156–158 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg071

CATMA: a complete *Arabidopsis* GST database

Mark L. Crowe[1,2,*], Carine Serizet[1,2], Vincent Thareau[1,3], Sébastien Aubourg[3], Pierre Rouzé[1], Pierre Hilson[2,3], Jim Beynon[4], Peter Weisbeek[5], Paul van Hummelen[6], Philippe Reymond[7], Javier Paz-Ares[8], Wilfried Nietfeld[9] and Martin Trick[1]

The John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK, [1]Laboratoire associé de l'INRA (France) and [2]Department of Plant Systems Biology, VIB, University Ghent, K. L. Ledeganckstraat, B-9000, Gent, Belgium, [3]Unité de Recherche en Génomique Végétale (URGV) INRA-CNRS, 2 rue Gaston Crémieux, 91057 Evry Cedex, France, [4]Department of Plant Biotechnology, Horticultural Research International, Wellesbourne, Warwick CV35 9EF, UK, [5]Department of Molecular Genetics, University of Utrecht, Padualaan 8, 3584CH Utrecht, The Netherlands, [6]MicroArray Facility, Flanders Interuniversity Institute of Biotechnology (VIB), Herestraat 49, 3000 Leuven, Belgium, [7]Gene Expression Laboratory, Institute of Ecology, University of Lausanne, 1015 Lausanne, Switzerland, [8]Department of Plant Molecular Genetics, Centro Nacional de Biotecnologia, Campus de Cantoblanco, 28049-Madrid, Spain and [9]Max-Planck-Institute for Molecular Genetics, Department Lehrach-Vertebrate Genomics, Ihnestrasse 73, D-14195 Berlin-Dahlem, Germany

Received August 12, 2002; Accepted October 14, 2002

ABSTRACT

The Complete Arabidopsis Transcriptome Micro Array (CATMA) database contains gene sequence tag (GST) and gene model sequences for over 70% of the predicted genes in the *Arabidopsis thaliana* genome as well as primer sequences for GST amplification and a wide range of supplementary information. All CATMA GST sequences are specific to the gene for which they were designed, and all gene models were predicted from a complete reannotation of the genome using uniform parameters. The database is searchable by sequence name, sequence homology or direct SQL query, and is available through the CATMA website at http://www.catma.org/.

INTRODUCTION

The Complete Arabidopsis Transcriptome MicroArray (CATMA) project (http://www.catma.org/) was formed in 2000 and now consists of groups from eight European countries. Our aim was to use the newly completed *Arabidopsis thaliana* genome sequence to develop a complete and specific microarray for *A. thaliana* by producing a specific gene sequence tag (GST) for every known or predicted gene found in the genome sequence, at the time believed to be 25 498 genes (1) and currently 29 084 genes (http://www.tigr.org). We believe that this approach will overcome many of the drawbacks of the use of ESTs and cDNAs as microarray probes: in particular that because the complete sequence of an EST clone is rarely known, their specificity to a particular gene cannot be guaranteed, and that known ESTs may represent only a fraction of the genes identified in a eukaryotic genome. The former is particularly important for genes which belong to gene families, around 65% of all *Arabidopsis* genes (1), where a full length clone may cross-hybridize to other family members.

We have designed the CATMA GSTs to be specific only to their target gene. Furthermore, because the full sequence of each GST is known, they can be used not only for microarray experiments but also for purposes such as RNAi work. This is aided by the introduction of extension primers onto each GST, allowing reamplification of, and introduction of cloning sites to, any GST using one pair from only twenty-four 3′ and sixteen 5′ PCR primers.

To track the ~30 000 GSTs and primer pairs expected to be generated by the project, and to allow easy dissemination of data about the GSTs, we developed a web-interfaced MySQL-driven database. We made this database publicly available in June 2002 and here, we describe the generation of the data and some of the features of the database.

GENE PREDICTION AND GST DESIGN

The data in the CATMA database is based primarily on a complete reannotation of the *Arabidopsis* genome sequence, which was carried out by CATMA members to overcome problems with the accuracy and consistency of the AGI annotation at the time (2). We performed this reannotation using the EuGène gene prediction software (3) to generate a uniform gene prediction model on which to design the CATMA GST set. EuGène was chosen after a survey of gene prediction software (4) as being the most suitable for the CATMA project: in particular EuGène rarely merges two genes into one, tending instead to split one gene into two. This

