D874–D878 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 7 November 2006*
doi:10.1093/nar/gkl753

GABI-Kat SimpleSearch: an *Arabidopsis thaliana* T-DNA mutant database with detailed information for confirmed insertions

Yong Li[1,2], Mario G. Rosso[1,2], Prisca Viehoever[1] and Bernd Weisshaar[1,*]

[1]Institute of Genome Research, Center for Biotechnology (CeBiTec), Bielefeld University, Universitaetsstrasse 25, D-33594 Bielefeld, Germany and [2]Max Planck Institute for Plant Breeding Research, Carl-von-Linne-Weg 10, D-50829 Koeln, Germany

Received August 15, 2006; Accepted September 25, 2006

ABSTRACT

Insertional mutagenesis approaches, especially by T-DNA, play important roles in gene function studies of the model plant *Arabidopsis thaliana*. GABI-Kat SimpleSearch (http://www.GABI-Kat.de) is a Flanking Sequence Tag (FST)-based database for T-DNA insertion mutants generated by the GABI-Kat project. Currently, the database contains >108 000 mapped FSTs from ~64 000 lines which cover 64% of all annotated *A.thaliana* protein-coding genes. The web interface allows searching for relevant insertions by gene code, keyword, line identifier, GenBank accession number of the FST, and also by BLAST. A graphic display of the genome region around the gene or the FST assists users to select insertion lines of their interests. About 3500 insertions were confirmed in the offspring of the plant from which the original FST was generated, and the seeds of these lines are available from the Nottingham Arabidopsis Stock Centre. The database now also contains additional information such as segregation data, gene-specific primers and confirmation sequences. This information not only helps users to evaluate the usefulness of the mutant lines, but also covers a big part of the molecular characterization of the insertion alleles.

INTRODUCTION

Insertional mutagenesis approaches by using transposable elements or Agrobacterium T-DNA, play important roles in plant functional genomics (1–4). The use of T-DNA as an insertional mutagen has several advantages (3,4). T-DNA integration results in stable mutations in the genome, as opposed to transposons which often excise after integration. Also, the low number of insertions per transformant significantly reduces the additional work required to remove unwanted mutations. The development of a simple *in planta* Agrobacterium transformation method for *Arabidopsis thaliana* allowed the high-throughput production of T-DNA insertion mutants in this model plant (5,6). A number of large T-DNA mutagenized *A.thaliana* populations have been generated that are intensively used for gene function search studies and other reverse genetics experiments (7–11). Initially, these populations were screened for the desired insertion mutants in the gene of interest by PCR. The PCR was performed with a gene-specific primer and a T-DNA-specific primer on DNA templates from large pools of mutant plants. The identification of a specific PCR product indicated the existence of a T-DNA insertion in the gene of interest, and the respective mutant was then identified by pool deconvolution. An alternative strategy, which gained popularity in recent years, was to amplify the DNA fragments flanking the T-DNA insertion sites of individual plants with special PCR-based methods, and to sequence-characterize the various insertion alleles. When the whole genome sequence of the species is available, the resulting Flanking Sequence Tags (FSTs) can be mapped to the pseudochromosomes and indexed in databases. As a result, the screening for a mutant insertion allele of a given gene is simplified to a search for the corresponding FST in the database. Obviously, the number of insertion mutants and FSTs in the database should be large enough to allow most of the genes being covered (3).

The aim of the GABI-Kat project was to build a large T-DNA mutagenized *A.thaliana* population with sequence-indexed insertion sites (11). In June 2002, GABI-Kat SimpleSearch, the web interface of the database containing data about the GABI-Kat population, was opened (12). Since then, GABI-Kat lines have been ordered by many scientists from all around the world, and thousands of confirmed insertion mutants have been delivered to the scientific community. This shows that the GABI-Kat population became one of the major reverse genetics resources for *A.thaliana* functional genomics.

