D656–D659 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki036

Brassica ASTRA: an integrated database for *Brassica* genomic research

Christopher G. Love[1,2], Andrew J. Robinson[1,2], Geraldine A. C. Lim[1,2], Clare J. Hopkins[1], Jacqueline Batley[1], Gary Barker[3], German C. Spangenberg[1,2] and David Edwards[1,2,*]

[1]Plant Biotechnology Centre, Primary Industries Research Victoria, La Trobe University, Bundoora 3086, Victoria, Australia, [2]Victorian Bioinformatics Consortium, Plant Biotechnology Centre, Primary Industries Research Victoria, La Trobe University, Bundoora 3086, Victoria, Australia and [3]School of Biological Sciences, University of Bristol BS8 1UG, UK

Received August 9, 2004; Revised and Accepted September 23, 2004

ABSTRACT

***Brassica* ASTRA is a public database for genomic information on *Brassica* species. The database incorporates expressed sequences with Swiss-Prot and GenBank comparative sequence annotation as well as secondary Gene Ontology (GO) annotation derived from the comparison with *Arabidopsis* TAIR GO annotations. Simple sequence repeat molecular markers are identified within resident sequences and mapped onto the closely related *Arabidopsis* genome sequence. Bacterial artificial chromosome (BAC) end sequences derived from the Multinational *Brassica* Genome Project are also mapped onto the *Arabidopsis* genome sequence enabling users to identify candidate *Brassica* BACs corresponding to syntenic regions of *Arabidopsis*. This information is maintained in a MySQL database with a web interface providing the primary means of interrogation. The database is accessible at http://hornbill.cspp.latrobe.edu.au.**

INTRODUCTION

Brassica comprises a diverse group of species including major vegetable and oilseed crops with a wide range of agronomic traits. *Brassica* species are within the same family (Brassicaceae, mustard family) as the fully sequenced and well-annotated model plant *Arabidopsis thaliana*. The ancestral lineages of *Arabidopsis* and *Brassica* diverged between 12.2 and 19.2 million years ago and the two species share extensive collinearity and 87% sequence identity between othologous exons (1).

The *Brassica* ASTRA database has been developed for the analysis and interrogation of *Brassica* genomic information, the identification of candidate genes for agronomic traits and comparative analysis with the *Arabidopsis* genome. The database uses a series of PERL scripts, which act as wrappers for sequence processing, annotation and management of a MySQL database. A web-based interface allows the researcher to interrogate and navigate through processed information on sequence annotation, molecular markers and comparative analysis with the genome of *A.thaliana*.

DATA AND PROCESSING

The primary sequence dataset comprises all the available public sequences for the major *Brassica* species. In the current release (v3.0), *Brassica* ASTRA contains 44 877 expressed sequences from *B.napus*, *B.oleracea*, *B.nigra* and *B.rapa* derived from the GenBank (2) and 1759 bacterial artificial chromosome (BAC) end sequences from *B.rapa* supplied by Professor Yong Pyo Lim (Chungnam National University, Korea).

Multiple annotation methods have been applied to the sequence data. The concurrent application of independent annotation methods enables automated annotation with confidence being derived from the consensus of derived results. Expressed sequences are annotated by BLASTn (3) comparison against GenBank plant nucleotide sequences, and BLASTx comparison against the complete Swiss-Prot database (4) using a cut-off value of $E < 10^{-5}$. The first and the top 10 BLAST hits are parsed to separate tables within the database.

Of the total expressed sequences, 90% were annotated using BLASTn against GenBank plant sequences, while 53% identified at least one significant match in the Swiss-Prot database.

