

# OryGenesDB: a database for rice reverse genetics

G. Droc, M. Ruiz, P. Larmande, A. Pereira<sup>1</sup>, P. Piffanelli, J. B. Morel<sup>2</sup>,  
A. Dievart, B. Courtois, E. Guiderdoni and C. Périn\*

CIRAD, UMR PIA 1096, TA 40/03 Avenue Agropolis 34398 Montpellier Cedex 5, France,

<sup>1</sup>Wageningen UR, Plant Research International, PO Box 16, 6700 AA, Wageningen, The Netherlands and

<sup>2</sup>CIRAD/INRA, UMR BGPI TA 41/K Campus International de Baillarguet 34398 Montpellier Cedex 05, France

Received June 9, 2005; Revised and Accepted September 12, 2005

## ABSTRACT

**Insertional mutant databases containing Flanking Sequence Tags (FSTs) are becoming key resources for plant functional genomics. We have developed OryGenesDB (<http://orygenesdb.cirad.fr/>), a database dedicated to rice reverse genetics. Insertion mutants of rice genes are catalogued by Flanking Sequence Tag (FST) information that can be readily accessed by this database. Our database presently contains 44166 FSTs generated by most of the rice insertional mutagenesis projects. The OryGenesDB genome browser is based on the powerful Generic Genome Browser (GGB) developed in the framework of the Generic Model Organism Project (GMOD). The main interface of our web site displays search and analysis interfaces to look for insertions in any candidate gene of interest. Several starting points can be used to exhaustively retrieve the insertions positions and associated genomic information using blast, keywords or gene name search. The toolbox integrated in our database also includes an ‘anchoring’ option that allows immediate mapping and visualization of up to 50 nucleic acid sequences in the rice Genome Browser of OryGenesDB. As a first step toward plant comparative genomics, we have linked the rice and *Arabidopsis* whole genome using all the predicted pairs of orthologs by best BLAST mutual hit (BBMH) connectors.**

## INTRODUCTION

Rice (*Oryza sativa* L.) has emerged as a model plant for cereal genomics particularly because of its compact genome (389 Mb) (1), the smallest among graminaceous crops, and the availability of vast genetic and molecular resources. Comparative genomics, notably in cereals (2),

offers additional clues to the function of candidate sequences by allowing the reciprocal transfer of information accumulated in other related species to and from rice.

The high quality, full length sequence of the 12 chromosomes of rice has been recently completed by the International Rice Genomics Sequencing Consortium (1) and independent automated annotations have revealed an unsuspected wealth of predicted genes, merely half of which have a clearly identified homolog in the *Arabidopsis* genome. The rice scientific community, which is organized around an International Rice Functional Genomics Consortium, is now facing the challenge of determining the function of most of the rice genes in the next decade. Reverse genetics, which provides a link between a candidate gene and a phenotype, represents the most straightforward experimental strategy to assign a biochemical, cellular, developmental or adaptive role to these sequences (3).

To facilitate the implementation of reverse genetics strategies for gene validation in rice and other cereals and integrate most of the available resources in a convenient platform, we developed OryGenesDB. OryGenesDB is a web accessible, user-friendly navigation tool based on the rice genome sequence originally created to readily retrieve in a single step all the publicly available features around a sequence query. This includes the positional information of flanking sequence tags (FSTs) of insertional mutagens, generated by the ongoing sequence cataloguing effort of rice insertional mutagenesis libraries (4).

## MATERIALS AND METHODS

### Programming and database implementation

OryGenesDB is comprised of web pages delivered by an apache HTTP server (<http://httpd.apache.org>) and the server integrates Perl CGI (5) to dynamically produce web pages based upon user’s input. The interface is generated by custom Perl code (6) that increasingly incorporates object-oriented coding practices to improve extensibility and re-usability of the individual software components. Bioperl (7) is used for specific tasks, such as parsing the Genbank files or TIGR XML

\*To whom correspondence should be addressed at TA40/03 Avenue Agropolis, 34398 Montpellier Cedex 5, France. Tel: +33 4 67 61 71 85; Email: perin@cirad.fr

files containing the *O.sativa* annotation. The interface interacts with a customized database backend utilizing Structured Query Language (SQL) and the open source MySQL (<http://www.mysql.com>) database engine.

