104–107 Nucleic Acids Research, 2000, Vol. 28, No. 1 

# UK CropNet: a collection of databases and bioinformatics resources for crop plant genomics

**Jo Dicks, Mary Anderson[1], Linda Cardle[2], Sam Cartinhour[3], Matthew Couchman, Guy Davenport, Jeremy Dickson[4], Mike Gale, David Marshall[2], Sean May[1,*], Hamish McWilliam, Andrew O'Malia, Helen Ougham[4], Martin Trick, Sean Walsh[1] and Robbie Waugh[2]**

John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK, [1]NASC, Division of Plant Science (UP), University of Nottingham, Nottingham NG7 2RD, UK, [2]Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK, [3]USDA-ARS and Department of Plant Breeding, Cornell University, Ithaca, NY 14853, USA and [4]Institute of Grassland and Environmental Research, Plas Gogerddan, Aberystwyth, Ceredigion SY23 3EB, UK

Received September 4, 1999; Accepted October 4, 1999

## ABSTRACT

**The UK Crop Plant Bioinformatics Network (UK CropNet) was established in 1996 in order to harness the extensive work in genome mapping in crop plants in the UK. Since this date we have published five databases from our central UK CropNet WWW site (http://synteny.nott.ac.uk/ ) with a further three to follow shortly. Our resource facilitates the identification and manipulation of agronomically important genes by laying a foundation for comparative analysis among crop plants and model species. In addition, we have developed a number of software tools that facilitate the visualisation and analysis of our data. Many of our tools are made freely available for use with both crop plant data and with data from other species.**

## INTRODUCTION

UK CropNet was established in 1996, initially with 3 years funding from the BBSRC PAGA II initiative and latterly with continued BBSRC funding until 2003. UK CropNet is a collaborative project with a membership of six groups, each making a major contribution to plant genome analysis in the UK. The six groups are *Arabidopsis* (University of Nottingham), Barley and Potato (Scottish Crop Research Institute), Brassicas (John Innes Centre), Cereals (John Innes Centre), Comparative Genome Analysis (John Innes Centre) and Forage Grasses (Institute of Grassland and Environmental Research). The project is tripartite involving the creation and population of databases, the development of novel graphical interfaces (with special emphasis on comparative mapping) and training.

A suite of new graphical tools for data interpretation and display has been developed. The tools, which have been written in the Java language to give platform independence, have been designed to maximise their use both within UK CropNet and throughout the wider bioinformatics community. They are currently being interfaced to the UK CropNet databases in prototype applications. It is hoped that these will become publicly available in the near future, once thorough testing and optimisation has been carried out.

## UK CropNet CROP PLANT DATABASES

The six UK CropNet groups have developed eight UK databases using the ACEDB (http://www.sanger.ac.uk/Software/Acedb/ ) database management system (DBMS) (Richard Durbin, Sanger Centre, UK and Jean Thierry-Mieg, CNRS, Montpellier, France). ACEDB is a popular choice of DBMS for crop plant information worldwide and by using it we gain compatibility with many international data sources. The UK CropNet databases are the *Arabidopsis* Genome Resource (AGR), BarleyDB, BrassicaDB, CerealsDB, Comparative Mapping (ComapDB), Forage Grasses (FoggDB), MilletGenes and SpudBase. Table 1 gives a summary of the information held within each of these resources, together with the availability of each database.

The UK CropNet WWW server (http://synteny.nott.ac.uk/ ) is the central repository for the UK CropNet databases and project information. We have established pages to outline our project mission, to give contact details for our members and latterly, to download UK CropNet software (http://synteny.nott.ac.uk/software.html ) (see software section below). The database section (http://synteny.nott.ac.uk/db.html ) gives access to the five public UK CropNet databases and mirrors of nine USDA-funded crop plant databases. Access over the WWW currently uses the recently developed AceBrowser interface (http://stein.cshl.org/AcePerl/AceBrowser/ ) (Lincoln Stein, Cold Spring Harbor Laboratory, NY). In future, we will provide CORBA-based interfaces (see software section below), so that our users may carry out fully interactive interrogation of the databases. To supplement the ACEDB interfaces, we also provide various searching facilities both on our WWW pages and across the multiple databases simultaneously.