*To whom correspondence should be addressed. Tel: +49 521 106 6873; Fax: +49 521 106 6423; Email: bernd.weisshaar@uni-bielefeld.de

During the last few years and since we described the basic FST analysis pipeline (12), we have constantly improved the database as well as the interface, and incorporated several major updates. This took place in addition to the regular increase in the amount of FST data held in the database. Here, we summarize and describe the improvements, updates and new features of the SimpleSearch tool, including the recent addition of detailed information on insertion sites from allele-specific sequence data and genetic segregation data from 3509 (number as of June 2006) lines with confirmed insertions.

DATABASE CONTENT

FSTs and annotation

The FST production pipeline and the annotation procedure were described in detail elsewhere (12,13). In brief, the genome sequences flanking the T-DNA insertion site (FSTs) were obtained by an adaptor-ligation PCR method which was adapted to high-throughput conditions (13). Each FST was mapped to the *A.thaliana* genome by BLAST (14), and annotated with the information of the corresponding BAC clone, and with the AGI gene code when the insertion qualifies as a 'gene hit' (12). We define 'gene hit' as the insertion site being located between 300 bp upstream of the ATG and 300 bp downstream of stop codon of a gene, and 'CDSi hit' as the insertion site being located between ATG and stop codon (CDS plus introns). *A.thaliana* genome sequence and annotation data from the TIGR version 5 dataset were used as the basis for mapping and annotation (15). Table 1 gives a summary of the data in the SimpleSearch database as of the current release (GK release 21 of June 25, 2006). There are >108 700 FSTs, and they were from >63 800 lines. Based on the TIGR v5 annotation, >16 900 genes (64% of all annotated *A.thaliana* protein-coding genes excluding pseudogenes) were covered by at least one 'gene hit' and >12 200 genes (46%) were covered by at least one 'CDSi hit'.

Data concerning confirmed insertions

One problem associated with the T-DNA mutagenized population is that not all the insertions deduced from FSTs that were obtained from T1 plants can be confirmed in the next (T2) generation. The T1 generation refers to plants that were selected for resistance conferred by the inserted T-DNA. Both SAIL (Syngenta population) and GABI-Kat reported a confirmation rate of ~76% (9,11). At present, the confirmation rate at GABI-Kat is 78%. This number is derived from ~6000 lines for which confirmation was attempted during the process of in-house confirmation to make sure that only confirmed insertion lines were delivered to users. The confirmation process consists of segregation analysis, genomic DNA extraction from T2 plantlets, PCR using a T-DNA primer and a gene-specific primer designed to fit the insertion site predicted on the basis of the original FST annotation, and sequencing of this allele-specific PCR product if the PCR was successful (11). For segregation analysis, we record the number of seeds plated, the number of seeds germinated and the number of resistant seedlings. This information is very useful as it gives an approximate number of T-DNA loci of the specified line after statistical evaluation according to Mendel's laws. In addition, distorted segregation patterns often indicate that the line contains a mutation in a gene important in pollen or female gametophyte development (16,17). We have stored this information for lines for which confirmation has been attempted in the database, and in total there are 6000 lines for which segregation data are available (Table 1). The database also contains information on ~20% failed confirmation attempts (search for the FST with GenBank accession No. 'CR405437' to see an example). The products resulting from a successful allele-specific PCR were sequenced from both directions, so usually two confirmation sequences were obtained for a confirmed insertion. Currently, there are 8840 confirmation sequences in the database, and they were stored together with the primer information so that the allele-specific PCR can be reproduced. Since July 2005, T3 seeds (seeds produced by T2 plants) of confirmed GABI-Kat lines are transferred to the Nottingham Arabidopsis Stock Centre (NASC). This allows direct user access to potential homozygous mutant materials, and the SimpleSearch database provides the molecular and genetic background information for these lines.

Table 1. Summary of data in the GABI-Kat SimpleSearch database[a]

Data type	Number of entries
FSTs	~108 700
Lines	~63 800
Lines with segregation data	~6 000
Lines available in NASC	3 509
Distinct genes covered	16 939
Distinct CDSi covered	12 239
Confirmation sequences	8 840

[a]Numbers are of GK release 21 as on 25 June 2006.