We also integrated in OryGenesDB a Genome Browser (GBrowse) based on the Generic Genome Browser developed by Generic Model Organism Project (GMOD) (8). GBrowse is a web-based application for displaying genomic annotations and other features. GBrowse is freely available under an open source licence (<http://www.gmod.org>).

### Database content

We have developed an automated pipeline to incorporate the data fields derived with public data from genome assembly and annotation projects. The workflow of the pipeline is displayed at <http://orygenesdb.cirad.fr/data.htm> and described below.

### Genomic data

The reference annotation layer consists of the 12 rice pseudo-chromosomes released by the Institute for Genomic Research (TIGR) by the division of the Rice Genome Annotation Database and Resource (<http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml>).

The rice genome and its annotation were downloaded from the TIGR FTP site. We also superimposed the Genbank annotations from IRGSP (9) on the TIGR annotations. The annotation data were assembled onto the pseudo-chromosomes and each feature was given new coordinates related to the newly assembled sequence. Dedicated Perl scripts were written to retrieve various attributes of genome features, in particular the gene names, positions in the pseudomolecules and description for all putative proteins. This information was then used as input files into our database.

We have identified 10679 putative pairs of orthologs between *O.sativa* and *Arabidopsis thaliana* (*E*-value cut off 0.1) using the Best Blast Mutual Hit (BBMH) strategy. The protein datasets all.pep and ATH1.pep files for rice (release 3.0) and *Arabidopsis* (release 5.0), respectively, were downloaded from TIGR for ortholog prediction. A reduced dataset of *Arabidopsis* annotation (Version 5.0) and all the predicted orthologs were added in rice and *Arabidopsis*

genome navigators in OryGenesDB to shuttle back and forth between both genomes.

From the FTP site of the Knowledge-based Oryza Molecular Biological Encyclopaedia (<http://cdna01.dna.affrc.go.jp/cDNA/>) we downloaded the distinct tabular text files (10) and mapped 19 552 unique rice full length cDNAs onto the 12 pseudo-chromosomes using KOME cDNA BAC mapping information.

The TIGR Gene Indices (<http://www.tigr.org/tdb/tgi.shtml>) is a collection of species-specific databases that use a highly refined protocol to analyze expressed sequence tag (11). We mapped the Tentative Consensus sequences from rice, maize, wheat, barley and sorghum on rice pseudo-chromosomes using BLASTN (12) with a cut off of  $1e-10$ . A similar procedure was followed to integrate data coming from the Rice Expression Database (<http://red.dna.affrc.go.jp/RED/>) (13).

### Flanking sequence tag mapping

OryGenesDB contains data generated by our group such as T-DNA and Ds FSTs deriving from the genomics initiative Génoplante (14) (<http://genoplante-info.infobiogen.fr/OryzaTagLine/>) and the European consortium Cereal Gene Tags (15). In addition, all of the public FSTs information available from other groups was also integrated (Table 1). The flanking sequences were aligned against the rice pseudo-chromosomes using BLASTN with a cut off of  $1e-10$ . If a flanking sequence had several hits in the rice genome, we chose the hit with the highest *E*-value. The mapped insertions were then assigned to BACs and genes using TIGR genome annotations. A gene was defined as beginning 800 bp before the ATG and ending at the 3'-untranslated region (3'-UTR).

