded from public sources or submitted by the partners in the European X chromosome transcript consortium, which doubled the amount of data assembled in IXDB. Over 250 000 names will lead to a positive match giving information about 100 000 different DNA objects. Nine chromosome spanning maps and 11 contig maps can be viewed. Table 1 gives the exact number of objects currently stored in IXDB.

Table 1. Number of genomic objects and relationships

YACs	47 475
PACs	5871
BACs	1141
COSMIDs	1991
cDNAs	11 877
Genes	502
STSs	5521
ESTs	13 850
Connections by experimental evidence	58 903
Direct connections	28 908
Connected objects	74 640
Objects positioned on maps	8852

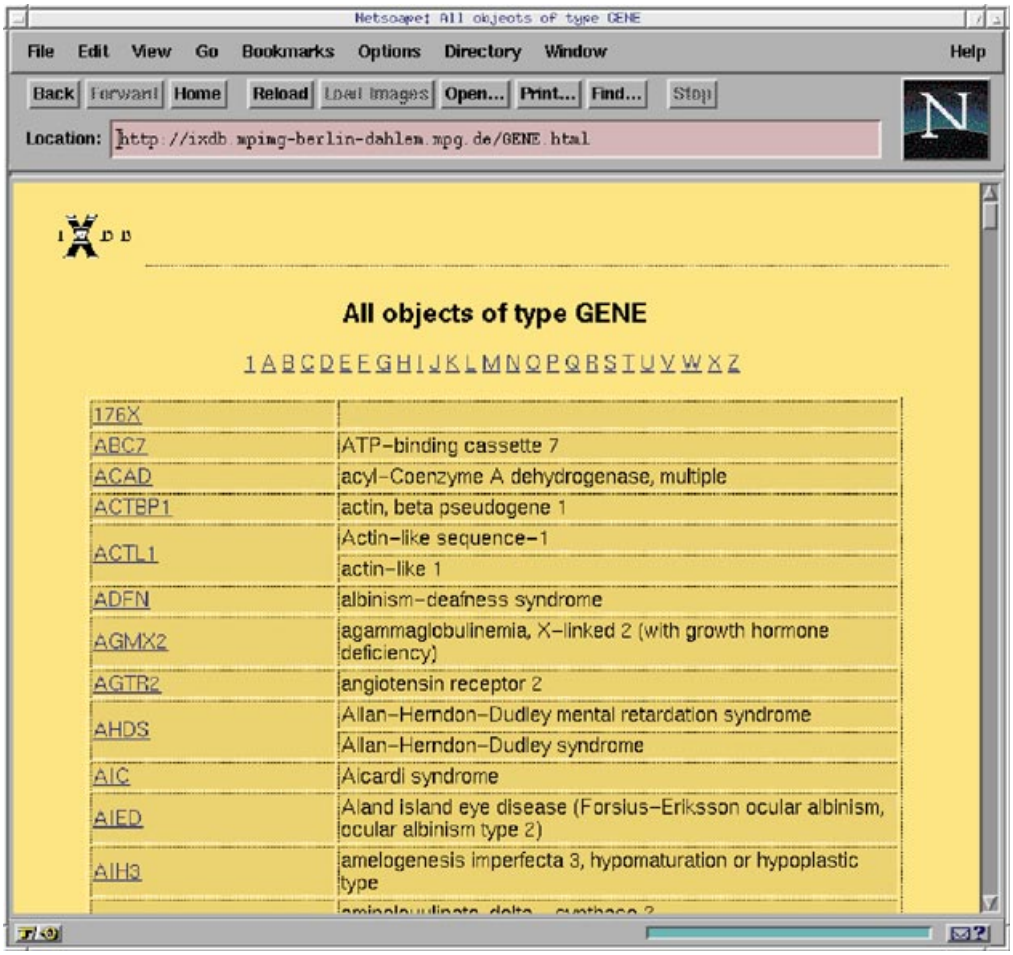

Figure 3. List of all genes stored in IXDB. By clicking on the gene name the user is led to the IXDB object report.