Intermediate Gene Ontology (GO) annotation is derived using the BLASTx comparison of database expressed sequences with GO annotated *Arabidopsis* peptide sequences from TAIR (5). Where a *Brassica* sequence has significant

*To whom correspondence should be addressed. Tel: +61 0 3 9479 5633; Fax: +61 0 3 9479 3618; Email: Dave.Edwards@dpi.vic.gov.au

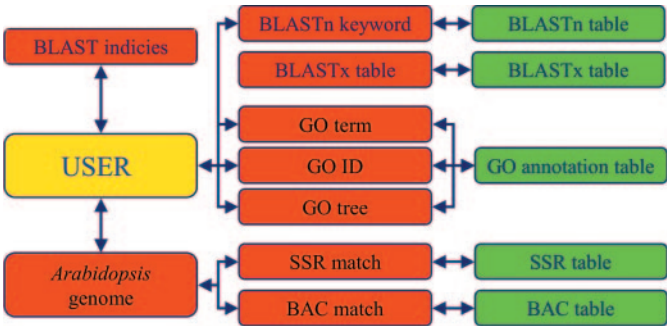

Figure 1. Work flow schematic diagram illustrating the multiple forms of data interrogation.

sequence identity ($E < 10^{-5}$) to a sequence from *Arabidopsis*, the *Arabidopsis* GO annotation is extrapolated to the corresponding *Brassica* sequence. Owing to the extent of gene duplication in *Arabidopsis*, a single *Brassica* expressed sequence tag (EST) may identify multiple *Arabidopsis* proteins. To enrich true orthologous comparisons, only the most significant BLAST match is processed. Only 43% of the expressed *Brassica* sequences have intermediate GO annotation. Sequences that have no GO annotation can be attributed either to matches with *Arabidopsis* proteins with no GO annotation (26%) or failure to identify a homologous *Arabidopsis* protein (31%).

Sequences are clustered and assembled using TIGR Gene Indices Clustering Tools (TGICL) (6) using strict criteria (98%

