© 2003 Oxford University Press

Nucleic Acids Research, 2003, Vol. 31, No. 1 179–182
DOI: 10.1093/nar/gkg060

GénoPlante-Info (GPI): a collection of databases and bioinformatics resources for plant genomics

Delphine Samson*, Fabrice Legeai, Emmanuelle Karsenty, Sébastien Reboux, Jean-Baptiste Veyrieras, Jeremy Just and Emmanuel Barillot

Génoplante-Info, Unité de Recherche Génomique-Info, INRA, Infobiogen, 523 Place des Terrasses,
F-91000 Evry, France

Received August 14, 2002; Revised and Accepted September 27, 2002

ABSTRACT

Génoplante is a partnership program between public French institutes (INRA, CIRAD, IRD and CNRS) and private companies (Biogemma, Bayer CropScience and Bioplante) that aims at developing genome analysis programs for crop species (corn, wheat, rapeseed, sunflower and pea) and model plants (*Arabidopsis* and rice). The outputs of these programs form a wealth of information (genomic sequence, transcriptome, proteome, allelic variability, mapping and synteny, and mutation data) and tools (databases, interfaces, analysis software), that are being integrated and made public at the public bioinformatics resource centre of Génoplante: GénoPlante-Info (GPI). This continuous flood of data and tools is regularly updated and will grow continuously during the coming two years. Access to the GPI databases and tools is available at http://genoplante-info.infobiogen.fr/.

INTRODUCTION

Génoplante is a major partnership program in plant genomics, which links public research in France (INRA, CIRAD, IRD and CNRS) and several private companies involved in crop improvement and protection (Biogemma, Bayer CropScience and Bioplante). Génoplante was established in 1999 with five years' funding from the partners, the Ministry of Research, and the Ministry of Agriculture and Fisheries. The partners aim to:

- Develop genome analysis programs for the five main crop species cultivated in France: corn, wheat, rapeseed, sunflower seed and pea. For each species, the Génoplante program is structured around four main topics:
  - analysis of the genome structure,
  - genetic studies on agronomic traits,
  - genetic studies on quality of plant products and
  - genetic studies on disease resistance.
- Develop so-called 'generic' research by:
  - studying the two plant models: *Arabidopsis* and rice,
  - inventing new biotechnological tools for genome analysis,
  - developing bioinformatics tools and databases and setting up the bioinformatics facilities for genome analysis (Génoplante-Info and RhoBioInf platforms) and
  - identifying specific genes of interest in other crops.

Details on the Génoplante program are given at http://www.genoplante.org/.

GénoPlante-Info (GPI) was created in early 2000 as the Génoplante bioinformatics resource centre for the public research partners of Génoplante. GPI is paralleled with the RhoBioInf bioinformatics resource centre on the private side. At GPI, we aim at integrating all the data and results produced by the ~100 Génoplante scientific projects. These data concern EST (Expressed Sequence Tags), FST (Flanking insertion Site Tag), allelic variability, BAC sequences, genome maps, gene and QTL mapping, proteomics and transcriptomics. We build and maintain a bioinformatics environment for analysis of these data. We also offer access to the software tools developed by the Génoplante teams. Public web access to the Génoplante data and tools is available at http://genoplante-info.infobiogen.fr/.

CONTENTS OF GENOPLANTE-INFO

Currently available on the GPI web site are several plant genomics databases:

- The FlagDB database (1), which characterizes a large collection of T-DNA insertion transformants of *Arabidopsis*: it currently comprises about 10 000 sequences flanking T-DNA insertion sites (FST) and interfaces to locate a query sequence on the *Arabidopsis* genome and to compare the results with FST mappings and gene predictions.
- The CATMA (2) database (Complete *Arabidopsis* Transcriptome Micro-Array, also at http://www.catma.org); CATMA was initiated in Génoplante, and is now a European consortium. A complete structural annotation of the *Arabidopsis* genome is offered in CATMA, obtained via new gene prediction software: Eugène, also developed partly

*To whom correspondence should be addressed. Tel: +33 60873704; Fax: +33 60873799; Email: delphine.samson@infobiogen.fr in Génoplante (see next section). For each predicted gene, a specific tag (GST) was computed, and primers designed whenever possible (21 120 GSTs). These two steps were carried out by *ad hoc* software developed by Génoplante: SPADS (see next section).