*To whom correspondence should be addressed. Tel: +44 115 951 3237; Fax: +44 115 951 3297; Email: sean_may@nasc.nott.ac.uk

**Table 1.** A summary of data published in the UK CropNet databases

| Database | Database contents | Where can I find this database? |
|---|---|---|
| **AGR** | Physical maps of chromosomes IV and V, Recombinant Inbred maps, DNA sequence (all publicly available Arabidopsis DNA sequence including 64 MB of canonical genome sequences, ESTs, BAC end sequences, complete mitochondrial genome, chloroplast sequences, BLAST analysis of approximately 40 MB of genome sequence against major databases), sequence contigs (anchored to RI maps) and germplasm resources. | UK CropNet WWW server<br>USDA mirror |
| **BarleyDB** | Nordic Barley database (59 trait studies and over 1100 germplasm accessions), EMBL and SwissProt sequences (1000 DNA and over 200 protein sequences for *Hordeum* species), bibliographic data and 25 barley linkage maps from other UK institutions. Data from European Barley Database (5 barley linkage maps, sequence data, primer data, trait study data, micro-malting data) will be added in Spring 2000. | UK CropNet WWW server<br>USDA mirror |
| **BrassicaDB** | Two genetic maps (derived from the segregation of 549 RFLP loci in two related populations), 2202 *B. napus* DNA sequences from EMBL (1709 ESTs and 493 genomic sequences coding for 276 different types of identified gene products), results of BLASTX analyses (32,167 proteins showing significant homology) of EST sequence data from the SwissProt/TrEMBL database and bibliographic information on 6,354 papers related to *B. napus*. | UK CropNet WWW server<br>USDA mirror<br>JIC Bioinformatics server (http://jiio16.jic.bbsrc.ac.uk) |
| **CerealsDB** | 2400 RFLP probes (insert sizes, copy numbers, chromosomal locations and polymorphism data with additional data on DNA end-sequence, putative gene function, primer sequences and PCR amplification conditions for the wheat anchor set), scanned and annotated images (DNA hybridization patterns and garden blots), genetic maps (wheat, rye and *Aegilops umbellulata*) and the corresponding segregation data, the JIC wheat germplasm and pedigree database, over 40,000 cereal sequences from EMBL and a catalogue of probe and worldwide locus naming conventions as a step towards unifying map nomenclature. | Selected information soon to be public… |
| **ComapDB** | Comparative mapping links between various monocots (including barley, foxtail millet, rice, wheat, maize, pearl millet, sugar cane, sorghum and oats) and dicots (arabidopsis and brassicas), conserved chromosomal segments of various monocots (relative to rice) and bibliographical data. | Soon to be public…<br><br>However, a prototype version is currently available on the JIC Bioinformatics server |
| **FoggDB** | Genetic maps, loci, sequences, associated data (including phenotypes, references and colleagues), trait scores, QTLs and images (phenotypes, disease symptoms, recombinant chromosomes for introgression mapping) for temperate forage grasses of importance to UK agriculture (ryegrasses and fescues) and also for tropical grasses possessing useful traits. | UK CropNet WWW server<br>USDA mirror |
| **MilletGenes** | Mapping data (11 pearl millet crosses, two foxtail millet crosses and a finger millet cross) with Downy mildew QTLs for all maps, RFLP and STS polymorphism data for all pearl millet probes, scanned autoradiographs (for RFLP probes) and gels (for sequence-tagged-site markers), probe insert sizes, copy numbers and locations for pearl millet and foxtail millet probes, DNA end-sequences, primer sequences and PCR amplification conditions for a subset of pearl millet probes and a set of JIC AFLP profiles. | UK CropNet WWW server<br>USDA mirror<br>JIC Bioinformatics server |
| **SpudBase** | 5 potato linkage maps, EMBL and SwissProt sequences (over 800 DNA and 200 protein sequences from *Solanum tuberosum*) and bibliographic data. | Soon to be public… |