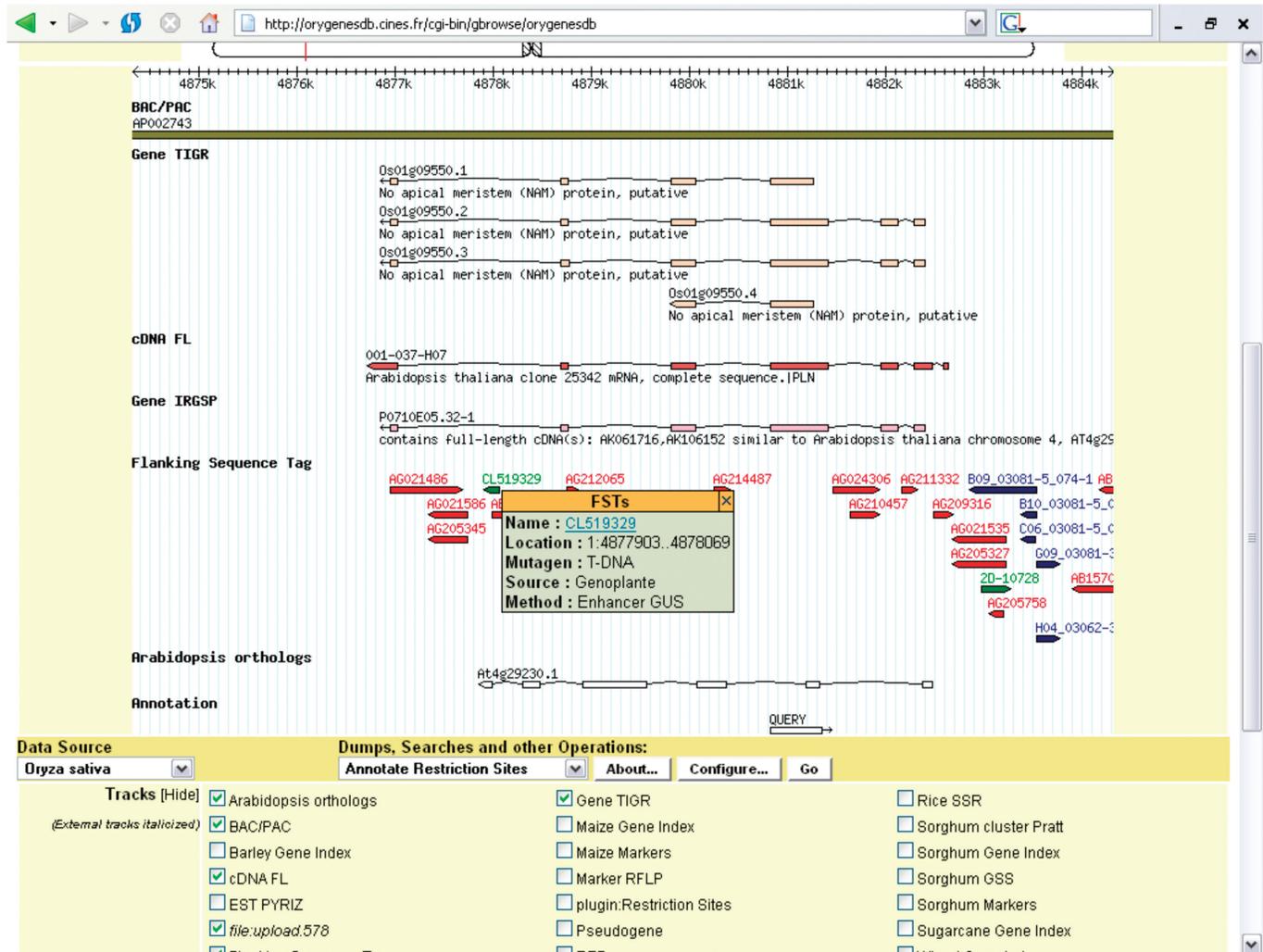
## RESULTS

### GBrowse interface

The main interface of OryGenesDB is based on GBrowse, the Generic Genome Browser (8), developed in the framework of the GMOD consortium. GBrowse is a powerful web-based application for displaying genomic annotations and other features, suitable for any genomes.