- The GPI EST database, which contains plant transcript sequences, and the description of the sample and library from which they originate. The database includes ESTs and mRNA from *Sorghum bicolor*, *Medicago truncatula*, *Arabidopsis*, *Zinnia elegans*, wheat, rice, rapeseed, pea and sunflower. The mRNAs have been extracted from EMBL, and the ESTs from dbEST and other sources: 1526 wheat ESTs come from a Génoplante *Triticum turgidum* ssp. *durum*, water-stressed root library (3), 1441 from *T. aestivum* and *Triticum monococcum* libraries coming from ITEC (http://wheat.pw.usda.gov/genome), 11 964 wheat ESTs from CerealsDB (4), 86 623 rice ESTs from the Beijing Genomics Institute (5), 1281 ESTs from a Génoplante *Arabidopsis* library enriched in plasmalemma and tonoplast encoding sequences (6), and 740 ESTs from a Génoplante *Zinnia* library. A global clustering of sequences was performed with the aim of producing gene specific contigs for each of the following species: *Sorghum*, *Medicago*, wheat, *Arabidopsis* and *Zinnia* (see Table 1). The clustering and a first annotation of the sequences were achieved by a suite of programs developed at GPI and based on other software programs mentioned below: the GPI EST pipeline. It includes: exclusion of the contaminated sequences [*Escherichia coli*, yeast, rRNAs or mitochondrial contents found with cross_match, (unpublished, see http://www.phrap.org)], elimination of low quality regions when the chromatograms are available (home made program: gpiseqnet), clipping and masking of the probable vector and of sequence repeats or low complexity regions using RepeatMasker (A. Smit, unpublished, http://repeatmasker.genome.washington.edu), EST clustering and annotation of the contig sequences. Clustering is a two-step procedure starting with a transitive clustering of sequence pairs with sufficient similarity for a given minimal length (e.g. 90% for 80 bp), and then aligning the sequences from each cluster, leading to one or more contigs with consensus sequences. Clustering is based on LASSAP (7) and contiging on CAP3 (8). An *ad hoc* algorithm (Bellerophon) is being designed to pinpoint candidate chimeric sequences from the contig graph structure; such sequences will be eliminated from the contiging process. Annotation is typically done by comparing the contig consensus to SWISS-PROT and SPTrEMBL with a package of GPI tools and the LASSAP BLAST2 algorithm. All the results of these above analyses are stored into the GPI EST database to ensure traceability and to have them readily available from the interface.
- The Plantgene database (Pierre Rouzé and Sébastien Aubourg, personal communication), a library of plant genes with cognate full-length cDNA. All the experimentally cloned full-length cDNA/mRNA available for *Arabidopsis* have been retrieved from EMBL, aligned to the genomic sequences and controlled and corrected with a pipeline of Perl scripts and by human expertise. A homogeneous set of non-redundant documented complete genes with a controled intron–exon structure has been produced. The same protocol has been used for 300 genes from cereals (Philippe Leroy, personal communication).
- A collection of tags obtained by serial analysis of gene expression (SAGE) on a root *Arabidopsis* library, with their level of expression (Hervé Sentenac, personal communication).
- A collection of 200 *Vitis vinifera* microsatellite sequences and primers developed from a Cabernet-Sauvignon cultivar (Didier Merdinoglu, Stéphane Decroocq, Anne-Françoise Adam-Blondon, personal communication).
- A description of the hydrophobic proteome of the chloroplast envelope (9): a subcellular proteomic approach was developed to identify the most hydrophobic envelope proteins from *Spinacia oleracea.* Fifty per cent of the identified proteins (27 out of 54) correspond to previously uncharacterized proteins, most of them being very likely envelope transporters. These proteins and previously known envelope transporters were used to design a test for the prediction of new envelop transporters. This test was run on the *Arabidopsis* proteome and identified more than fifty new proteins as putative envelope transport systems.
- A collection of Amplified Consensus Genetic Markers whose primers were designed for 119 target genes spread over the *Arabidopsis* genome, and which have been shown to amplify *Brassica napus*, *Brassica oleracea* or *Brassica rapa* genes. For 90 genes, PCR amplification on 20 to 24 different cultivars was conducted and sequencing of the PCR products led to the detection of 301 SNPs for 47 genes (Dominique Brunel, personal communication).

