Nucleic Acids Research, 2004, Vol. 32, Database issue ***D347–D350***
DOI: 10.1093/nar/gkh134

FLAGdb^{++}: a database for the functional analysis of the *Arabidopsis* genome

Franck Samson, Véronique Brunaud, Sylvain Duchêne, Yannick De Oliveira, Michel Caboche, Alain Lecharny and Sébastien Aubourg*

Unité de Recherche en Génomique Végétale (URGV), UMR INRA 1165, CNRS 8114, Université d'Evry Val d'Essonne, 2 rue Gaston Crémieux, CP 5708, F-91057 Evry Cedex, France

Received August 12, 2003; Revised October 16, 2003; Accepted October 27, 2003

ABSTRACT

FLAGdb^{++} is dedicated to the integration and visualization of data for high-throughput functional analysis of a fully sequenced genome, as illustrated for *Arabidopsis*. FLAGdb^{++} displays the predicted or experimental data in a position-dependent way and displays correlations and relationships between different features. FLAGdb^{++} provides for a given genome region, summarized characteristics of experimental materials like probe lengths, locations and specificities having an impact upon the confidence we will put in the experimental results. A selected subset of the available information is linked to a locus represented on an easy-to-interpret and memorable graphical display. Data are curated, processed and formatted before their integration into FLAGdb^{++}. FLAGdb^{++} contains different options for easy back and forth navigation through many loci selected at the start of a session. It includes an original two-component visualization of the data, a genome-wide and a local view, which are permanently linked and display complementary information. Density curves along the chromosomes may be displayed in parallel for suggesting correlations between different structural and functional data. FLAGdb^{++} is fully accessible at http://genoplante-info.infobiogen.fr/FLAGdb/.

INTRODUCTION

Sequencing projects provide putative genes arrayed along linear or circular representations of chromosomes. Thus, in some databases, the visualization of genes is based on a 1D display of gene neighbourhood [e.g. for *Arabidopsis*, TAIR (1) and the GMOD-AtIDB project (2)]. Nevertheless the gene context is far from being limited to neighbouring regions and often involves elements located on different chromosomes. With FLAGdb^{++} we intend to explore further the potentialities of a 1D topological representation and its use for high-throughput (HTP) functional analysis of the *Arabidopsis* genome. *In silico* functional annotation is mainly a process of navigation through various databases followed by a synthesis of all retrieved information. Much as genome sequence annotators desired dedicated graphical tools, such as ARTEMIS (3), HTP functional genomicists need warehouse databases with graphical displays that aid navigation and information synthesis. FLAGdb^{++} is being developed to help analysis of raw data from different HTP functional approaches such as massive sequencing of T-DNA mutants, genome-wide characterization of gene families and transcriptome studies. Selected, valued, quality-labelled and eventually weighed data that are thought to be of direct interest for functional analyses are progressively layered onto the same database scheme underlain by the sequences of the five *Arabidopsis* chromosomes (4).

TECHNICAL DESCRIPTION OF THE DATABASE

FLAGdb^{++} has been implemented in the Relational Database Management System ORACLE v8.i. The database architecture to manage the data around the genomic sequences is based on the previously constructed MICADO (5) and FLAGdb/FST (6) databases. All the graphical interfaces are developed with JAVA JDK 1.4. JAVA WEB START technology is used to facilitate FLAGdb^{++} installation and the automatic upgrade of new releases. The connectivity between the locally installed application and the database is built on client-server architecture using JDBC protocols.