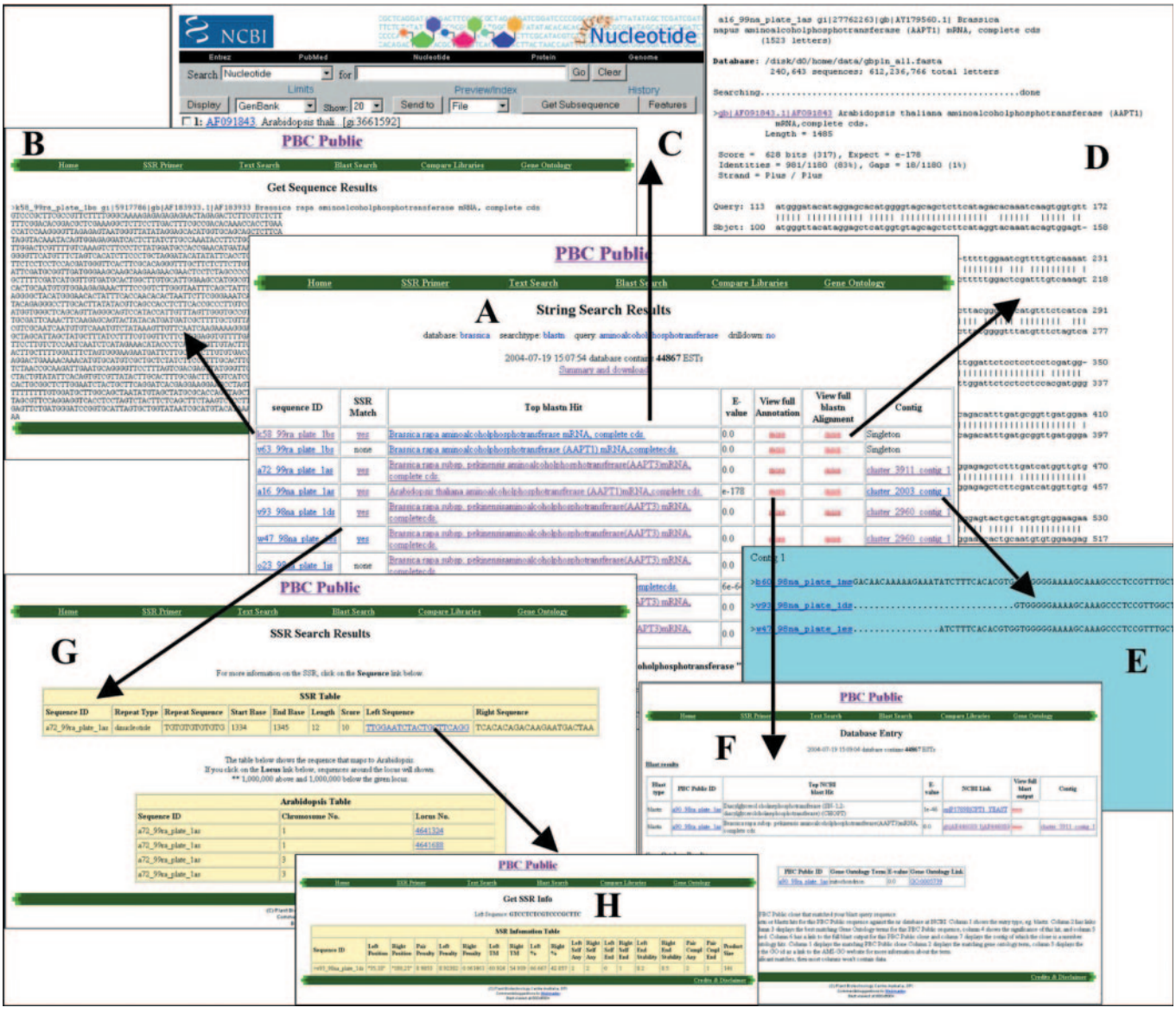

Figure 2. Query result views illustrating the integration of data within the database. The result summary (**A**) links to the candidate EST sequence (**B**), the GenBank record for the most significant match (**C**), the complete BLAST annotation results (**D**), sequence alignment of contig representatives (**E**) and a view of all the functional annotation for that sequence record (**F**). Identified SSRs link to information and on primer sequence for their amplification, comparative physical location on the *Arabidopsis* genome (**G**) and detailed information for SSR amplification (**H**).

minimum percentage identity between overlaps, minimum overlap length of 40 bases and no more than 20 unmatched overhangs) and alignments are linked to constituent sequences in the database. A total of 44 877 *Brassica* expressed sequences assembled into 6192 distinct contigs representing 28 256 sequences, with the remaining 16 621 classified as singletons, giving a potential of 22 813 transcripts present in the database.

Database sequences are mined for the presence of simple sequence repeat (SSR) molecular markers using the SSRPrimer tool (7). The sequences containing SSRs, along with PCR primers for their amplification are parsed directly into the database. Sequences containing SSRs are compared with the *Arabidopsis* genome using BLASTn to identify potential homologous genomic regions between the two species. SSRs were identified in 6625 (15%) of ESTs. As expected for expressed sequences, trinucleotide repeats were the most abundant form (69%) followed by dinucleotides (19%), tetranucleotides (7%) and pentanucleotide (5%) repeats. A total of 5070 SSR containing *Brassica* sequences matched similar sequences within *Arabidopsis*, distributed across chromosome 1 (1185), chromosome 2 (814), chromosome 3 (1126), chromosome 4 (701) and chromosome 5 (1244). Expressed sequences, SSRs and PCR primer sequences may also be downloaded from the database as flat files.

QUERY TOOLS AND USER INTERFACE

The web query interface provides multiple routes to interrogate the database (Figure 1). Submitting the accession number for a sequence will retrieve the full annotation for that sequence. *Brassica* sequences may also be identified by submitting query amino acid or DNA sequences to the BLAST form, which will then return a list of corresponding matches and links to their annotation. Sequence annotation, derived from the BLAST comparisons of resident sequences with the GenBank or Swiss-Prot databases, may be searched using key words. The search may be limited to annotation for the best BLAST match for each sequence or extended to include the 10 most significant matches.