**Table 1.** Rice insertion resources integrated in OryGenesDB

Institution	Mutagen	Source	No. of flanking sequences	No. of mapped sequences
CIRAD-INRA-IRD-CNRS, Genoplante CerealGene Tags, European Union National Institute of Agrobiological Sciences	T-DNA	<a href="http://genoplante-info.infobiogen.fr/OryzaTagLine/">http://genoplante-info.infobiogen.fr/OryzaTagLine/</a>	7480	7140
	Ac/Ds	(15)	1381	1381
	Tos17	<a href="http://tos.nias.affrc.go.jp/">http://tos.nias.affrc.go.jp/</a>	18 024	17 933
Zhejiang University	T-DNA	<a href="http://www.blackwell-synergy.com/doi/full/10.1046/j.1365-313X.2003.01860.x">http://www.blackwell-synergy.com/doi/full/10.1046/j.1365-313X.2003.01860.x</a>	1017	917
National University of Singapore	Ac/Ds	<a href="http://www.blackwellpublishing.com/products/journals/suppmat/TPJ/TPJ1948/TPJ1948sm.htm">http://www.blackwellpublishing.com/products/journals/suppmat/TPJ/TPJ1948/TPJ1948sm.htm</a>	1469	1434
Postech PMBBRC	T-DNA	<a href="http://141.223.132.44/pfg/index.php">http://141.223.132.44/pfg/index.php</a>	11 741	11 741
	Ac/Ds	<a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucleotide&amp;cmd=search&amp;term=Rice[ORGN]+CG408311:CG409382[ACCN]">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucleotide&amp;cmd=search&amp;term=Rice[ORGN]+CG408311:CG409382[ACCN]</a>	1072	1040
University of California at Davis	Ac/Ds	<a href="http://www-plb.ucdavis.edu/Labs/sundar/rice/">http://www-plb.ucdavis.edu/Labs/sundar/rice/</a>	1191	1170



**Figure 1.** Insertions within Os01g09550.1. Predicted gene Os01g09550.1, cDNA 001-037-H07 and *Arabidopsis* ortholog At4g29230.1 overlap flanking sequence data. Red, green and blue colors represent Tos17, T-DNA and Ac/Ds insertions, respectively. A contextual popup is open each time the mouse moves under a feature and links or feature-specific actions can be activated. See ([http://orygenesdb.cirad.fr/upload.htm#upload\\_ggb](http://orygenesdb.cirad.fr/upload.htm#upload_ggb)) for a tutorial and a more complete description. The text file containing the mapped sequences can be modified and is directly integrated in OryGenesDB using the GBrowse displaying third party tool. A list of clickable annotated landmarks is displayed and the mapping feature can be visualized as a glyph in the annotation layer (QUERY). See ([http://orygenesdb.cirad.fr/upload.htm#upload\\_ggb](http://orygenesdb.cirad.fr/upload.htm#upload_ggb)) for a tutorial and a more complete description.

At the top of the browser a graphical representation of the 12 rice chromosomes was added. Clicking on a given chromosome allows the user to quickly access to a specific chromosomal region. To select a more specific region of the genome, the user enters its reference in the text field labeled 'Landmark or Region' (e.g. chromosome name, gene name, clone name or accession). GBrowse then fetches the region of the genome specified by the user's search criteria.

Once a genomic region is displayed, the user can navigate through it using a navigation tool bar which allow scrolling and zooming using different scales ranging from base pairs to mega base pairs. Indeed, GBrowse provides a navigation bar which allow scrolling and zooming through arbitrary regions of the genome. Moreover, to get more information on a specific feature, landmarks on each track contain additional links (Figure 1). This could be a page on the browser's website (e.g. IRGSP genes), or a page on an external website (e.g.

NCBI link for the T-DNA insertion lines) or a contextual pop-up window.