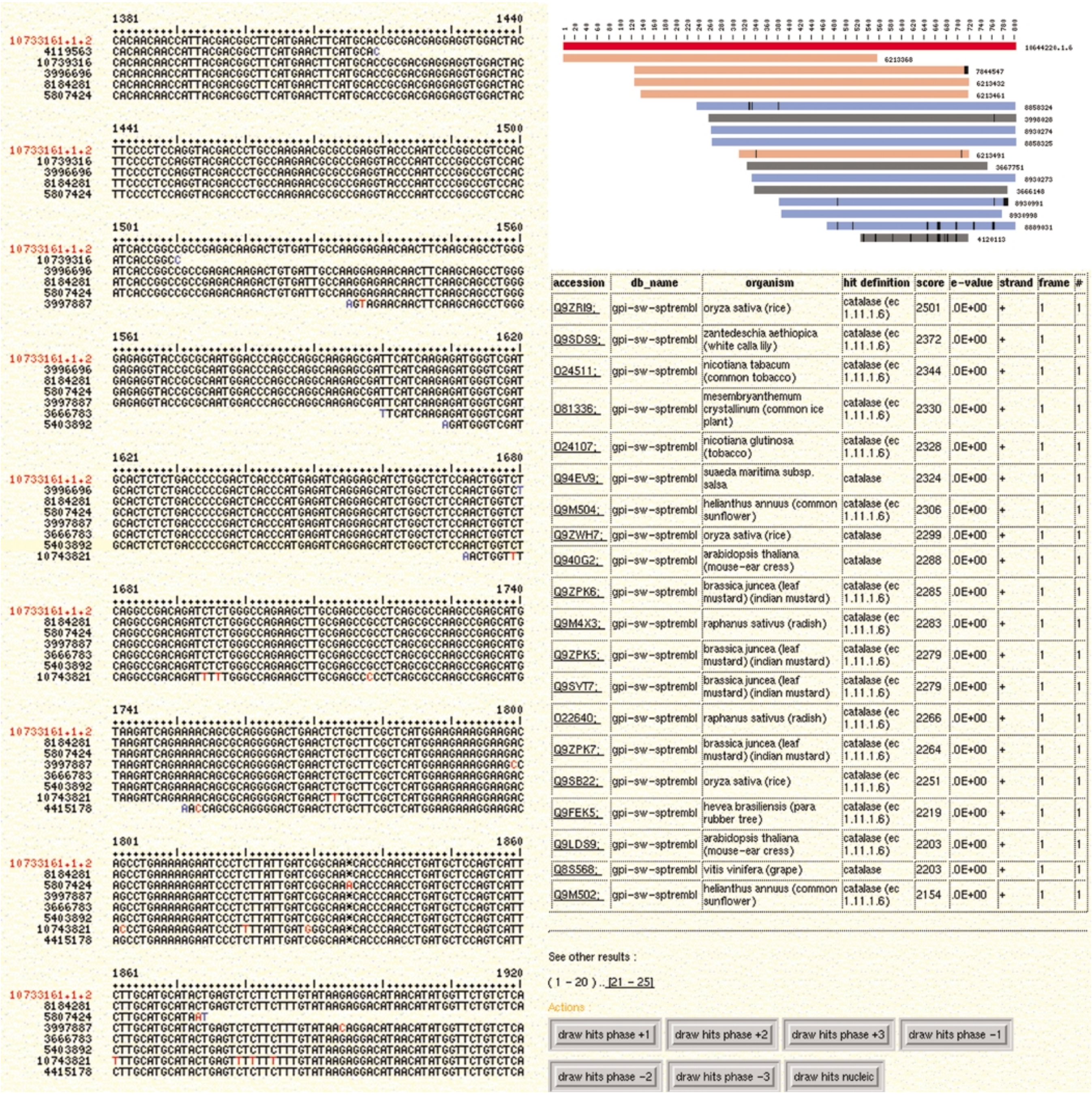
Table 1. Current status of the results of EST clustering at GPI. Contigs for maize, rapeseed, pea, sunflower and rice are will soon follow