In addition to being accessible through the UK CropNet WWW server, the databases may be accessed through the website of the USDA-ARS Center for Bioinformatics and Comparative Genomics (http://genome.cornell.edu/ ). In a reciprocal agreement, nine USDA crop plant databases are mirrored at the UK CropNet site. This provides faster access to these databases for UK users and also provides a cross-querying search facility in which the UK CropNet and USDA databases can be queried singly or in combination.

## UK CropNet SOFTWARE

We have adopted a Java/CORBA framework for our software, to enable cross-platform applicability. We endeavour to write our software so that it can be interfaced to any data source and consequently used by a wider community. We have written Java graphical components that will give the user considerable power for both general and fine-grained comparative analyses of many different types of data. Our Circles Engine (http://jiio8.jic.bbsrc.ac.uk/bioinformatics/developers/miscellaneous/circleengine/index.html ) has a special multi-layered software specification that enables it to display entities that have a ring structure. One instance of it is the Circular Genome Display (CGD) (http://jiio8.jic.bbsrc.ac.uk/bioinformatics/developers/comparative/cgmd/index.html ) (e.g., Fig. 1) which is a dynamic, interactive version of the display seen in a series of papers by Moore and colleagues and is the 'gateway' to our other comparative displays. CGD has a new CORBA interface to the ComapDB database and consequently displays may be drawn 'on-the-fly' within a user's WWW browser

The GridMap (http://synteny.nott.ac.uk/gridmap/grid-top.htm ) is a powerful generic tool that represents similarities and differences between pairs of objects, including genomes and sequences, in a grid form. It is particularly useful for graphical