In 1998 the focus of the data integration was set according to the final goal of the whole project, the construction of a chromosome spanning bacterial clone and transcript map. The X chromosome mapping data from the Sanger Centre, consisting of numerous objects on BAC, PAC and cosmid level together with the corresponding chromosome spanning map, was downloaded and integrated into the existing data set. In order to continue the EST mapping the data from the ESTmap and the X chromosome related UniGene cluster have been added and cross-referenced with the source databases. In addition to that, the X chromosome specific data of the Human SNP Database including a genetic map from the Whitehead Institute was added.

IXDB has also received a fair amount of data from the members of the European X chromosome transcript consortium. Most of this is publicly available by appearance of this volume including the data from the GSC in Jena, the University Hospital Nijmegen, the ICGM CHU Cochin, the Department of Medical Genetics of the University of Helsinki, and the TIGEM. Other groups have also submitted data, but have chosen to keep their data private in IXDB until the middle of 1999.

Furthermore a lot has been done on curating and updating the existing data in IXDB. Data related to genes was updated according to the approved gene symbols of the HUGO Nomenclature Committee. This helped to clear the confusion raised by the existence of many different gene symbols for the same gene. However, these unofficial symbols still exist as synonyms and can be searched for in IXDB. Data from several other sources, such as CGM in St Louis was also updated. Links to Unigene, to the TIGR database, to the Imageclone database, to GeneCards, to the Human Gene Mutation Database and the Mouse Genome Database were added.

CONCLUSIONS

The announced (and postponed) termination of GDB (5), the human genome database, has raised the question of which is the best way to store and make available the overwhelming quantity of data produced in the human genome project. Although attention and funding has moved to sequencing and functional studies, mapping data still forms the information backbone of these techniques and will continue to do so. Many projects continue to rely on the availability of comprehensive and high-quality mapping data. We believe that chromosome-specific databases can play a dominant role in this respect. They are small enough to be administered by a single group, they do not require major commitments of funding agencies, and the amount of data can still be curated by a group of experts. On the other hand, they store information on a level of granularity that is sufficient for

projects which do not need to have access to the mapping data of the whole genome.

We see IXDB as a successful example for a chromosome specific database. To the best of our knowledge, no other single chromosome database is comparable to IXDB in terms of integration depth and width of scope. However, only a handful of people are actively involved in the administration and development of IXDB. We believe that this proves the feasibility of the 'chromosome-specific' strategy.

AVAILABILITY

IXDB is available at http://ixdb.mpimg-berlin-dahlem.mpg.de/ . Further information on used software or data submissions can be obtained from this webpage or via Email to: xteam@mpimg-berlin-dahlem.mpg.de. Please refer to this article if your research is assisted by the Integrated X chromosome database.

ACKNOWLEDGEMENTS

This work is supported by the EC grant CT961134 and BMBF grant 01KW9608.

REFERENCES

1 Roest Crollius,H., Ross,M.T., Grigoriev,A., Knights,C.J., Holloway,E., Misfud,J., Li,K., Playford,M., Gregory,S.J., Humphray,S.J. *et al.* (1996) *Genome Res.*, **6**, 943–955.
2 Schuler,G.D., Boguski,M.S., Stewart,E.A., Stein,L.D., Gyapay,G., Rice,K., White,R.E., Rodriguez-Tome,P., Aggarwal,A., Bajorek,E. *et al.* (1996) *Science*, **274**, 540–546.
3 Leser,U., Lehrach,H. and Roest Crollius,H. (1998) *Bioinformatics*, **14**, 83–90.
4 Leser,U., Wagner,R., Grigoriev,A., Lehrach,H. and Roest Crollius,H. (1998) *Nucleic Acids Res.*, **26**, 108–111.
5 Letovsky,S.I., Cottingham,R.W., Porter,C.J. and Li,P.W.D. (1998) *Nucleic Acids Res.*, **26**, 94–99.