Species	mRNA from EMBL	ESTs	Contigs
Sorghum	65	82 961 from dbEST	23 965
Medicago	103	168 391 from dbEST	36 935
Wheat	580	217 856 from dbEST 1526 from Génoplante 11 964 from CerealsDB	59 615
Zinnia	35	6 from dbEST 405 from Génoplante	254
Arabidopsis	24 622	174 624 from dbEST 1281 from Génoplante	29 140

INTERFACES AND TOOLS

Several bioinformatics software tools have been developed in the framework of Génoplante and are accessible at GPI:

- Eugène (10) is a gene prediction software program based on an acyclic directed graph weighted by interpolated Markov models, a start codon prediction program, Netstart (11), and two splice site prediction programs, NetGene2 (12,13) and SplicePredictor (14). Eugène has shown better specificity and better sensitivity than any other gene prediction software on the *Arabidopsis* genome (10).



```
             1381                                                      1440
             +++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|
10733161.1.2 CACAACAACCATTACGACGGCTTCATGAACTTCATGCACCGCGACGAGGAGGTGGACTAC
     4119563 CACAACAACCATTACGACGGCTTCATGAACTTCATGCAC
    10739316 CACAACAACCATTACGACGGCTTCATGAACTTCATGCACCGCGACGAGGAGGTGGACTAC
     3996696 CACAACAACCATTACGACGGCTTCATGAACTTCATGCACCGCGACGAGGAGGTGGACTAC
     8184281 CACAACAACCATTACGACGGCTTCATGAACTTCATGCACCGCGACGAGGAGGTGGACTAC
     5807424 CACAACAACCATTACGACGGCTTCATGAACTTCATGCACCGCGACGAGGAGGTGGACTAC

             1441                                                      1500
             +++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|
10733161.1.2 TTCCCCTCCAGGTACGACCCTGCCAAGAACGCGCCGAGGTACCCAATCCCGGCCGTCCAC
    10739316 TTCCCCTCCAGGTACGACCCTGCCAAGAACGCGCCGAGGTACCCAATCCCGGCCGTCCAC
     3996696 TTCCCCTCCAGGTACGACCCTGCCAAGAACGCGCCGAGGTACCCAATCCCGGCCGTCCAC
     8184281 TTCCCCTCCAGGTACGACCCTGCCAAGAACGCGCCGAGGTACCCAATCCCGGCCGTCCAC
     5807424 TTCCCCTCCAGGTACGACCCTGCCAAGAACGCGCCGAGGTACCCAATCCCGGCCGTCCAC

             1501                                                      1560
             +++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|
10733161.1.2 ATCACCGGCCGCCGAGACAAGACTGTGATTGCCAAGGAGAACAACTTCAAGCAGCCTGGG
    10739316 ATCACCGGCC
     3996696 ATCACCGGCCGCCGAGACAAGACTGTGATTGCCAAGGAGAACAACTTCAAGCAGCCTGGG
     8184281 ATCACCGGCCGCCGAGACAAGACTGTGATTGCCAAGGAGAACAACTTCAAGCAGCCTGGG
     5807424 ATCACCGGCCGCCGAGACAAGACTGTGATTGCCAAGGAGAACAACTTCAAGCAGCCTGGG
     3997887                                   AGTAGAACAACTTCAAGCAGCCTGGG

             1561                                                      1620
             +++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|
10733161.1.2 GAGAGGTACCGCGCAATGGACCCAGCCAGGCAAGAGCGATTCATCAAGAGATGGGTCGAT
     3996696 GAGAGGTACCGCGCAATGGACCCAGCCAGGCAAGAGCGATTCATCAAGAGATGGGTCGAT
     8184281 GAGAGGTACCGCGCAATGGACCCAGCCAGGCAAGAGCGATTCATCAAGAGATGGGTCGAT
     5807424 GAGAGGTACCGCGCAATGGACCCAGCCAGGCAAGAGCGATTCATCAAGAGATGGGTCGAT
     3997887 GAGAGGTACCGCGCAATGGACCCAGCCAGGCAAGAGCGATTCATCAAGAGATGGGTCGAT
     3666783                                        TTCATCAAGAGATGGGTCGAT
     5403892                                                 AGATGGGTCGAT

             1621                                                      1680
             +++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|
10733161.1.2 GCACTCTCTGACCCCCGACTCACCCATGAGATCAGGAGCATCTGGCTCTCCAACTGGTCT