*To whom correspondence should be addressed. Email: mark.crowe@bbsrc.ac.uk

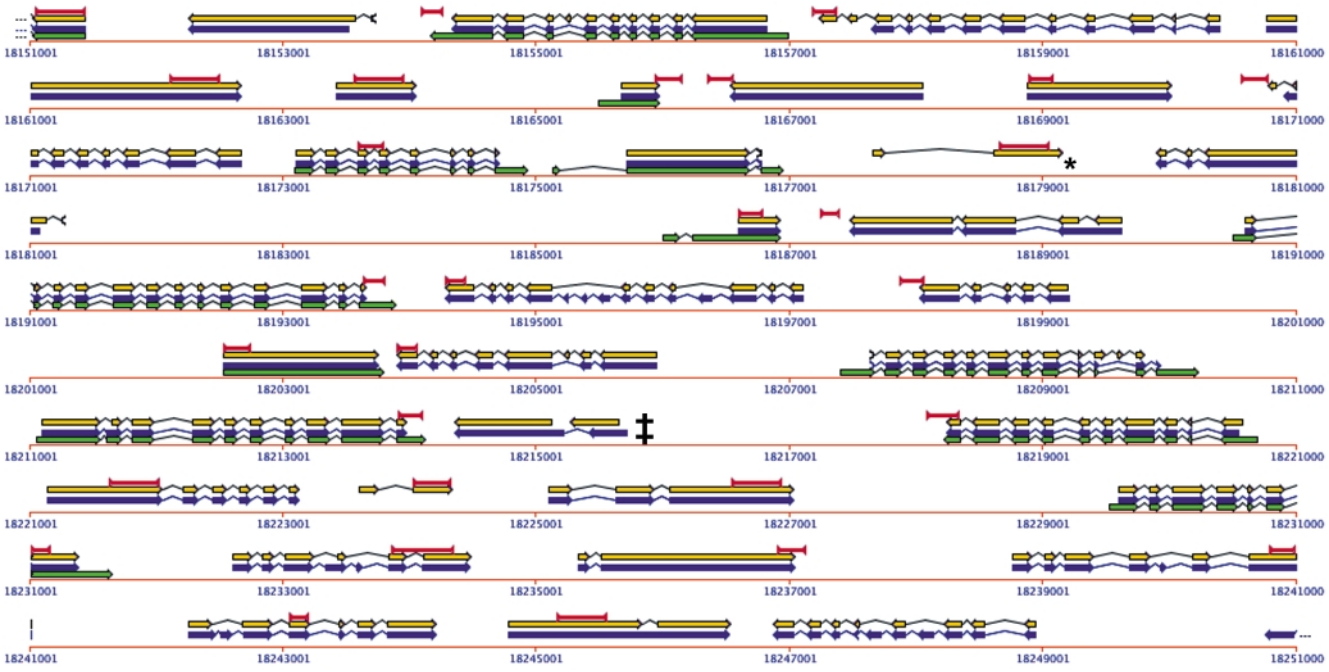

Figure 1. Comparison of EuGène and AGI gene predictions and cDNA sequences in a region of *Arabidopsis thaliana* chromosome 2. Orange arrows, EuGène gene model; blue arrows, AGI gene model; green arrows, cDNA sequences; red bars, CATMA GSTs. Examples can be seen where only EuGène predicts a gene model (*) or where EuGène splits an AGI gene model (‡). cDNA sequences include 5′ and 3′ non-coding regions, so extend up- and downstream of the predicted gene models, which contain only coding sequence.

tendency reduces the probability of designing only one GST for two adjacent genes but with the slight penalty that in some cases, two GSTs will be designed for a single gene. We supplemented automated gene predictions with verified full length cDNA sequences and more intensively, individually analysed gene predictions. We predicted a total of around 29 600 genes using this approach. Figure 1, generated using the FLAGdb++ tool (5), shows an example of the genes predicted by EuGène in a region near the end of chromosome 2, with comparisons to AGI gene models and GST and cDNA sequences. Cases can be seen where EuGène includes a new gene model or splits an AGI prediction, but overall there is good consistency between the two predictions and, where present, with the cDNA sequences.

We designed GSTs from the reannotated gene set using SPADS (Specific Primers and Amplicon Design Software) (6), which optimises and automates the choice of a genome-specific GST for each gene and designs PCR primers to amplify this GST. The principal requirements for a GST were that it should be between 150 and 500 bp in length and show no more than 70% identity to any other part of the *Arabidopsis* genome. In the first round of GST design, the results from which make up the initial release of the database, we designed GSTs for just over 21 000 genes, or ∼70% of the predicted total.

CONTENTS

Entries in the database are keyed by the CATMA GST identification code, and all have a GST sequence associated with them as well as the predicted or experimentally verified gene transcript sequence. In addition, the primer sequences used to generate the GST products, the BAC templates from which they were generated and the results of the PCR amplification are available. There is also a range of supplementary information, such as gel images and other quality control data, details of the synthesising laboratory, the specificity of the GST against the *A. thaliana* genome, matches to the EMBL (7) and SWISS-PROT database (8) and comments about the GST. Wherever possible, the CATMA genes and GSTs are related to a corresponding AGI (*Arabidopsis* Genome Initiative, http://www.arabidopsis.org) gene model. We assign an AGI gene to a CATMA gene based on two parameters: position of the predicted genes in the genome and homology of the GST to the AGI gene model. If the chromosome positions of the two gene models overlap and the GST shows more than 95% identity to the AGI model over at least 150 bp, then the AGI gene model is considered a match for that CATMA gene model. However, no further correlation between the CATMA and AGI gene models is implied by the assignment of a match.