Data content of FLAGdb^{++}

The general concept of FLAGdb^{++} is the anchoring of different kinds of data to genomic sequences, either complete chromosomes or sequenced BAC clones. Integrated data are of a heterogeneous nature but are all described by at least a coordinate set relative to a genomic sequence. Data are selected and curated as a function of their interest for functional analysis. They can be the result of experimental work or generated by bioinformatics prediction software. Figure 1 describes the data integrated into the version 1.4 of FLAGdb^{++}. The general idea is to keep the different data independent as much as possible. Thus, each novel piece of data is mapped to the genomic sequences as a new feature independently of the other features in order to avoid transitive

*To whom correspondence should be addressed. Tel: +33 1 60 87 45 16; Fax: +33 1 60 87 45 49; Email: aubourg@evry.inra.fr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

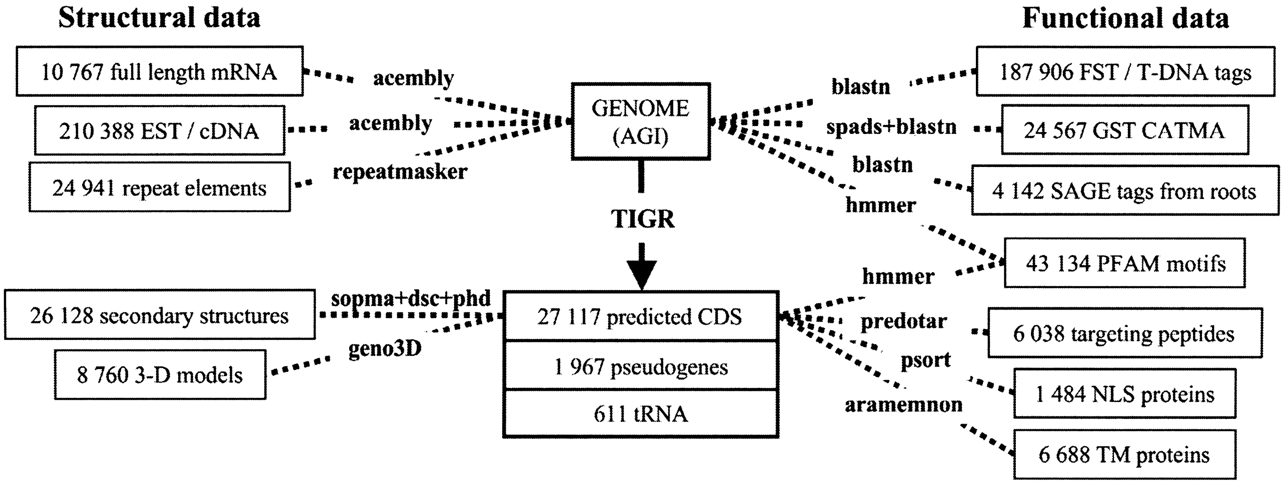


Figure 1. Schematic representation of the FLAGdb^{++} data (v1.4), the treatments performed for their integration into the database and their dependencies. Data are selected depending on the following criteria: FST: the best BLASTn hit, a minimum FST length of 15 bp and a minimum of 90% identity with a chromosome locus; GST: designed by SPADS (7) are all 100% identical to only one locus; SAGE tags are mapped by BLASTn to loci 100% identical to them (31% match to only one locus); Pfam domains have been searched for by HMMER (8), first using the predicted proteome and second using the six-frame translated chromosomes; Nuclear Localization Signals (NLSs) are indicated for predicted proteins giving a nuclear PSORT (v1) score > 0.9 (9). The 2D and 3D structures were predicted using software available at NPS@ (http://npsa-pbil.ibcp.fr) and sequence comparisons with PDB (10); Transmembrane segments come from the ARAMEMNON database (11) and the predictions of signal peptides from PREDOTAR (v2). More details are available in the HTML documentation.