     3996696 GCACTCTCTGACCCCCGACTCACCCATGAGATCAGGAGCATCTGGCTCTCCAACTGGTCT
     8184281 GCACTCTCTGACCCCCGACTCACCCATGAGATCAGGAGCATCTGGCTCTCCAACTGGTCT
     5807424 GCACTCTCTGACCCCCGACTCACCCATGAGATCAGGAGCATCTGGCTCTCCAACTGGTCT
     3997887 GCACTCTCTGACCCCCGACTCACCCATGAGATCAGGAGCATCTGGCTCTCCAACTGGTCT
     3666783 GCACTCTCTGACCCCCGACTCACCCATGAGATCAGGAGCATCTGGCTCTCCAACTGGTCT
     5403892 GCACTCTCTGACCCCCGACTCACCCATGAGATCAGGAGCATCTGGCTCTCCAACTGGTCT
    10743821                                                    AACTGGTTT

             1681                                                      1740
             +++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|
10733161.1.2 CAGGCCGACAGATCTCTGGGCCAGAAGCTTGCGAGCCGCCTCAGCGCCAAGCCGAGCATG
     8184281 CAGGCCGACAGATCTCTGGGCCAGAAGCTTGCGAGCCGCCTCAGCGCCAAGCCGAGCATG
     5807424 CAGGCCGACAGATCTCTGGGCCAGAAGCTTGCGAGCCGCCTCAGCGCCAAGCCGAGCATG
     3997887 CAGGCCGACAGATCTCTGGGCCAGAAGCTTGCGAGCCGCCTCAGCGCCAAGCCGAGCATG
     3666783 CAGGCCGACAGATCTCTGGGCCAGAAGCTTGCGAGCCGCCTCAGCGCCAAGCCGAGCATG
     5403892 CAGGCCGACAGATCTCTGGGCCAGAAGCTTGCGAGCCGCCTCAGCGCCAAGCCGAGCATG
    10743821 CAGGCCGACAGATTTTTGGGCCAGAAGCTTGCGAGCCCCCTCAGCGCCAAGCCGAGCATG

             1741                                                      1800
             +++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|
10733161.1.2 TAAGATCAGAAAACAGCGCAGGGGACTGAACTCTGCTTCGCTCATGGAAGAAAGGAAGAC
     8184281 TAAGATCAGAAAACAGCGCAGGGGACTGAACTCTGCTTCGCTCATGGAAGAAAGGAAGAC
     5807424 TAAGATCAGAAAACAGCGCAGGGGACTGAACTCTGCTTCGCTCATGGAAGAAAGGAAGAC
     3997887 TAAGATCAGAAAACAGCGCAGGGGACTGAACTCTGCTTCGCTCATGGAAGAAAGGAAGCC
     3666783 TAAGATCAGAAAACAGCGCAGGGGACTGAACTCTGCTTCGCTCATGGAAGAAAGGAAGAC
     5403892 TAAGATCAGAAAACAGCGCAGGGGACTGAACTCTGCTTCGCTCATGGAAGAAAGGAAGAC
    10743821 TAAGATCAGAAAACAGCGCAGGGGACTGAACTTTGCTTCGCTCATGGAAGAAAGGAAGAC
     4415178           AACCAGCGCAGGGGACTGAACTCTGCTTCGCTCATGGAAGAAAGGAAGAC

             1801                                                      1860
             +++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|
10733161.1.2 AGCCTGAAAAAGAATCCCTCTTATTGATCGGCAA*CACCCAACCTGATGCTCCAGTCATT
     8184281 AGCCTGAAAAAGAATCCCTCTTATTGATCGGCAA*CACCCAACCTGATGCTCCAGTCATT
     5807424 AGCCTGAAAAAGAATCCCTCTTATTGATCGGCAAACACCCAACCTGATGCTCCAGTCATT
     3997887 AGCCTGAAAAAGAATCCCTCTTATTGATCGGCAA*CACCCAACCTGATGCTCCAGTCATT
     3666783 AGCCTGAAAAAGAATCCCTCTTATTGATCGGCAA*CACCCAACCTGATGCTCCAGTCATT
     5403892 AGCCTGAAAAAGAATCCCTCTTATTGATCGGCAA*CACCCAACCTGATGCTCCAGTCATT
    10743821 ACCCTGAAAAAGAATCCCTTTTATTGATGGGCAA*CACCCAACCTGATGCTCCAGTCATT
     4415178 AGCCTGAAAAAGAATCCCTCTTATTGATCGGCAA*CACCCAACCTGATGCTCCAGTCATT

             1861                                                      1920
             +++++++++|+++++++++|+++++++++|+++++++++|+++++++++|+++++++++|
10733161.1.2 CTTGCATGCATACTGAGTCTCTTCTTTGTATAAGAGGACATAACATATGGTTCTGTCTCA
     8184281 CTTGCATGCATACTGAGTCTCTTCTTTGTATAAGAGGACATAACATATGGTTCTGTCTCA
     5807424 CTTGCATGCATAAT
     3997887 CTTGCATGCATACTGAGTCTCTTCTTTGTATAACAGGACATAACATATGGTTCTGTCTCA
     3666783 CTTGCATGCATACTGAGTCTCTTCTTTGTATAAGAGGACATAACATATGGTTCTGTCTCA
     5403892 CTTGCATGCATACTGAGTCTCTTCTTTGTATAAGAGGACATAACATATGGTTCTGTCTCA
    10743821 TTTGCATGCATACTGAGTTTTTTTTTTGTATAAGAGGACATAACATATGGTTCTGTCTCA
     4415178 CTTGCATGCATACTGAGTCTCTTCTTTGTATAAGAGGACATAACATATGGTTCTGTCTCA
```

