GABI-Kat SimpleSearch: a flanking sequence tag (FST) database for the identification of T-DNA insertion mutants in Arabidopsis thaliana

Yong Li, Mario G. Rosso, Nicolai Strizhov†, Prisca Viehoever and Bernd Weisshaar∗,
‡

Max Planck Institute for Plant Breeding Research, Carl-von-Linne-Weg 10, D-50829 Koeln, Germany

Received on December 19, 2002; revised on February 17, 2003; accepted on February 20, 2003

ABSTRACT

Summary: GABI-Kat SimpleSearch is a database of flanking sequence tags (FSTs) of T-DNA mutagenized Arabidopsis thaliana lines that were generated by the GABI-Kat project. Sequences flanking the T-DNA insertion sites were aligned to the A.thaliana genome sequence, annotated with information about the FST, the insertion site and the line from which the FST was derived. A web interface permits text-based as well as sequence-based searches for relevant insertions. GABI-Kat SimpleSearch aims to help biologists to quickly find T-DNA insertion mutants for their research.

Availability: http://www.mpiz-koeln.mpg.de/GABI-Kat/

Contact: weisssha@mpiz-koeln.mpg.de

In plants, an effective way of disrupting genes is insertional mutagenesis using transposons or T-DNA. Collections of mutant lines containing transposon or T-DNA insertions are valuable resources for plant functional genomics. Several transposon (Wisman et al., 1998; Parinov et al., 1999; Genetrap DB, http://genetrap.cshl.org/) and T-DNA (Krysan et al., 1999, SIGnAL, http://signal.salk.edu/tabout.html; FLAGdb/FST, http://genoplante-info.infobiogen.fr; Samson et al., 2002) mutagenized populations have been created for Arabidopsis thaliana, the model system in which many aspects of plant biology are studied. The transposon or T-DNA integrated into the plant genome not only disrupts a gene that might be located at the insertion site, but also provides a tag to facilitate the identification of that gene. The availability of the almost complete A.thaliana genome sequence generated by The Arabidopsis Genome Initiative (2000), and PCR techniques that enable us to recover plant genomic DNA sequences flanking the insertion site (flanking sequence tag, FST), make it possible to locate the exact position of insertions in the A.thaliana genome.

The GABI-Kat project is building a large T-DNA mutagenized A.thaliana population (accession Columbia, the sequenced genotype) of finally 70,000 plants with sequence-characterized insertion sites. The sequence trace files derived from PCR-generated DNA fragments spanning insertion sites were processed using phred (Ewing et al., 1998). T-DNA vector sequences in the candidate FSTs were masked using cross_match, and sequences shorter than 30 nt were discarded. Candidate FSTs passing this filter were aligned to the A.thaliana genome sequence by BLASTN (Altschul et al., 1997). Up-to-date nuclear genome sequence data (BAC sequences of the minimal tiling path) and gene annotation data were obtained from MATDB (MIPS Arabidopsis thaliana Database, Schoof et al., 2002). Only sequences with BLAST expect values lower than 5e-4 were considered as good FSTs. BLAST reports were parsed and the expected insertion site were calculated. MATDB gene annotation data were used to determine if a given insertion site was within a gene. An FST qualifies as a ‘gene hit’ when the insertion site is located between 300 bp upstream of the ATG and 300 bp downstream of stop codon of an annotated gene. We use the term ‘CDSi hit’ when an insertion site is located between ATG and stop codon (insertions in CDS and included introns).

Annotated FSTs and gene annotation data were stored in a relational database (MySQL). The data are accessible through a web interface termed GABI-Kat SimpleSearch (http://www.mpiz-koeln.mpg.de/GABI-Kat/). On the web site there are several static HTML pages providing general information of the project, the procedure to obtain seeds and other relevant information. The database queries were implemented by a set of PHP scripts running in an Apache/PHP environment. The PHP module of the GD
library was used for drawing the graphics. In addition to GABI-Kat SimpleSearch, FST data are also available from the GSS division of EMBL/GenBank/DDBJ.

The goal of SimpleSearch is to allow quick access from plant biologists to the T-DNA insertion mutants generated by GABI-Kat. Two ways to start a search are available: (i) a text-based search to find ‘gene hits’; and (ii) a sequence-based search using BLASTN or TBLASTN against all FSTs. The text-based search accepts either AGI gene codes (e.g. At1g23450) or a keyword as input. Up to 20 AGI gene codes can be entered at once. The key word search performs a substring search in the gene annotation text field. This feature helps to find FSTs related to a given gene family, provided that the gene annotation is accurate. The sequence-based search allows accessing FSTs mapping to intergenic regions that might potentially be of interest to users, for example because a not annotated gene is assumed to exist. The results from both search methods will eventually be presented in a table of entries matching the search criteria. The columns in the table are the GABI-Kat line ID linked to the FST page, the AGI gene code linked to MATDB, the gene annotation text and a link to the graphic display of the expected insertion site. The FST page displays the FST in FASTA format, along with data from the BLAST output including the presence/absence of T-DNA sequences in the original sequence read. To assist users in evaluating insertion sites relative to genes of interest, the graphic view displays the genome fragment around the gene or tagged sequence at the expected insertion site as an image. Annotated genes with exon–intron structure and other FSTs located in this region are shown, and BLAST expect values of the FSTs are colour-coded. The image can be zoomed out and back in again to display two differently sized genome fragments. Users can request seeds of an insertion line and seeds will be delivered if the T-DNA insertion can be confirmed, and the SimpleSearch web site contains additional features such as seed request tracking and personalized access to confirmation sequences.

The February 2003 release of SimpleSearch includes 41 389 FSTs from 26 375 T-DNA transformed A.thaliana lines; annotation is based on MATDB release 11012003. A total of 9862 different genes (37% coverage) have been hit, of which 6458 qualify as CDSi hits (24% coverage). The mean length of FSTs is 239 bp, and in 70% of the FSTs T-DNA-derived sequence was detected. GABI-Kat SimpleSearch will be updated regularly with new FSTs. Until the final goal of 70 000 analyzed lines is reached, the FSTs from an average of about 6000 transformed lines will be entered into SimpleSearch every 3 months. Also, new MATDB releases containing improved gene annotation data and additional genomic sequences from the few remaining gaps in the genomic sequence will be incorporated. Future developments will address the completeness of the annotation information incorporated into SimpleSearch by including additional data from sources like TIGR (http://www.tigr.org/), TAIR (http://www.arabidopsis.org/) and CATMA (http://www.catma.org/).

ACKNOWLEDGEMENTS

The authors thank all members of the GABI-Kat and the ADIS teams for greenhouse and laboratory work. Thanks also to Thomas Rosleff Soerensen and Martin Werber for helpful discussions. This work is supported by the BMBF in the context of the German plant genomics program GABI (Förderkennzeichen 0312273).

REFERENCES