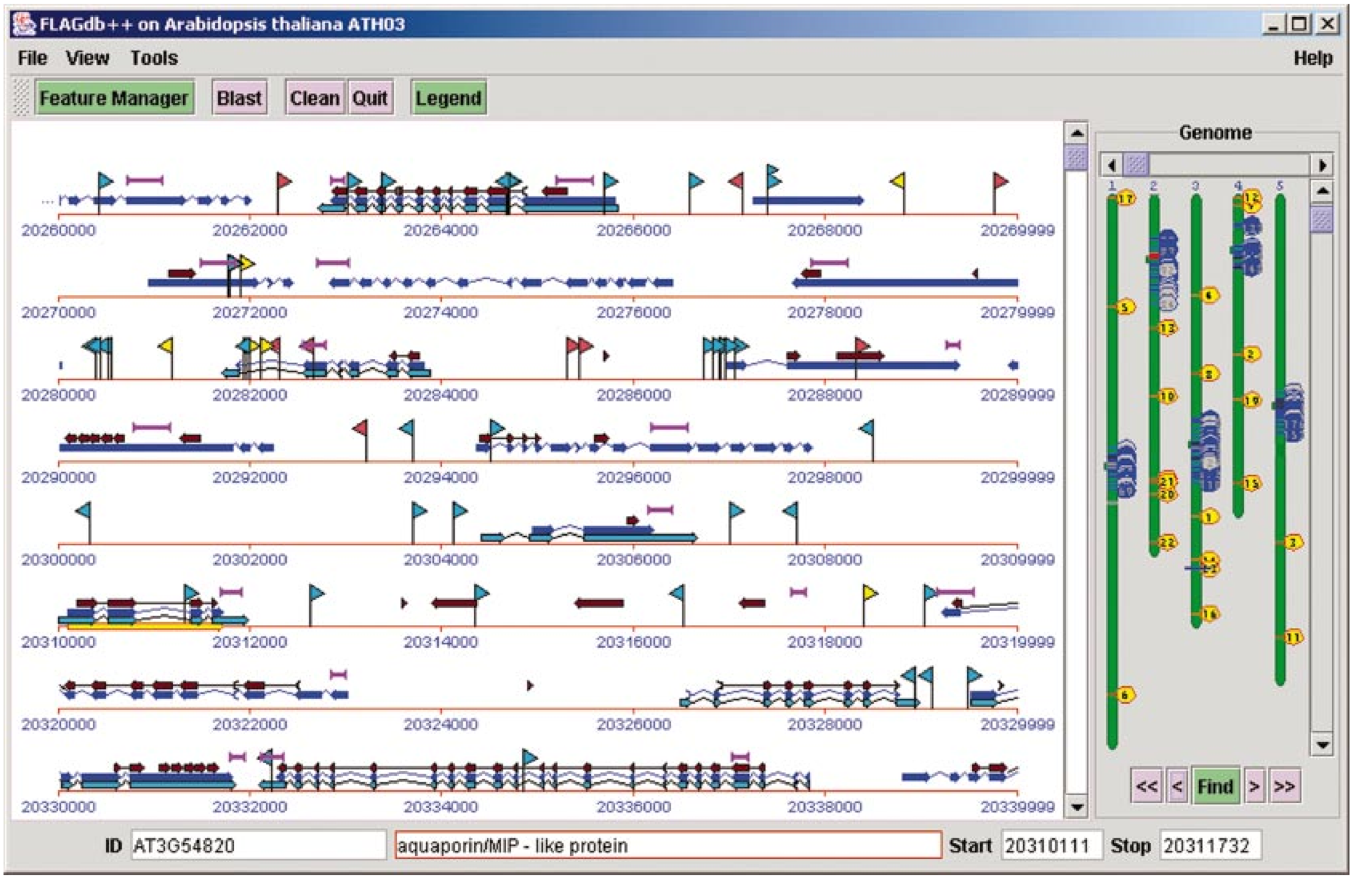
errors coming from erroneous automatic annotation (12). For instance, the detection of the protein motifs in FLAGdb^{++} has been realized by using HMMER and all the Hidden Markov Model (HMM) profiles of Pfam (8) not only on all the predicted Coding Sequence (CDS) of *Arabidopsis* but also on the six frames of the five translated chromosomes. In this way, motifs that fall outside characterized genes are also detected, pointing out putative genes or pseudogenes missed by current annotation. This approach allowed us to highlight 9878 Pfam motif occurrences in regions previously described as intergenic.