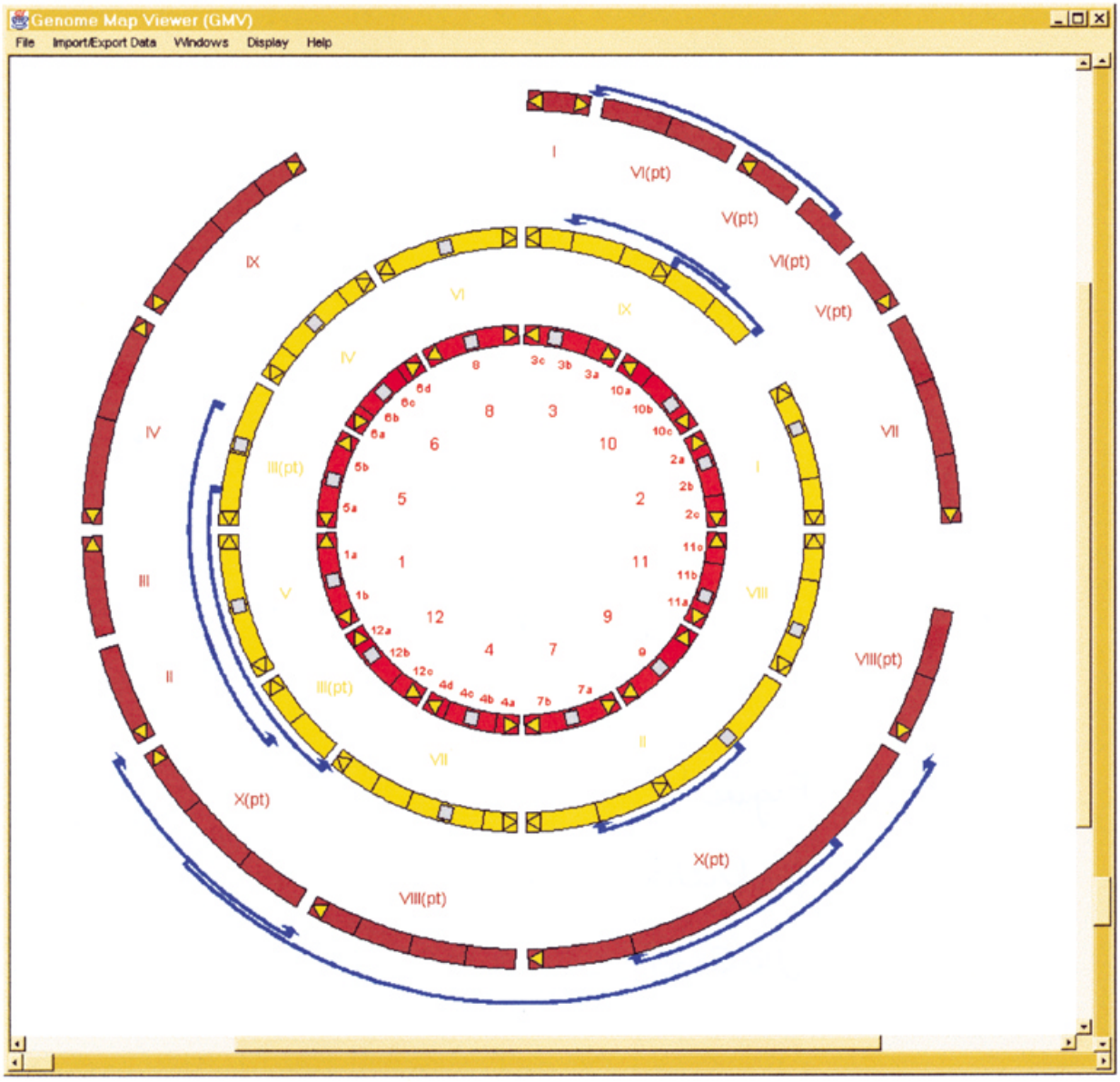


**Figure 1.** A Circular Genome Display of rice, foxtail millet and sugar cane.

analysis of comparative mapping data and can be used to represent displays such as Oxford Grids and Species Grids. The Recombinant Viewer (http://synteny.nott.ac.uk/software.html ) is a tool for analysing genetic segregation data from a mapping population. By applying user-definable rules to colour-code symbols according to their context it can highlight areas of special interest such as double crossover events or highly heterozygous plants. The flexibility of data type built in to this viewer also enables it to be used as a tool to display multiple DNA sequence alignments. 'Gel Sketch' is a simple tool which displays a representation of a gel when given a table of bands with associated molecular weights. The Pairwise Comparative Map (PCM) (http://synteny.nott.ac.uk/pairwise/Welcome.html ) displays two maps of any type and draws lines between the homologous loci on each map (Fig. 2). Maps may be resized, inverted or swapped. Clicking on a locus highlights it along with any homologues on the same map (paralogues) and on the opposite one (orthologues). The Multi PCM allows a number of PCMs to be placed in a scrollpane. Highlighting a locus in one PCM has the effect of highlighting all the homologues across all of the PCMs. The maps are colour co-ordinated across the Multi PCM to unify the maps originating from the same database. A new prototype application in Java 3D enables us to visualize high-density genomic data. We have developed a Perl module, and more recently a Java interface, to carry out Quick and Dirty (QAD) genetic mapping. This can rapidly map new markers using a simplified approach based on contextual comparison with existing scoring data. All these