ACCESSING THE DATABASE

Access to the database is freely available to any researcher on completion of a brief registration form. This should be accessed through our website main page (http://www.catma.org/). A browser capable of supporting cookies is required, and it is recommended that javascript be enabled.

There are three main methods available for searching the CATMA database. The first is a simple search by name to retrieve any or all of the data associated with a specific gene or GST. The name used may be a CATMA GST identification code, a CATMA gene name or an AGI gene name. A list of several gene/GST names can be searched at once, or a single name supplied for more detailed information.

The second search method is sequence similarity. A target sequence can be used to perform WU-BLASTN (W.Gish, http://blast.wustl.edu/) searching against the GST, gene transcript model or full gene (including intron) sequences. Again several sequences can be submitted at once, or more detail can be obtained by using just a single sequence. Finally, the database may be searched directly using SQL queries; this is a powerful, albeit less user-friendly, way of performing searches, and is frequently used by CATMA database users.

RECENT AND FUTURE DEVELOPMENTS

CATMA is an ongoing project, and we will continue to design GSTs towards our ultimate goal of a GST for every *Arabidopsis* gene. As we design new batches of GSTs, they will be amplified and tested by members of the CATMA consortium. Once each batch has been thus verified, we will add information about those GSTs and their associated genes to the database. Users of the database can register for an email notification when significant updates are made to the database.

There are three main sources for these additional GSTs. Primarily, there is the design of GSTs for genes which were missed in the first round; the majority of these are members of gene families that are highly similar at the nucleotide level, and so for which it is difficult to generate specific GSTs. We will be using a variety of strategies to design GSTs for these genes, including the introduction of a putative 3′ UTR and increasing the identity cutoff above its current 70%; all such deviations from the standard conditions will be flagged in the database. The second group of extra GSTs are replacements for those which failed to amplify properly in the PCR, for which we will design alternative primer pairs. Finally, as EST and cDNA sequencing projects add to or alter our gene set and improvements in the gene prediction software result in the identification of extra genes (both by CATMA and other groups), we will design GSTs for these genes. Already we are producing GSTs for ~1500 AGI gene predictions which do not correspond to any CATMA genes.

Gradual changes will be made to add to the search capabilities of the database—for example, we have recently introduced searching by EMBL/GenBank accession number as well as by CATMA and AGI gene names. Finally, as the CATMA GSTs start to be widely used in microarray experiments, we would expect to link the results of those experiments back to the GST database.

ACKNOWLEDGEMENTS

We thank all the researchers involved in CATMA for their hard work. A full list of contributors is available at http://www.catma.org/acknowledgements.html. We also thank Thomas Schiex of INRA Toulouse for optimization of EuGène and Franck Samson, Véronique Brunaud and Alain Lecharny of INRA Evry for the FLAGdb++ image. The CATMA project was supported by the following funding agencies and organizations: BBSRC Investigating Gene Function Initiative grant 208/IGF12424 (GARNet); Institut National de la Recherche Agronomique (INRA); Vlaams Interuniversitair Instituut voor Biotechnologie (VIB); Génoplante; Division Earth and Life Sciences ALW from the Netherlands Research Council NOW; University of Lausanne; Spanish Ministery of Science and Technology, grant BIO2000-3094-E; BMBF grant 0312274 (GABI).

REFERENCES

1. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature*, **408**, 796–815.
2. Haas,B.J., Volfovsky,N., Town,C.D., Troukhan,M., Alexandrov,N., Feldmann,K.A., Flavell,R.B., White,O. and Salzberg,S.L. (2002) Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.*, **3**, research0029.1–0029.12.
3. Schiex,T., Moisan,A. and Rouzé,P. (2001) EuGène: An Eucaryotic Gene Finder that combines several sources of evidence. In Gascuel,O. and Sagot,M.-F. (eds), *Comput. Biol.*, LNCS 2066, 111–125.
4. Pavy,N., Rombauts,S., Déhais,P., Mathé,C., Ramana,D.V., Leroy,P. and Rouzé,P. (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics*, **15**, 887–899.
5. Samson,F., Brunaud,V., Balzergue,S., Dubreucq,B., Lepiniec,L., Pelletier,G., Caboche,M. and Lecharny,A. (2002) FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of *Arabidopsis thaliana* T-DNA transformants. *Nucleic Acids Res.*, **30**, 94–97.
6. Thareau,V., Déhais,P., Rouzé,P. and Aubourg,S. (2001) Automatic design of gene specific tags for transcriptome studies. In Duret,L., Gaspin,C. and Schiex,T. (eds), JOBIM 2001, Toulouse, 195–196.
7. Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R., Redaschi,N., Stoehr,P., Tuli,M.A., Tzouvara,K. and Vaughan,R. (2002) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **30**, 21–26.
8. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.