WEB INTERFACE

In addition to the search for insertion alleles by the AGI gene (or locus) code or keyword and the sequence-based search by BLAST, we recently introduced the feature of searching by line ID or GenBank accession number. This gives a more thorough access to the data in database. Searching by line ID or GenBank accession number will directly lead to the FST data display page (Figure 1). On top of the FST page is information for the line from which the FST was generated. These include the vector used to transform the line, with a link to the GenBank entry of the vector at NCBI, information about line availability and when available segregation data. The 'line availability' field tells if a given line is dead (e.g. because no viable seeds were produced by the T1 plant), or if it is available from GABI-Kat or NASC. When a line is available from NASC, the NASC code is given and linked to the NASC seeds stock detail page of the line, so that the user can order it easily. The segregation data are presented as three numbers: total number of seeds evaluated, number of seeds germinated and number of resistant seedlings. Since more than one FSTs were produced for many T1 plants in the population, the line information page includes data for one or more FST-derived (predicted) insertion sites or loci. For each displayed

Line specific information

Line ID	048E04
Vector Used	pAC161
Line Availability	available from NASC (N404564)
Segregation Analysis	50:50:37
Confirmed for Hit	At5g43175

Gene hit *At5g43175*

Sequence (A. th genome BLAST matches underlined)	>29-K016082-022-048-E04-8409 TGATATCTAAATAAATATTGGTAGTCCTTAAAATAATATTTATGACTTTCAGGTACATCT TAAAAGAAGATTAAAATTAAAATGCTGTTAGTCCAAATTGTCTTTAGCATATAATAATTA CAAAAAGAGTGTCAAGTTCATAGAACGCAATAGACAGAAAAATTGAACAACTCAAACAAA AAATAACCATAATCATGAATATAGAACGAGAAAAAAAACAAGTACTGTAAGTTATGTTAT TCCAATGATTTTGTCTTAAATAAGCCGAGACAAAAGATTGTGATGTAGTCCCATATTCAG ACCATTGTGAGCAAGAGGTGCATACATCCATAGATCTTCTGAACTCAAGAGCTACATTGC AAAATTTCCAAAAACGACAAAAGTGTTTGGTTAGGAAAAATGGTAAAACAGAGTTTAGGT AAGAGTTTTACATGGTTTCTTGCAGTACAAGAAACGATCAATACTCCAACAAGAAAAAAA GTCGTTACCTTGATTTGAAGCTGCAGGAACTTCACGTAATGGACAGCATCTTCCAGCATT GTGCTTATATCGACCTTTGTCCCATTAGGAACTAGGCTCTGCAAGGTCTTTAGCCTATCG TTTATCCTTTCTCTTCGTTTCTGCAACATTCATCAACATTATTATTCCACAAAACCTAAA AACTCAAATCAACAATGAAAATATTTGTTTAGGTTATCTTGATT
GenBank Accession	AL936383
Graphic View	
T-DNA Border	0 base(s) before start of sequence
BLAST e value	0.0
Hit Clone Code (BAC ID)	MMG4
Hit Gene Code	At5g43175 [TIGR] [TAIR] [MIPS] [SIGnAL]
Insertion Classification	3' region
Confirmation Status	confirmed, show confirmation sequences

Line 048E04 has other detected FST hit, show all hits of this line

Figure 1. Screenshot of the FST page that results from a search for GenBank accession number AL936383. The data are from line 048E04. The meaning of the three numbers describing the result of the genetic segregation data is found by following the link on the numbers. The upper part shows the line-specific information followed by the FST DNA sequence data in the lower part of the page. For the chosen line, data on additional insertions other than the one displayed are available. Therefore, a link to information about all the insertions from the line is given at the bottom.

insertion, the FASTA format sequence of the original FST, the respective GenBank accession number (which is linked to the GenBank entry at NCBI), a link to the graphic locus view (Figure 2), the information about the location of the T-DNA border/plant DNA junction if detected, the BAC clone code and the confirmation status are shown. If the insertion site qualifies as a gene hit, links to the respective TIGR, TAIR (18), MIPS MAtDB (19) and SIGnAL (10) gene pages are provided. For confirmed lines available in NASC, a link is provided to a page showing all the allele-specific confirmation sequences derived from the respective insertion site. Unlike the original FST sequences, which are of varied quality, the confirmation sequences are generally of high quality and length. The sequences representing the T-DNA were not trimmed from the confirmation sequences. Therefore, the user has access to the exact T-DNA border/plant genome junction structures of the insertion sites.