## ORYGENESDB QUERY INTERFACE

### 'Gene Search'

'Gene Search' allows retrieving a set of gene/cDNA information with a keyword search through feature annotations. All the matched FSTs inside or in the promoter region of the selected genes are displayed. The user may enter one or several terms in the search text box to restrict the search, including boolean operators (AND, OR and NOT). The 'Flanking Sequence Tag' form enables the user to select the insertion type (T-DNA, Tos17, Ds, Ac or all), 'Gene Annotations' searches keyword terms in text annotation against one or several databases (TIGR, IRGSP, full length cDNA or all),

'Region' restricts the search for FSTs in a specific gene region (promoter, gene, 3'-UTR or all) and 'Orientation' restricts the search in the forward, reverse or both gene orientations. All the results are displayed as a table with details on FSTs, their positions, associated features and a direct link to the rice genome browser. The output table can be downloaded as an Excel™ formatted file.

#### 'Domain Search'

In the 'Domain Search' interface, the genes are selected according to one or more PFAM domains (16). The type of flanking insertion elements, the gene region of insertion and the gene orientation can be used to restrict the search as in the 'Gene Search'.

#### 'BLAST Search'

A FST search in a specific gene can also be done by searching the *O.sativa* genome with a set of query sequences. The user can cut and paste, or type a sequence into the large text window, or upload a file containing nucleotide or protein sequences in FASTA format. By running NCBI-BLAST (12,17) against the DNA or protein database of *O.sativa*, the user can retrieve a specific sequence and localize nearby FSTs. The resulting pages will display the gene and related FST information.

Information resulting from these searches is retrieved as tables which can be downloaded as Excel™ files (Download Result). A link to the output BLAST alignments is also available (View Alignment). The output tables display details on the BLAST matches (hit name, expect value, percentage of similarity and so on). When the selected database is TIGR, the output table contains details on the corresponding FSTs, their positions, a link to the Genome Browser and external related links, and the resulting sequences can be downloaded in FASTA format. The sequence query matches can be downloaded in GFF format to be used as personal annotations in GBrowse.

#### 'Adding annotations to the Rice Genome'

Personal FASTA sequences can be uploaded and viewed in the context of the rice or Arabidopsis genomes. Web forms allow to submit up to 50 nucleic FASTA sequences that can be searched (by BLASTN) on rice or *Arabidopsis* BAC/PAC sequences. The BLAST results are converted in GFF format and integrated directly in GBrowse (Figure 1).

#### 'Locus Search'

This tool was designed to facilitate the batch retrieval of FSTs, for a given set of genes, cDNAs, proteins, transcripts or any other feature contained in OryGenesDB. Using this tool the user can in one single query retrieve efficiently and exhaustively all matching FSTs.

## DISCUSSION

OryGenesDB is a genome database developed for reverse genetics in rice. It is a repository to store insertion data produced by our laboratory (14) and some other related transposon/T-DNA/retrotransposon insertion lines projects

[for a review see (4)]. Most of the rice insertional projects were first developed as institutional databases containing FSTs and few basic tools to query FSTs. Few of them, namely TIGR, RiceGE, Flagdb++ and Gramene integrate a Genome Browser and contain FSTs from several labs. OryGenesDB to date represents the most populated database with >44 166 FSTs and, with RiceGE, is the only one specifically dedicated to rice reverse genetics.

OryGenesDB comes with a dedicated toolbox specifically designed for reverse genetics applications. The set of user friendly interfaces, complementary to GGB tools, developed for OryGenesDB provides the user with several powerful ways to search for insertions in candidate genes. This ensures to search, as exhaustively as possible, for all insertions and their positions for a given list of candidates. As output, an Excel™ formatted file containing all insertions with their position, the matching features and a hyperlink to GGB navigator can be downloaded making them immediately available for further analyses. Raw sequences can also be directly mapped to the rice pseudomolecules using the 'adding annotation tool'. A multi-FASTA file is uploaded and a BLASTN is performed against the whole rice genome. This tool generates the GFF code using the BLAST output and matching sequences can be mapped using the GGB third party annotation tool.