FLAGdb^{++} may be used to analyse data from various transcriptome approaches and particularly those based on the CATMA project (13). For this reason, in the database, features like gene-specific Gene Sequence Tags (GSTs) or Serial Analysis of Gene Expression (SAGE) tags are defined by several qualifiers describing not only their origin and details about their nature but, frequently, also giving information on the confidence level we may have in their specificity, the quality of their prediction or the existence of other mapping possibilities, in order to help the user in his interpretation of experimental results. For instance, due to the small size of SAGE tags, 14 bp in the set displayed in FLAGdb^{++}, some of them may match several positions in the genome. In FLAGdb^{++}, different graphical representations make it easy to distinguish SAGE tags mapping at either a unique locus or multiple chromosomal positions and to visualize the various tagged loci. The same approach is used to facilitate the exploitation of Flanking Sequence Tag (FST) resources.

INTERFACES, TOOLS AND QUERIES

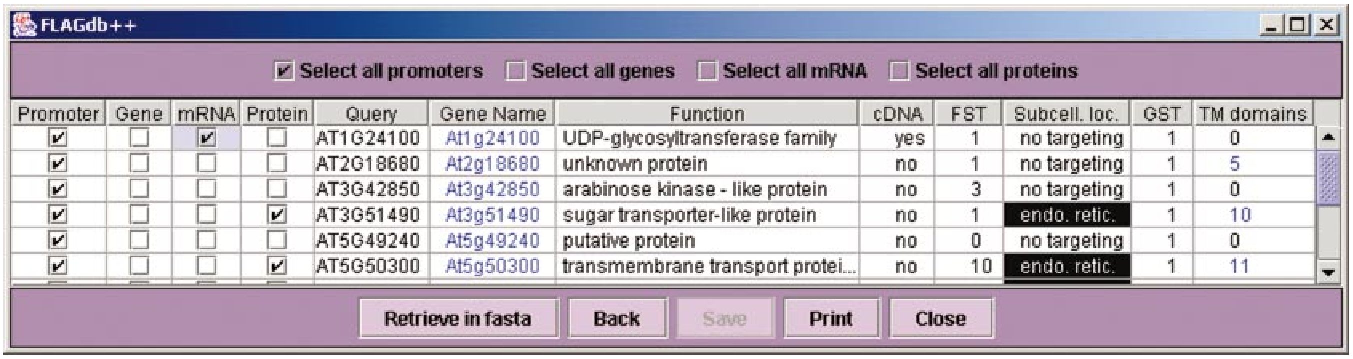
Interfaces and tools in FLAGdb^{++} provide easy navigation through genomic data by allowing a global view of the information. The main interface includes an original two-component visualization of the data, a local and a genome-wide view, whcih are permanently linked and display complementary information (Fig. 2). Features are mapped onto the genome-wide view with a colour code for functionally different characteristics. The genomic interactive view can display lists of genes of interest, representations of families, BLAST results and links between repeat elements (Fig. 2). This system of visualization and the navigation tools are of particular interest in the functional studies of genomes that are modelled by duplication of large fragments and bursts of gene families (14). Indeed, the chromosome representation of all the hits from BLAST or HMMER results allows the consideration of a sequence in the context of its family. Figure 2 shows two different kinds of genome distribution: aquaporin paralogues are dispersed all over the genome whilst ATHILA retrotransposons are specific to peri-centromeric regions. The interfaces and tools have been developed in order to help biologists elaborate new hypotheses and new questions by suggesting correlations between different structural and functional data. An important conceptual effort has been made to design graphical representations of the data that are directly informative about quality and knowledge content.

FLAGdb^{++} offers a direct visualization of BLAST results from a batch query of up to 100 different sequences, with scores and ranks linked to both the genome-wide and the local views. With the aim of comparing and taking advantage of the complementarities between the different features, FLAGdb^{++} generates tables of results displaying organized information corresponding to different genomic elements from a list of accession numbers or sequences. These 'Batch Info' forms or synthetic tables (Fig. 3) give users a convenient tool to compare and retrieve a variety of information about sets of different genomic elements and are useful orientation tables



ID AT3G54820 aquaporin/MIP - like protein Start 20310111 Stop 20311732

Figure 2. The main window of FLAGdb^{++}. In this example, the local view, displaying 100 kb of chromosome 3 sequence, shows CDS and mRNA (accordingly to the TIGR annotation, in blue and green arrows, respectively), GST (purple lines), FST (triangular flags with specific colours for different FST projects) and Pfam motifs (brown arrows). The global genomic view displays the distribution of ATHILA retrotransposons (PF03078; blue spots) and a BLASTP result obtained with an aquaporin protein as query (yellow spots on the chromosomes numbered according to the E-value of the hit).