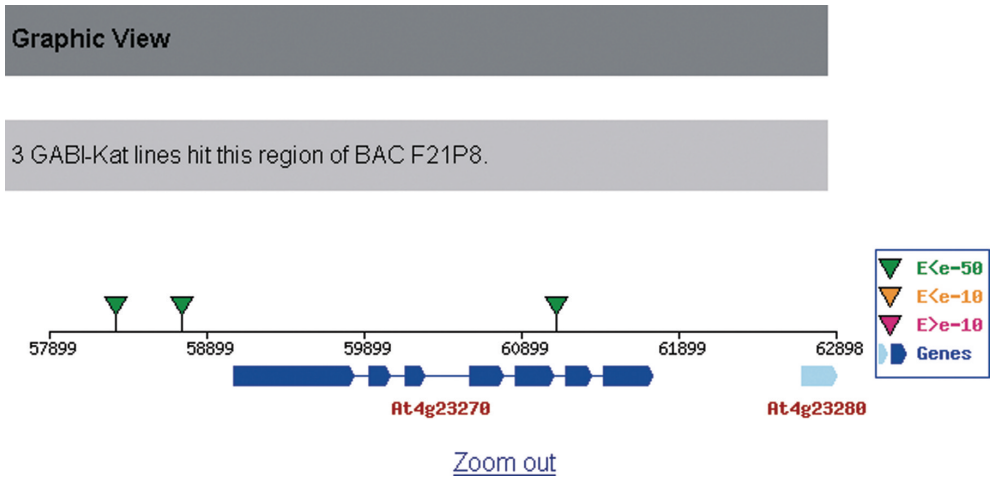

Figure 2. Screenshot of the graphical view page of gene At4g23270. All annotated genes and FSTs in the genome region centered on the locus At4g23270 are presented in an image map. Mouse-over text on the FSTs and the genes tells the line ID and the gene annotation text, respectively. Clicking on the FST icon leads to the respective FST page.

For all the different kind of searches, results are organized around the FST page. When coming to the FST page from searching GenBank accession number or from the BLAST search result page, only the corresponding FST is displayed. If there are more than one insertion from the same line, a link to showing all the insertions of the line is given. When coming to the FST page from searching line ID, all the insertions of the line are shown and for each insertion the best FST is displayed.

AVAILABILITY

The GABI-Kat SimpleSearch database is freely available at http://www.GABI-Kat.de and can be queried through the web interface described above. The FST data were submitted to EMBL/GenBank/DDBJ as genomic survey sequence, and can also be downloaded in a flat file format from our web site. Finally, users can find our FST data at external *A.thaliana* web sites such as MIPS MAtDB (19), FLAGdb++ (20) and SIGnAL (10). However, it is worth noting that minor differences do exist in the FST annotation on these sites when compared with SimpleSearch as they do not necessarily use an identical annotation procedure. Also, the external databases that rely on the FST data in GenBank may indicate the existence of an insertion in a given gene, but the respective line is dead which in turn means that the insertion allele described by the FST does not exist any more.