accession	db_name	organism	hit definition	score	e-value	strand	frame	#
Q9ZRI9;	gpi-sw-sptrembl	oryza sativa (rice)	catalase (ec 1.11.1.6)	2501	.0E+00	+	1	1
Q9SDS9;	gpi-sw-sptrembl	zantedeschia aethiopica (white calla lily)	catalase (ec 1.11.1.6)	2372	.0E+00	+	1	1
O24511;	gpi-sw-sptrembl	nicotiana tabacum (common tobacco)	catalase (ec 1.11.1.6)	2344	.0E+00	+	1	1
O81336;	gpi-sw-sptrembl	mesembryanthemum crystallinum (common ice plant)	catalase (ec 1.11.1.6)	2330	.0E+00	+	1	1
O24107;	gpi-sw-sptrembl	nicotiana glutinosa (tobacco)	catalase (ec 1.11.1.6)	2328	.0E+00	+	1	1
Q94EV9;	gpi-sw-sptrembl	suaeda maritima subsp. salsa	catalase	2324	.0E+00	+	1	1
Q9M504;	gpi-sw-sptrembl	helianthus annuus (common sunflower)	catalase (ec 1.11.1.6)	2306	.0E+00	+	1	1
Q9ZWH7;	gpi-sw-sptrembl	oryza sativa (rice)	catalase	2299	.0E+00	+	1	1
Q940G2;	gpi-sw-sptrembl	arabidopsis thaliana (mouse-ear cress)	catalase	2288	.0E+00	+	1	1
Q9ZPK6;	gpi-sw-sptrembl	brassica juncea (leaf mustard) (indian mustard)	catalase (ec 1.11.1.6)	2285	.0E+00	+	1	1
Q9M4X3;	gpi-sw-sptrembl	raphanus sativus (radish)	catalase (ec 1.11.1.6)	2283	.0E+00	+	1	1
Q9ZPK5;	gpi-sw-sptrembl	brassica juncea (leaf mustard) (indian mustard)	catalase (ec 1.11.1.6)	2279	.0E+00	+	1	1
Q9SYT7;	gpi-sw-sptrembl	brassica juncea (leaf mustard) (indian mustard)	catalase (ec 1.11.1.6)	2279	.0E+00	+	1	1
O22640;	gpi-sw-sptrembl	raphanus sativus (radish)	catalase (ec 1.11.1.6)	2266	.0E+00	+	1	1
Q9ZPK7;	gpi-sw-sptrembl	brassica juncea (leaf mustard) (indian mustard)	catalase (ec 1.11.1.6)	2264	.0E+00	+	1	1
Q9SB22;	gpi-sw-sptrembl	oryza sativa (rice)	catalase (ec 1.11.1.6)	2251	.0E+00	+	1	1
Q9FEK5;	gpi-sw-sptrembl	hevea brasiliensis (para rubber tree)	catalase (ec 1.11.1.6)	2219	.0E+00	+	1	1
Q9LDS9;	gpi-sw-sptrembl	arabidopsis thaliana (mouse-ear cress)	catalase (ec 1.11.1.6)	2203	.0E+00	+	1	1
Q8S568;	gpi-sw-sptrembl	vitis vinifera (grape)	catalase	2203	.0E+00	+	1	1
Q9M502;	gpi-sw-sptrembl	helianthus annuus (common sunflower)	catalase (ec 1.11.1.6)	2154	.0E+00	+	1	1