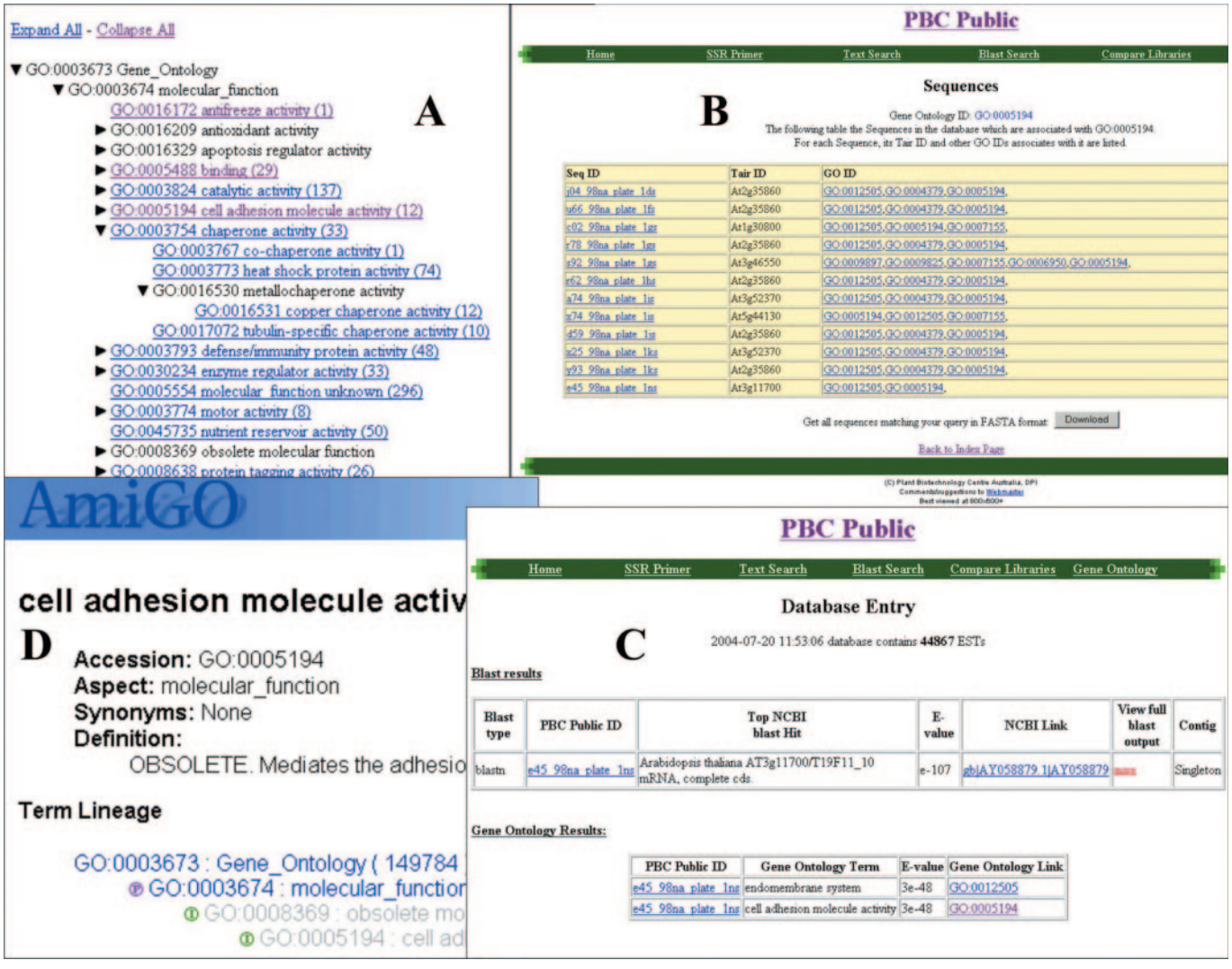

Figure 3. Browsing the GO tree structure. The molecular function hierarchical tree of GO annotation (**A**), links to a list of the sequences annotated with a specific GO term (**B**), with further links to the complete database entry for each sequence (**C**). GO IDs link to the external Ami-GO website (**D**).