DISCUSSION

FST-based insertion mutant databases exist for various species for which insertional mutagenesis could be carried out in large scale. For *A.thaliana*, most of the insertion mutant databases were initially developed for holding the data generated in their own projects. While they are quite similar in the FST annotation and query interface, each database has its own unique features. Like GABI-Kat SimpleSearch, FLAGdb initially was a project database of mapped FSTs from the Versailles T-DNA population (8). Currently, the FST data of FLAGdb has been integrated into FLAGdb++ (http://urgv.evry.inra.fr/projects/FLAGdb++/HTML/index.shtml), a database with JAVA front-end integrating different high-throughput functional genomics data for *A.thaliana* (20). ATIDB (Arabidopsis thaliana insertion database; http://atidb.org) is a federated database for archiving FSTs of transposon or T-DNA insertions for *A.thaliana* from different sources including SIGnAL, SAIL, GABI-Kat, FLAGdb and the John Innes Centre (21). Besides the various search options, ATIDB provides analysis tools to study the insertion distribution at a genome scale. The T-DNA Express at SIGnAL (http://signal.salk.edu/cgi-bin/tdnaexpress) is currently the most comprehensive database of mapped FSTs for *A.thaliana* (10). It integrates not only FSTs from sources collected by ATIDB mentioned above, but also include FST data from the Wisconsin T-DNA population (7) and the RIKEN transposon population (22). So far, >360 000 insertion sites have been stored in T-DNA Express, and cover ~90% of all *A.thaliana* genes. The search interface at T-DNA Express provide different ways to access the insertion data. A useful tool at this site is 'iSect toolbox'. Among the many functions this tool offers, a very important one is to design genomic primers to confirm the T-DNA insertions. Although the GABI-Kat SimpleSearch database concentrates on data generated from GABI-Kat, it allows searches and result presentations comparable to these large and complex databases.

What distinguishes SimpleSearch from other databases is that it includes detailed information on single GABI-Kat lines like segregation data, line availability information, insertion site-specific primers and confirmation sequences for >3500 confirmed insertion sites. These data show the exact T-DNA border/plant genome junction structures. This detailed information not only helps users to evaluate the usefulness of the mutant lines, but also covers a big part of the molecular characterization of the insertions. The transfer of confirmed lines to NASC is an ongoing process at GABI-Kat. In addition to confirming lines that are requested by users, we have started to confirm a number of key

GABI-Kat lines. These key lines were identified by analysing the results from the SIGnAL, SAIL, Wisconsin and GABI-Kat FST datasets for unique T-DNA insertion sites in the *A.thaliana* accession Columbia at the level of insertion alleles predicted to cause a knock-out of the respective gene. The seeds of these confirmed lines will also become available from NASC and this will contribute significantly to the goal of finding at least one mutant for every *A.thaliana* gene. Results from the genetic and molecular insertion confirmation pipeline will constantly be added to SimpleSearch.

A few further developments to further strengthen the SimpleSearch database were planned for the future. We are tracking published papers which made use of GABI-Kat mutant lines, and these references are currently listed on our web site on a HTML page . In most of these publications, it was clearly mentioned which GABI-Kat line was used. We plan to store the reference to these publications in the database and link them to the respective lines for display on the FST page. While GABI-Kat SimpleSearch will continue to function as a project database, we are also exploring the possibility for better interoperability with other *A.thaliana* database by ways such as a BioMOBY service (23).

ACKNOWLEDGEMENTS

We thank Nicolai Strizhov, Christa Höricke-Grandpierre and Heike Hagemeyer for PCR, members of the GABI-Kat team for greenhouse and seed handling work, members of the DNA core facility at MPI for Plant Breeding Research and the Chair of Genome Research at Bielefeld University for sequencing. This work was supported by the German Federal Ministry of Education and Research (BMBF) in the context of the German plant genomics program GABI (Förderkennzeichen 0312273). Funding to pay the Open Access publication charges for this article was provided by BMBF/PTJ grant 'GABI-Kat' (0312273) to BW.

Conflict of interest statement. None declared.