See other results :

(1 - 20) .. [21 - 25]

Figure 1. Examples of graphical interface for navigating in GPI EST contigs. On the drawing, the consensus sequence is shown in red, and the aligned ESTs are drawn below. Mismatches are highlighted in black, zooming in is possible up to the sequence level. The colour code can match the library, or cultivar origin. Annotations (blast similarities) and other features (contamination suspicion, repeated elements . . .) can also be viewed with the graphical interface. Left: sequence translations. Below the drawing, the table presents the consensus annotation.

- Predotar is a neural network program (Ian Small, personal communication) that predicts the subcellular localization (plastid, mitochondria, endoplasmic reticulum) of proteins from their N-terminal signal. It has a very high specificity and has been run on the complete *Arabidopsis* proteome.
- SPADS, Specific Primer and Amplicon Design Software (15), generates GSTs (Gene Specific Tags) from the sequence of a genome and its structural annotation. A GST consists in a genome-specific pair of primers amplifying a genome-specific gene sequence, thus allowing the discrimination of genes from the same family.
- The GPI EST Web interface is a navigation environment for the analysis of EST data, that was developed at GPI. Users have access to many simple or advanced query facilities for selecting EST contigs of interest. Requests can concern annotation (e.g. SWISS-PROT keyword based

queries), expression profiles (relative occurrence of sequences from each library in a given contig), virtual *in silico* library substraction ... Interactive graphical interfaces (Fig. 1) concern EST alignments in contigs and annotations. Moreover, the user can naturally do his own analysis (pick primers, do a similarity search in over 80 databases, or align selected sequences) on the raw or translated sequences.

All Génoplante sequences released by GPI are also accessible through a web BLAST server and a web ClustalW server. These servers have been developed with the Pise generator (16).

FUTURE DIRECTIONS

GPI progressively integrates the data generated by the Génoplante programs into its bioinformatics environment and releases it in the public domain. The web site is regularly updated with these new releases. Considerable effort is being made on the design of interfaces integrating all the data and offering comprehensive viewpoints from sequence and map to expression and function.

On the 1st July 2003 Génoplante will release sequences for about 200 000 maize and wheat ESTs. Sequences from several EST libraries for sunflower, rice, rapeseed and pea will also be released during the next two years.

Mapping data for markers, QTL and genes of interest from the Génoplante's species of predilection will also be made public in the coming months, as proteomics (2D gels and mass spectroscopy) and transcriptomics (micro- and macroarrays) data for *Arabidopsis*. *Arabidopsis* FSTs are continuously being produced in Génoplante and are being released every three months. Rice FSTs will also follow soon. Allelic variability, mapping and synteny data for maize, wheat, rapeseed and other species will be released in the near future too.