The results for both annotation and sequence searches are presented as web tables providing links to further information on specific sequences (Figure 2). These results include:

Sequence ID: Links to the complete sequence in the FASTA format.
SSR Match: Links to information on identified SSRs, PCR primers for their amplification and predicted homologous regions on the *Arabidopsis* genome.
Primary Annotation: Top BLAST hit, *E*-value and the NCBI link.
Full Annotation: Includes all annotation available for that record (BLASTn, BLASTx and GO annotation).
Full Blast Output: Includes complete BLAST annotation alignments with hyperlinks to each original NCBI record.
Contig: Links to the aligned sequences that make up a contig assembly.

One major limitation of comparative GenBank or Swiss-Prot annotation is the inconsistency in the annotation of sequences submitted to these databases. The use of GO annotation derived from intermediate mapping overcomes this limitation since GO uses a precise vocabulary (8). The *Brassica* ASTRA GO tables may be searched using a GO term or GO ID number. Furthermore, a hierarchical GO annotation tree may be browsed to identify lists of *Brassica* sequences annotated with specific GO terms (Figure 3). Searching the *Brassica* GO annotation, either using GO term, or GO ID or tree browse methods produces a table of sequences, which have been annotated with the specified GO term. These tables link to the complete annotation of each identified sequence. The presence of multiple, distinct forms of annotation provides a level of confidence in the result. GO annotation searches maintain hyperlinks to the AMI-GO website to provide detailed information on specific GO terms.

Brassica ASTRA hosts tools aiding the physical and genetic comparison between the *Brassica* and *Arabidopsis* genomes. For both tools, users specify regions of the *Arabidopsis* genomic sequence. The *Brassica* SSR search will identify lists of predicted homologous *Brassica* expressed genes that contain SSRs. Further links provide information on PCR primers for amplification of the SSR in *Brassica* and the subsequent comparative mapping of the locus. The *Brassica* BAC end tool identifies *Brassica* BACs that map onto specified regions of the *Arabidopsis* genome. Where both ends of a single BAC map within a 500 000 bp region of the *Arabidopsis* genome and demonstrate opposite orientation to each other with respect to the *Arabidopsis* genome, the BAC result is highlighted. These cases suggest the presence of conserved intervening sequences present on the unsequenced portion of the identified BAC clone. Within an initial set of 635 BACs, 78 mapped onto one or more locations on the *Arabidopsis* genome using the above criteria, representing 9% of the *Arabidopsis* genome. End sequences for a further 100 000 BACs are being produced within the Multinational *Brassica* Genome Sequencing Project, representing a predicted 14-fold coverage of the *Arabidopsis* genome.

FUTURE DEVELOPMENTS

The structure of this database offers the flexibility to expand and integrate a variety of further forms of genomic information including gene expression and genetic mapping data, predicted single nucleotide polymorphisms (9) as well as sequence data resulting from the Multinational *Brassica* Genome Project. There is also significant scope for further integration with other new and established databases and work is currently under way to promote links between *Brassica* ASTRA and features on the *Arabidopsis* Ensembl genome viewer (10).

ACKNOWLEDGEMENTS

We would like to express our gratitude to Professor Yong Pyo Lim (Chungnam National University, Korea) for the provision of BAC end sequence data for *Brassica rapa*.

REFERENCES

1. The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis* thaliana. *Nature*, **408**, 796–815.
2. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res*., **31**, 23–27.
3. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
4. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
5. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al*. (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res*., **31**, 224–228.
6. Pertea,G., Huang,X., Liang,F., Antonescu,V., Sultana,R., Karamycheva,S., Lee,Y., White,J., Cheung,F., Parvizi,B. *et al*. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.
7. Robinson,A.J., Love,C.G., Batley,J., Barker,G. and Edwards,D. (2004) Simple sequence repeat marker loci discovery using SSR primer. *Bioinformatics*, **20**, 1475–1476.
8. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.
9. Barker,G., Batley,J., O'Sullivan,H., Edwards,K.J. and Edwards,D. (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using AutoSNP. *Bioinformatics*, **19**, 421–422.
10. James,N., Craigon,D., Gill,G., Schildknecht,B., Sun,G.A. and May,S. (2004) AtEnsembl—a new *Arabidopsis* genomic resource. In *Plant and Animal Genomes XII Conference*, Town & Country Convention Center, San Diego, CA, January 10–14, p. 999.