☑ Select all promoters □ Select all genes □ Select all mRNA □ Select all proteins

Promoter	Gene	mRNA	Protein	Query	Gene Name	Function	cDNA	FST	Subcell. loc.	GST	TM domains
☑	□	☑	□	AT1G24100	At1g24100	UDP-glycosyltransferase family	yes	1	no targeting	1	0
☑	□	□	□	AT2G18680	At2g18680	unknown protein	no	1	no targeting	1	5
☑	□	□	□	AT3G42850	At3g42850	arabinose kinase - like protein	no	3	no targeting	1	0
☑	□	□	☑	AT3G51490	At3g51490	sugar transporter-like protein	no	1	endo. retic.	1	10
☑	□	□	□	AT5G49240	At5g49240	putative protein	no	0	no targeting	1	0
☑	□	□	☑	AT5G50300	At5g50300	transmembrane transport protei...	no	10	endo. retic.	1	11

Figure 3. Table for BLAST results or queries with feature lists. A typical example of a search with a gene cluster from a microarray experiment (clicking on the 'Retrieve in FASTA' button would generate a file with all the promoter, one mRNA and two protein sequences in FASTA format).

offering many links to FLAGdb^{++} genome or local view and to other databases like GenBank, SIGnAL, GABI-Kat (15), ARAMEMNON (11), Pfam (8), PDB (10), MIPS, TAIR (1) and TIGR. A flexible sequence retriever tool allows users to download different types of sequence features or regions like promoters, mRNA, genes or proteins to easily switch to phylogeny studies or search for *cis*-acting regulatory elements.

A 'Help' menu in the main interface gives access to documentation on different available data, their visualization and associated tools.

APPLICATIONS

Statistical analysis of the large collection of T-DNA pre-insertion sites characterized from FSTs of T-DNA-tagged mutant lines (16) and managed in FLAGdb^{++} has led to a deeper insight into T-DNA insertion by illegitimate recombination (17). It also demonstrated the interest of one of the FLAGdb^{++} features, i.e. the visual correlations at a genomic scale, in the density curve window, between occurrences of different features like T-DNA integration sites on one side, and genes on the other side. Besides these studies that are based on extensive queries, FLAGdb^{++} is dedicated mainly to biologists who are involved in different aspects of genome function. A non-exhaustive list of possible usages includes the analysis of syntenic regions between *Arabidopsis* and other plant genomes in candidate gene approaches, searching for unpredicted expressed regions matching with transcripts, evidence for differential splicing and alternative polyadenylation sites and primary access to information for large groups of genes from microarray experiments. For all these applications, data mining is made easy and efficient by the specific design of the FLAGdb^{++} two-component navigation and visualization tools. Transcriptome analysis data are now being intensively generated for *Arabidopsis* since different kinds of microarrray are available (spotted cDNA, GSTs or long oligonucleotides). Nevertheless, wrong or incomplete gene annotations still generate either the absence of targets for unpredicted or wrongly predicted genes and uncertainties on the specificity of some other targets. Evaluation of these probe-dependent restrictions is either already included in FLAGdb^{++} [data associated with mapped GSTs for CATMA (13)] or is easily obtained by querying FLAGdb^{++}.

WORK IN PROGRESS AND FUTURE PLANS

The FLAGdb^{++} structure can easily be adapted to organize data produced on any sequenced genome. Work is under progress to tailor FLAGdb^{++} to the rice genome data. We expect that FLAGdb^{++} will fruitfully participate in the production of other fundamental and biologically important results on the gene context, gene interactions and regulatory networks in *Arabidopsis*. A chromosome-wide representation of a set of microarray experimental results will soon be available. We also expect that the possibilities offered by the two-component visualization tool will be largely exploited in the near future to visualize gene interactions, regulatory networks, proteomes and phenotype–genotype relationships.