IMPLEMENTATION

GPI has developed conceptual models for managing ESTs, mapping and synteny, transcriptome, allelic variability, and genomic sequence data. Proteome data modeling is under progress. Conception was done in UML (Unified Modeling Language) and submission formats have been defined in flat files or XML. Databases are developed under ORACLE 8i or Postgres and most software is written in Perl or Java. Database connectivity is based on JDBC or DBD/DBI. The Génoplante-Info environment is accessible at http://genoplante-info.infobiogen.fr

ACKNOWLEDGEMENTS

This work is supported by the Génoplante program. We thank Benjamin Preciado and the Infobiogen team (Guy Vaysseix, Francis Capy, François Laissus, Didier Gillet, Xavier Benigni, Jean-Marc Plaza and Claire Valencien) for their support in system and network administration, and in public data and software installation and maintenance. We are very grateful to our Génoplante colleagues for the submission of their data and/or tools to GPI, their participation in many bioinformatics meetings and task forces, and for their help in defining and enriching the GPI environment. In particular, we would like to express our gratitude towards Johann Joets, Nicolas Sajot, Philippe Lessard, Philippe Dufour, Philippe Leroy, Olivier Gigonzac, Evelyne James, Denis Scala, Virginie Chataigner, Pierre Rouzé, Pierre Hilson, Vincent Thareau, Véronique Brunaud, Franck Samson, Sébastien Aubourg, Alain Lecharny, Ian Small, Thomas Schiex, Jean-Philippe Galaud, Philippe Joudrier, Dominique Brunel, Anne-Françoise Adam-Blondon, Norbert Rolland, Cécile Fizames, Alain Gojon, Michel Caboche, Jean-Loup Risler, Catherine Christophe and many others.

REFERENCES

1. 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.
2. Crowe,M., Serizet,C., Thareau,V., Aubourg,S., Rouzé,P., Hilson,P., Beynon,J., Weisbeek,P., van Hummelen,P., Reymond,P., Paz-Ares,J., Nietfeld,W. and Trick,M. (2003) CATMA—a complete *Arabidopsis* GST database. *Nucleic Acids Res.*, **31**, 156–158.
3. Labhilili,M., Joudrier,P. and Gauthier,M.-F. (1995) Characterization of cDNAS encoding *Triticum durum* dehydrins and their expression patterns in cultivars that differ in drought tolerance. *Plant Science*, **112**, 219–230.
4. Dicks,J., Anderson,M., Cardle,L., Cartinhour,S., Couchman,M., Davenport,G., Dickson,J., Gale,M., Marshall,D., May,S., McWilliam,H., O'Malia,A., Ougham,H., Trick,M., Walsh,S. and Waugh,R. (2000) UK CropNet: a collection of databases and bioinformatics resources for crop plant genomics. *Nucleic Acids Res.*, **28**, 104–107.
5. Yu,J., Hu,S., Wang,J., Ka-Shu Wong,G., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.
6. Galaud,J.-P., Carrière,M., Pauly,N., Canut,H., Chalon,P., Caput,D. and Pont-Lezica,R.F. (1999) Construction of two ordered cDNA libraries enriched in genes encoding plasmalemma and tonoplast proteins from a high-efficiency expression library. *Plant J.*, **17**, 111–118.
7. Glémet,E. and Codani,J.J. (1997) LASSAP, a LArge Scale Sequence compArison Package. *Comput. Appl. Biosci.*, **13**, 137–143.
8. Huang,X. and Madan,A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
9. Ferro,M., Salvi,D., Riviere-Rolland,H., Vermat,T., Seigneurin-Berny,D., Grunwald,D., Garin,J., Joyard,J. and Rolland,N. (2002) Integral membrane proteins of the chloroplast envelope: identification and subcellular localization of new transporters. *Proc. Natl Acad. Sci. USA* **99**, 11487–11492.
10. Schiex,T., Moisan,A. and Rouzé,P. (2001) EuGène: an eucaryotic gene finder that combines several sources of evidence. *Lect. Notes Comp. Science*, **2006**, 111–125.
11. Pedersen,A.G. and Nielsen,H. (1997) Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA, pp. 226–233.
12. Hebsgaard,S.M., Korning,P.G., Tolstrup,N., Engelbrecht,J., Rouzé,P. and Brunak,S. (1996) Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Res.*, **24**, 3439–3452.
13. Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, **220**, 49–65.
14. Kleffe,J., Hermann,K., Vahrson,W., Wittig,B. and Brendel,V. (1996) Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences. *Nucleic Acids Res.*, **24**, 4709–4718.
15. Thareau,V., Déhais,P., Rouzé,P. and Aubourg,S. (2001) Automatic design of specific gene tags for transcriptome studies. In Duret,L. Gaspin,C. and Schiex,T. (eds), *JOBIM 2001*, Toulouse, 195–196.