REFERENCES

1. Tissier,A.F., Marillonnet,S., Klimyuk,V., Patel,K., Torres,M.A., Murphy,G. and Jones,J.D.G. (1999) Multiple independent defective Suppressor-mutator transposon insertions in *Arabidopsis*: a tool for functional genomics. *Plant Cell*, **11**, 1841–1852.
2. Wisman,E., Hartmann,U., Sagasser,M., Baumann,E., Palme,K., Hahlbrock,K., Saedler,H. and Weisshaar,B. (1998) Knock-out mutants from an En-1 mutagenized *Arabidopsis thaliana* population generate phenylpropanoid biosynthesis phenotypes. *Proc. Natl Acad. Sci. USA*, **95**, 12432–12437.
3. Krysan,P.J., Young,J.C. and Sussman,M.R. (1999) T-DNA as an insertional mutagen in *Arabidopsis*. *Plant Cell*, **11**, 2283–2290.
4. Azpiroz-Leehan,R. and Feldmann,K.A. (1997) T-DNA insertion mutagenesis in *Arabidopsis*: going back and forth. *Trends Genet*., **13**, 152–156.
5. Bechtold,N. and Pelletier,G. (1998) *In planta* Agrobacterium-mediated transformation of adult *Arabidopsis thaliana* plants by vacuum infiltration. *Methods Mol. Biol*., **82**, 259–266.
6. Clough,S.J. and Bent,A.F. (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J*., **16**, 735–743.
7. Sussman,M.R., Amasino,R.M., Young,J.C., Krysan,P.J. and Austin-Phillips,S. (2000) The *Arabidopsis* knockout facility at the University of Wisconsin-Madison. *Plant Physiol*., **124**, 1465–1467.
8. 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.
9. Sessions,A., Burke,E., Presting,G., Aux,G., McElver,J., Patton,D., Dietrich,B., Ho,P., Bacwaden,J., Ko,C. *et al*. (2002) A high-throughput *Arabidopsis* reverse genetics system. *Plant Cell*, **14**, 2985–2994.
10. Alonso,J.M., Stepanova,A.N., Leisse,T.J., Kim,C.J., Chen,H., Shinn,P., Stevenson,D.K., Zimmerman,J., Barajas,P., Cheuk,R. *et al*. (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653–657.
11. Rosso,M.G., Li,Y., Strizhov,N., Reiss,B., Dekker,K. and Weisshaar,B. (2003) An *Arabidopsis thaliana* T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Mol. Biol*., **53**, 247–259.
12. Li,Y., Rosso,M.G., Strizhov,N., Viehoever,P. and Weisshaar,B. (2003) GABI-Kat SimpleSearch: a flanking sequence tag (FST) database for the identification of T-DNA insertion mutants in *Arabidopsis thaliana*. *Bioinformatics*, **19**, 1441–1442.
13. Strizhov,N., Li,Y., Rosso,M.G., Viehoever,P., Dekker,K.A. and Weisshaar,B. (2003) High-throughput generation of sequence indexes from T-DNA mutagenized *Arabidopsis thaliana* lines. *Biotechniques*, **35**, 1164–1168.
14. Altschul,S.F., Madden,T.I., Schaffer,A.A., Zhang,J., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
15. Wortman,J.R., Haas,B.J., Hannick,L.I., Smith,R.K. Jr., Maiti,R., Ronning,C.M., Chan,A.P., Yu,C., Ayele,M., Whitelaw,C.A. *et al*. (2003) Annotation of the *Arabidopsis* genome. *Plant Physiol*., **132**, 461–468.
16. Yadegari,R. and Drews,G.N. (2004) Female gametophyte development. *Plant Cell*, **16**, S133–S141.
17. McCormick,S. (2004) Control of male gametophyte development. *Plant Cell*, **16**, S142–S153.
18. 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.
19. Schoof,H., Ernst,R., Nazarov,V., Pfeifer,L., Mewes,H.W. and Mayer,K.F. (2004) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res*., **32**, D373–D376.
20. Samson,F., Brunaud,V., Duchene,S., De Oliveira,Y., Caboche,M., Lecharny,A. and Aubourg,S. (2004) FLAGdb++: a database for the functional analysis of the *Arabidopsis* genome. *Nucleic Acids Res*., **32**, D347–D350.
21. Pan,X., Liu,H., Clarke,J., Jones,J., Bevan,M. and Stein,L. (2003) ATIDB: *Arabidopsis thaliana* insertion database. *Nucleic Acids Res*., **31**, 1245–1251.
22. Kuromori,T., Hirayama,T., Kiyosue,Y., Takabe,H., Mizukado,S., Sakurai,T., Akiyama,K., Kamiya,A., Ito,T. and Shinozaki,K. (2004) A collection of 11 800 single-copy Ds transposon insertion lines in *Arabidopsis*. *Plant J*., **37**, 897–905.
23. Wilkinson,M., Schoof,H., Ernst,R. and Haase,D. (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics web services. The PlaNet Exemplar Case. *Plant Physiol*., **138**, 5–17.