SUPPLEMENTARY DATA

A tutorial is available at http://www.evry.inra.fr/tutorial/. It provides the installation protocol for the interface and some examples on how FLAGdb^{++} can be useful for helping analyses of large gene clusters from HTP approaches. This tutorial is extensively linked to full HTML documentation describing the different data and the tools contained in FLAGdb^{++}.

CITING FLAGDB^{++}

Please cite this paper when referencing the FLAGdb^{++} database.

ACKNOWLEDGEMENTS

The authors are grateful to collaborators for producing the biological data or computer predictions integrated into FLAGdb^{++}, especially Ian Small, Jean Thierry-Mieg, Christophe Geourjon, Gilbert Deléage, Vincent Thareau, Carine Serizet, Sandrine Balzergue, Loïc Lepiniec, Georges Pelletier, Joseph Ecker, Bernd Weisshaar, Rainer Schwacke and Alain Gojon. The authors thank Kelly Martinez and Clémence Bruyère for their help during the FLAGdb^{++} implementation; Bertrand Dubreucq for the tests under MAC OS X; Michael Hodges for reading the manuscript; Fabrice Legeai, Delphine Samson, Sébastien Reboux, Isabelle Luyten and Aymeric Duclert for delivering access to the database. This work is supported in part by a GENOPLANTE grant.

REFERENCES

1. 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.
2. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
3. Berriman,M. and Rutherford,K. (2003) Viewing and annotating sequence data with Artemis. *Brief. Bioinform.*, **4**, 124–132.
4. *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
5. Biaudet,V., Samson,F. and Bessières,P. (1998) MICADO: an on-line integrative database for microbial genomes. *Microb. Comp. Genomics*, **3**, 71.
6. 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.
7. Thareau,V., Déhais,P., Serizet,C., Hilson,P., Rouzé,P. and Aubourg,S. (2003) Automatic design of gene-specific sequence tags for genome-wide functional studies. *Bioinformatics*, in press.
8. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
9. Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **2**, 34–36.
10. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 1489–1491.
11. Schwacke,R., Schneider,A., van der Graaff,E., Fischer,K., Catoni,E., Desimone,M., Frommer,W.B., Flugge,U.I. and Kunze,R. (2003) ARAMEMNON, a novel database for *Arabidopsis* integral membrane proteins. *Plant Physiol.*, **131**, 16–26.
12. Kyrpides,N.C. and Ouzounis,C.A. (1999) Whole-genome sequence annotation: 'Going wrong with confidence'. *Mol. Microbiol.*, **32**, 886–887.
13. Crowe,M.L., Serizet,C., Thareau,V., Aubourg,S., Rouzé,P., Hilson,P., Beynon,J., Weisbeek,P., van Hummelen,P., Reymond,P. *et al.* (2003) CATMA: a complete *Arabidopsis* GST database. *Nucleic Acids Res.*, **31**, 156–158.
14. Raes,J., Vandepoele,K., Simillion,C., Saeys,Y. and Van de Peer,Y. (2003) Investigating ancient duplication events in the *Arabidopsis* genome. *J. Struct. Funct. Genomics*, **3**, 117–129.
15. 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.
16. Balzergue,S., Dubreucq,B., Chauvin,S., Le-Clainche,I., Le Boulaire,F., de Rose,R., Samson,F., Biaudet,V., Lecharny,A., Cruaud,C. *et al.* (2001) Improved PCR-walking for large-scale isolation of plant T-DNA borders. *Biotechniques*, **30**, 496–504.
17. Brunaud,V., Balzergue,S., Dubreucq,B., Aubourg,S., Samson,F., Chauvin,S., Bechtold,N., Cruaud,C., DeRose,R., Pelletier,G. *et al.* (2002) T-DNA integration into the *Arabidopsis* genome depends on sequences of pre-insertion sites. *EMBO Rep.*, **3**, 1152–1157.