Several further developments of OryGenesDB are underway. We will first continue to add new sequenced insertions publicly available from our laboratory or from other international projects. An ongoing project aims to generate a whole proteome ortholog prediction between *A.thaliana* and *O.sativa* by phylogenomics and integrate both genome ortholog predictions in OryGenesDB. This approach will be extended later to other plants of agronomic interest. As a first step in that direction, we added 10 679 pairs of orthologs predicted by BBMH between rice and *A.thaliana* and linked our Rice navigator with an *A.thaliana* navigator in OryGenesDB through these putative orthologs showing the potential of OryGenesDB for plant comparative genomics. OryGenesDB is now the fastest growing database among those dedicated to rice reverse genetics and will greatly help to take up the next challenge of determining the function of most of the rice genes in the next decade.

## ACKNOWLEDGEMENTS

We would like to thank Laetitia Regnier from CINES (Centre Informatique National de l'Enseignement Supérieur), Montpellier France for hosting OryGenesDB. We acknowledge the support of the EU 5th framework project QLG2-CT-2001-01453 CerealGene Tags and of the Generation Challenge Program. Funding to pay the Open Access publication charges for this article was provided by CIRAD (Centre de coopération internationale en recherche agronomique pour le développement).

*Conflict of interest statement.* None declared.

## REFERENCES

1. International Rice Genome Sequencing Project (2005), The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

2. Feuillet,C. and Keller,B. (2002) Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution. *Ann. Bot. (Lond)*, **89**, 3–10.
3. Bouchez,D. and Hofte,H. (1998) Functional genomics in plants. *Plant Physiol.*, **118**, 725–732.
4. Hirochika,H., Guiderdoni,E., An,G., Hsing,Y.I., Eun,M.Y., Han,C.D., Upadhyaya,N., Ramachandran,S., Zhang,Q., Pereira,A. *et al.* (2004) Rice mutant resources for gene discovery. *Plant Mol. Biol.*, **54**, 325–334.
5. Guelich,S., Gundavaram,S. and Birznieks,G. (2000) *CGI Programming with Perl*, 2nd edn.
6. Christiansen,T., Torkington,N., Sebastobol,C.A. and Wall,L. (1998) *Perl Cookbook*, 2nd edn. Sebastobol CA.
7. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
8. 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.
9. Yuan,Q., Ouyang,S., Liu,J., Suh,B., Cheung,F., Sultana,R., Lee,D., Quackenbush,J. and Buell,C.R. (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res.*, **31**, 229–233.
10. Kikuchi,S., Satoh,K., Nagata,T., Kawagashira,N., Doi,K., Kishimoto,N., Yazaki,J., Ishikawa,M., Yamada,H., Ooka,H. *et al.* (2003) Collection, mapping, and annotation of over 28 000 cDNA clones from japonica rice. *Science*, **301**, 376–379.
11. Lee,Y., Tsai,J., Sunkara,S., Karamycheva,S., Perlea,G., Sultana,R., Antonescu,V., Chan,A., Cheung,F. and Quackenbush,J. (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res.*, **33**, D71–D74.
12. 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.
13. Yazaki,J., Kishimoto,N., Ishikawa,M. and Kikuchi,S. (2002) Rice Expression Database: the gateway to rice functional genomics. *Trends Plant Sci.*, **7**, 563–564.
14. Sallaud,C., Gay,C., Larmande,P., Bes,M., Piffanelli,P., Piegu,B., Droc,G., Regad,F., Bourgeois,E., Meynard,D. *et al.* (2004) High throughput T-DNA insertion mutagenesis in rice: a first step towards *in silico* reverse genetics. *Plant J.*, **39**, 450–464.
15. van Enckevort,L.J.G., Droc,G., Piffanelli,P., Greco,R., Gagneur,C., Weber,C., González,V.M., Cabot,P., Fornara,F., Berri,S. *et al.* (2005) EU-OSTID: A collection of transposon insertional mutants for functional genomics in rice. *Plant Mol. Biol.*, in press.
16. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
17. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., 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.