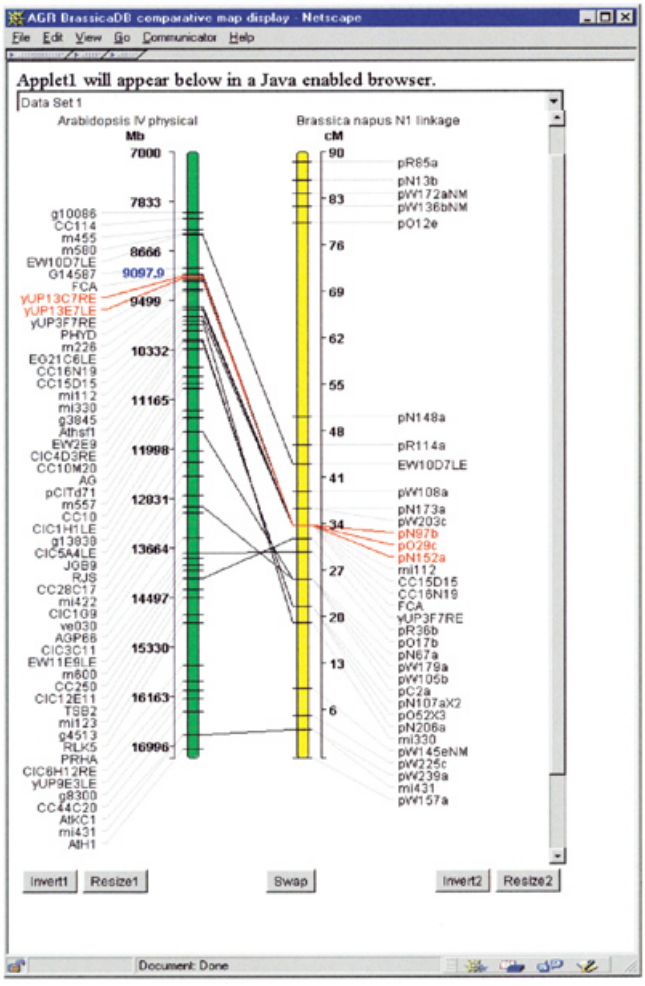


**Figure 2.** A Pairwise Comparative Map of the physical map of *Arabidopsis* chromosome IV and the genetic map of *Brassica napus* N1.

displays have been designed to give maximum power for displaying UK CropNet data but with the wider community also in mind. Consequently, our displays are configurable and are made publicly and freely available through the UK CropNet software repository.

Inter-operability within UK CropNet, to facilitate comparative mapping and genomics, centres on our newly developed CORBA interface to ACEDB (CITA) (http://jiio16.jic.bbsrc.ac.uk/BrassicaDB/CITA/index.html ). This provides the infrastructure for cross-database querying. Prototype CITA servers have been written in Perl and C that allow simultaneous connections to multiple ACEDB databases. In addition, a general-purpose Java client (GFace) that is launched through a user's WWW browser has been developed. This allows both simple and advanced querying, the display of simple text objects as well as the invocation of UK CropNet graphical displays such as the PCM and GridMap. GFace thus reinstates, and enhances, for the remote user much of the graphical interactivity that local use of an ACEDB database offers and which, to some extent, has been lost in the Webace and AceBrowser implementations. A new tool, ARCADE (A Real-time Comparative Analysis Display Environment), will use CITA to facilitate querying of multiple databases, so that complex comparative queries and analyses may be performed. The functional specification of ARCADE is currently being written in collaboration with other interested parties. Our data interfaces will use the forthcoming standards to be made by the OMG's Life Sciences Special Interest Group.

## FUTURE INITIATIVES

### Access to the databases

We will enable CITA access (through the Gface client) to both the UK CropNet databases and the USDA crop plant databases at the UK CropNet WWW site.

### User-training

We will shortly begin training courses for current and potential users of the UK CropNet resources. Training will begin with a widely publicised one-day dissemination event to alert potential users to the databases and resources of UK CropNet and those of our sister UK Animal Bioinformatics Network (http://www.ri.bbsrc.ac.uk/bioinformatics/databases.html ). Following this, two-day training meetings will be held at UK CropNet sites at three-monthly intervals. Users will register for their site of interest (for their species of interest) 3 months in advance. Users with pertinent data to be input will be particularly encouraged to register and attend.

## ACKNOWLEDGEMENT

We would like to thank the BBSRC for their continued support of this project.